Applications of Data Mining in Engineering, Management and Medicine

Neha Kaul

ARCLER PRESS

APPLICATIONS OF DATA MINING IN ENGINEERING, MANAGEMENT AND MEDICINE

APPLICATIONS OF DATA MINING IN ENGINEERING, MANAGEMENT AND MEDICINE

Neha Kaul

ARCLER PRESS

www.arclerpress.com

Applications of Data Mining in Engineering, Management and Medicine

Neha Kaul

Arcler Press

2010 Winston Park Drive,

2nd Floor

Oakville, ON L6H 5R7

Canada

www.arclerpress.com

Tel: 001-289-291-7705
001-905-616-2116
Fax: 001-289-291-7601
Email: orders@arclereducation.com

e-book Edition 2019

ISBN: 978-1-77361-615-5 (e-book)

This book contains information obtained from highly regarded resources. Reprinted material sources are indicated and copyright remains with the original owners. Copyright for images and other graphics remains with the original owners as indicated. A Wide variety of references are listed. Reasonable efforts have been made to publish reliable data. Authors or Editors or Publishers are not responsible for the accuracy of the information in the published chapters or consequences of their use. The publisher assumes no responsibility for any damage or grievance to the persons or property arising out of the use of any materials, instructions, methods or thoughts in the book. The authors or editors and the publisher have attempted to trace the copyright holders of all material reproduced in this publication and apologize to copyright holders if permission has not been obtained. If any copyright holder has not been acknowledged, please write to us so we may rectify.

Notice: Registered trademark of products or corporate names are used only for explanation and identification without intent of infringement.

© 2019 Arcler Press

ISBN: 978-1-77361-250-8 (Hardcover)

Arcler Press publishes wide variety of books and eBooks. For more information about Arcler Press and its products, visit our website at www.arclerpress.com

ABOUT THE AUTHOR

Neha Kaul is an experienced Java consultant currently residing in Paris, France and working for one of the leading financial companies in France. She received her double Master's Degree in Computer and Communication Networks and Information Technology from Telecom SudParis and University Paris-Saclay in 2016. She is a recipient of the prestigious Telecom Scholarship for Excellence provided by Fondation Telecom, France. She received the Bachelor of Engineering degree in Computer Engineering from the University of Pune, India in 2011. From 2011 to 2014, she was employed as a Software Engineer with Geometric Ltd, Pune, India. Her major interests include Advanced Java frameworks, Logging Frameworks, Network Security and Data Mining.

TABLE OF CONTENTS

PREFACE

The development of Information Technology sector and the rise in innovation has generated a large amount of data stored in numerous databases in various locations. The development of advanced database management systems has lead to an explosion of data that is stored worldwide. This data is being used in many different applications and fields.

However, we are unable to turn this data into knowledgeable and useful information so as to improve the existing business processes and reduce costs of applications, business, etc. The stored data may be in different formats, like documents, audio/video, numbers, text, figures, hypertext formats, etc. As the data is available in the different types, it should be stored, maintained and analyzed in the correct way. To take full advantage of the available surplus of data; data retrieval only is not sufficient.

A tool or technique capable of automatic summarization of data, extraction, and recovery of the essence of information stored and discovery of possible repeating patterns in raw data is required. With the continuously growing size of data that is being stored in databases, files, and other repositories, it is of the utmost importance, to possess a powerful tool for analysis and interpretation of the data that is capable of extracting interesting information from the data that could help in the decision-making processes. The solution to this problem is 'Data Mining.' Data mining is the analysis of large sums of data "Big data" in order to extract useful information; its main purpose is to make efficient future decisions by learning from existing data. The data cumulated by the different sectors such as Medicine, Engineering, Corporate, Energy, etc., offers promising research options for 'Data Mining.'

In this book, we initially present Data Mining in brief, followed by some Data Mining techniques. Further, the book covers the applications of data mining in the fields of Management, Engineering, and Medicine. Different types of applications in these fields have been detailed along with practical helpful examples.

1 CHAPTER

INTRODUCTION TO DATA MINING

CONTENTS

1.1. WHAT IS DATA MINING?

Data Mining, in simple words, is a means of analyzing data and extracting useful information from the data at hand. The extracted data can further be used by organizations for several purposes such as prediction of trends, developing strategies, improving customer relations, reduce financial costs, and so on.

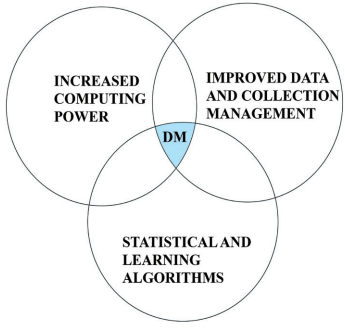
There exist several definitions of Data Mining. Some of them are as follows:

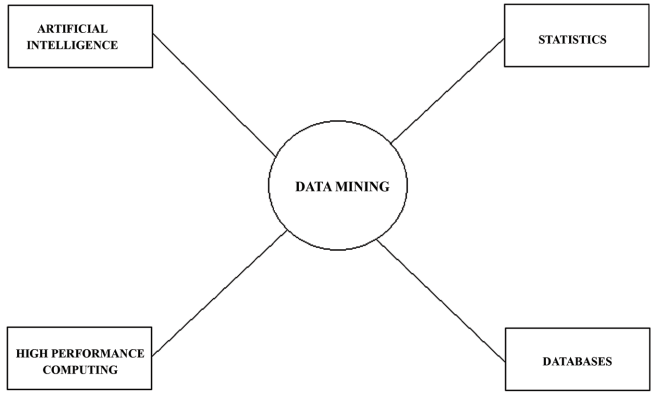
Data mining is a process used by companies to turn raw data into useful information. By using software to look for patterns in large batches of data, businesses can learn more about their customers and develop more effective marketing strategies as well as increase sales and decrease costs. Data mining depends on effective data collection and warehousing as well as computer processing (Staff, 2018).

Data mining is the process of sorting through large data sets to identify patterns and establish relationships to solve problems through data analysis. Data mining tools allow enterprises to predict future trends ("What is data mining? – Definition from WhatIs.com," 2018).

Data mining is the computing process of discovering patterns in large data sets involving methods at the intersection of machine learning, statistics, and database systems ("Data mining," 2018).

Data mining is the process of finding anomalies, patterns and correlations within large data sets to predict outcomes. Using a broad range of techniques, you can use this information to increase revenues, cut costs, improve customer relationships, reduce financial and operational risks and more ("What Is Data Mining?," 2018).





In the process of data mining, smart methods are applied on large quantities of data so as to extract useful information from the large sets of data. The final objective is to extract useful data that can be further manipulated and employed for future use.

Several methods of data mining have been defined so far. The complete detailed process of data mining consists of several steps which are discussed in the upcoming sections.

The term 'Data Mining' is a misnomer, because the real goal of this process is the extraction of patterns and knowledge from large amounts of data, not the extraction (mining) of data itself (Han et al., 2000). A better name that was proposed for data mining is 'Knowledge mining from data' as we actually mine/dig up knowledge from large quantities of data. But, this name was deemed to be too long. Another proposition made was 'Knowledge Mining,' but the patrons of this subject at the time of its invention felt that this name doesn't clearly explain or emphasize the 'mining' aspect of this field.

As mining refers to the discovery of exquisite gems from a wide depth of raw unrefined data, the name 'Data Mining,' a misnomer, that contains both the terms 'data' and 'mining' became popular with the crowd.

Other names given to this process are Knowledge Discovery in Databases (KDD), data/pattern analysis, knowledge extraction, data dredging, data archeology, information harvesting, and so on (Han et al., 2000). The most common synonym of the term 'Data Mining' is KDD.

(Han et al., 2000).

1.1.1. History of Data Mining

Although the field of Data mining appears to be a relatively new and upcoming field, its roots can be easily traced to the 17th Century. At this point of time, a popular method used to identify patterns within data was the Bayes Theorem. This theorem presents a way to relate the current event probability to its previous probability. Further, in the 18th century, another approach similar to that of data mining called as 'Regression Analysis' was being used for pattern finding. This method served to estimate relationships between variables in a system.

In the year 1937, a scientist named Alan Turing spoke of a powerful machine capable of performing high-speed computations similar to our present day computers (Turing, 1937). By the end of the 18th century and the beginning of the 19th century, this concept became a reality with the rise of the Computer Age. The advent of computers permitted the collection, storage, and processing of large quantities of data.

Meanwhile, advanced research in the fields of Neural Networks, Artificial Intelligence, Genetic Algorithms, Statistics and Machine Learning was taking place. With computers generating more and more data, there was an emergence of methods of data storage. More and more sophisticated databases came into the picture. As the power and capacities of the databases kept increasing in the late 1980's, many establishments started keeping records of transactional data. After the successful establishment of Database Management systems, the database technology gradually moved towards advanced database systems such as Data Warehousing, Web-Based Databases, etc. These systems integrated powerful data models object-

oriented models, relational models, deductive models, etc.

This led to an increase in research in the field of advanced database systems and later the introduction of data warehouses containing records of petabytes of information. With the emergence of the World Wide Web (www), a burst of information arrived and contributed a lot of electronic information stored in data warehouses. This vast data was too large and complex to break down and analyze using conventional statistical approaches.

In order to find a solution, several conferences and workshops were conducted. The intent behind these workshops was to leverage and find a way to take the advances made in the fields of Neural Networks, Artificial Intelligence, Genetics and Machine Learning could be applied to analyze large voluminous sets of data. This led to the term 'Data Mining' being coined in the late 1980s within the research communities. By 1990, data mining was considered to be an acknowledged sub-process of an extensive process called as Knowledge Discovery in Database or KDD.

This rise in favorability in this area of research can be accredited to several factors such as the advances in technology, the increasing processing power of computers and vast data storage capabilities. The ready availability of powerful means of data storage combined with powerful computers with high computational capacities meant that the processing of massive volumes of data/information using regular desktop machines was an achievable target.

Eventually, the entire search process led to the foremost International Conference on Data Mining and Knowledge Discovery in 1995 which was held in Montreal. This was soon followed by the launch of a journal called as *Data Mining and Knowledge Discovery* in 1997. During this time, data mining began to make a strong presence in not only the Database domain, but also in the retail and financial domains. Many enterprises dedicated to Data Mining were formed and products were developed. One of the very first applications of 'Data Mining' was its use to detect credit card fraud. Now, it is an actively emerging field and the principle of 'Data Mining' is being applied in various domains.

1.2. TERMINOLOGY USED IN DATA MINING

1.2.1. Data Mining

It refers to extraction of desired information from huge data available in www or databases. It has many applications of which few of them are

market analysis, customer retention, fraud detection, science exploration, disease analysis, etc. ("Data Mining Terminologies | Data Mining Glossary of Terms," 2018).

1.2.2. Knowledge Discovery

It has broad functionalities which include data cleaning, data selection, data integration, data transformation, data mining, pattern evaluation, etc. which shall be discussed in upcoming sections.

1.2.3. Knowledge Base

It is the storage based on the pattern search like the cache in the computer network. This helps in providing quick results for the search when similar patterns are being searched in the future ("Data Mining Terminologies | Data Mining Glossary of Terms," 2018).

1.2.4. Data Mining Engine

It is the main component in data mining system. It performs many core functions viz. association, classification, characterization, prediction, cluster analysis, etc. ("Data Mining Terminologies | Data Mining Glossary of Terms," 2018).

1.2.5. Aggregation

This is a process used in data mining to search, collect and then present the collected data.

1.2.6. Database

A database is a structure that is used to store data digitally. It serves as a collection of data in the digital form and it provides a structure for the organization of the data. Data is fed to a database and it is usually accessed through the use of a database management system (DBMS) ("Big Data A to Z: A glossary of Big Data terminology," 2018).

1.2.7. DBMS

It is software that permits access to the data stored within a database. It collects the data and then provides access in a format that is structured.

1.2.8. Data Center

A data center is a physical location that stores data in servers and other storage devices. It is a large facility that houses several servers containing petabytes or even more quantities of data. A typical center like this can have a single owner that controls all the servers or several organizations that use some of the servers at the center.

1.2.9. Data Collection

This is any method that captures the raw data.

1.2.10. Data Warehouse

A data warehouse is a place that stores data which is further used for the reporting and analysis ("Big Data A to Z: A glossary of Big Data terminology," 2018).

1.2.11. Online Analytical Processing (OLAP)

The process of analyzing multidimensional data using three operations: consolidation (the aggregation of available), drill-down (the ability for users to see the underlying details), and slice and dice (the ability for users to select subsets and view them from different perspectives) ("Big Data A to Z: A glossary of Big Data terminology," 2018).

1.2.12. Online Transaction Processing (OTP)

OLP is a means to supply the end users access to substantial quantities of transactional data in a particular way that enables them to find meaning from the data.

1.2.13. Predictive Analysis

The process of predictive analysis is the use of statistical formulas or functions on one/several sets of data so as to be able to predict trends or events.

1.2.14. Predictive Modeling

The process of predictive modeling is the development of a model that will probably predict a possible outcome or find a trend.

1.2.15. Structured Query Language (SQL)

It is a programming language that provides access to databases and allows manipulation of the elements of the databases. This language permits manipulation of databases with the help of several operations such as insertions, updating, modifications, creation, etc. ("Big Data A to Z: A glossary of Big Data terminology," 2018).

1.2.16. Transactional Data

Transactional data is data that changes in an unpredictable manner and it usually changes rapidly. It is considered to be data that is extracted from transactions. Some of the most common examples of transactional data include bank accounts data, trades, interest on several items, subscriptions, product shipments, invoices, etc.

1.2.17. Data Mining Software

Data mining software is a software program that analyzes data and looks for relationships, pattern, and correlations within the data based on the end user request. Data mining software can find collections/clusters of information analyze this information and find logical relationships, associations, and patterns to come to conclusions about the trends found in the data.

1.2.18. Data Warehousing

The centralization of data into a database, or a database server performed by an organization is called as data warehousing. By means of a data warehouse, an enterprise can select specific data to be analyzed and used.

1.3. DATA MINING PROCESS

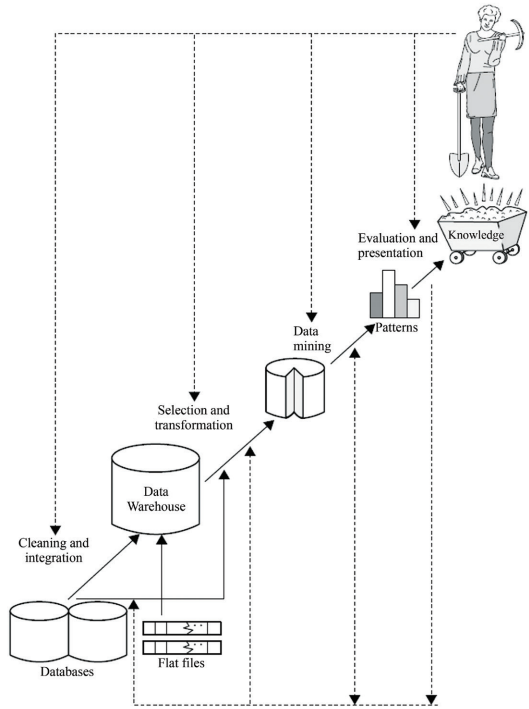
Data mining is currently a well-established discipline. The knowledge discovery in databases (KDD) process has the following stages (Han et al., 2000):

- *Data Cleaning: A data* cleaning process is a form of data preprocessing where noise and inconsistent data are eliminated from the data.
- *Data Integration:* This step involves the integration of several data sources which is a form of data preprocessing.
- *Data Selection:* This is an important step where the data to be

worked upon is selected. It is in this step where data that is relevant to the analysis/mining task is fetched.

- *Data Transformation:* At this stage, the selected data is consolidated and transformed into a form that is suitable for future mining operations with the use of aggregation or summary methods.
- *Data mining:* This is the process that performs the mining of data by means of applying intelligent methods on the data in order to extract useful information or knowledge from it.
- *Interpretation/ Pattern Evaluation:* This stage pertains to the identification of patterns of interest that represent valuable knowledge.
- *Knowledge Presentation:* This is the final step in the KDD process where techniques such as visualization and proper methods to represent the mined learning or knowledge are used to present the final mined information to the end user.

The steps of KDD are as shown below:



(Han et al., 2000).

As we can see in the above image, Data Mining is a step in the KDD process.

1.4. CRISP–DM PROCESS MODEL

Another model that defines the industry-specific standardized steps of data mining has been defined by the CRISP-DM model. It stands for the cross-industry standard process for data mining ("Cross-industry standard process for data mining," 2018). It was defined in the year 1996 as a general model that is a slight variation to above-mentioned process as per industry standards. It defines a total of six stages as given in the following subsections (Dubitzky, 2008).

1.4.1. Business Understanding

This is the initial stage that concentrates on the comprehension of the requirements and objectives of the task at hand from a business perspective. This is crucial as it helps uncover the true end result demanded by the client. Comprehensive analysis of the requirements is essential to determining the factors to be included in the project plan and also to ensure that the project provides the right answers to the right questions. Additionally, in this step, the translation of these requirements into a data mining problem definition is done. A sample example of a data mining goal could be the following – "Predict if the customer will open a bank account with this bank, given their Age, Salary, City, Job Domain, Marital Status, Mortgage." Plus, an initial project plan to attain these objectives is designed in this step. The plan describes an outline of the specific steps to be taken, risk assessment, a proposed timeline and the tools and techniques that would be needed to support the project.

The sub-steps of this step are:

- ascertain the business objectives;
- assess the situation at hand;
- determine the goals of data mining;
- design/create a project plan.

1.4.2. Data Understanding

This step commences with data collection and getting acquainted with the data. If required, data loading and integration is performed as well. In case some problems are encountered during this step, they are properly

documented. This step focuses on finding insights into the data, identifying problems within the quality of data, detecting intriguing/useful subsets within the data set that helps form hypothesis for hidden facts within the data. The key question that is asked is 'Does the given data satisfy the pertinent requirements or not?' A basic understanding of the data is obtained in this step and the succeeding steps build on this understanding. During the verification of the data, common things that are checked are whether the data is incomplete, if some of the fields are blank, spellings of values, ambiguous attributes with similar meanings, and so on.

This step can roughly be broken down into:

- collection of data;
- description of data;
- data quality verification;
- data exploration.

1.4.3. Data Preparation

This step consists of all the activities that are required to construct the dataset that will be sent to the data mining tool and hence it is the final dataset that is obtained from the raw data. This phase is composed of five steps namely data selection, data cleaning, Data construction, data integration, and data formatting. Initially, it is decided what data will be used to solve the mining problem. The decision depends on several factors such as its relevance to the final mining task, its quality and additional technical constraints that may have been applied on the data. An important part of the selection process is determining why a particular data should be included or excluded. Additionally, during selection, it should be decided if one or several attributes of a particular data are valuable than its other attributes or not. During the process of cleaning the data, clean subsets of data are selected and problems encountered during data cleaning are documented. A decision to estimate the missing data can be taken in case of unclean data subsets. The subsequent step is the construction of the data which encompasses the preparation of operations to be performed on the data. An example of an operation is the creation of a new blank record for customers who have not made transactions for the last year.

The data integration step performs the duty of combining information from numerous tables or records in order to create new records/values. The integration of data helps reduce the data that will be mined. The integration

process also performs aggregation of data. For instance, in case of shopping data of a customer, there may be several records for the purchases made by the customer. This data could be aggregated by adding it into a new table with the information about the purchases and a new column named 'number of purchases' that stores the total number of purchases of the customer. The last step is the data-formatting step where the design or format of the data can be changed. Changes in the data format may include simple things like trimming of data, removal of special/illegal characters, reorganization, re-ordering of data, etc.

This step can be divided into the following steps:

- data selection;
- data cleaning;
- data construction;
- data integration;
- data formatting.

1.4.4. Modeling

The modeling step is the most important step in the CRISP–DM lifecycle. In this particular step, one or more modeling techniques are chosen and applied to the data. Different parameters are set and different models can be built for the same data-mining task. This is done as some models have specific needs regarding the form of the data. The selected model is then tested and validated to estimate its quality and validity for the data-mining problem at hand. During the testing of the model, empirical tests are performed to evaluate the strength of the model. The model is then built and assessed as per the domain. The success of the data modeling and discovery techniques are judged during the assessment.

In brief, this step has the following small steps:

- selection of the modeling technique;
- generation of test design;
- creation of the model;
- assessment of the model.

1.4.5. Evaluation

This step serves to evaluate the model and review its construction in order to verify if it achieves the business object and will properly solve

the data mining problem. The results are evaluated to validate the business requirements. The models are assessed based on the business success criteria and the models that meet these criteria and chosen to be the final models that will be applied. It is At this stage, that one reviews the model and determines if certain business cases/issues have not been fully considered by the model. Lastly, the next steps pertaining to the deployment process are discussed. A list of all the possible actions that need to be taken next can be listed at this point.

In short, the steps are:

- evaluation of the results;
- review process;
- determination of the new steps.

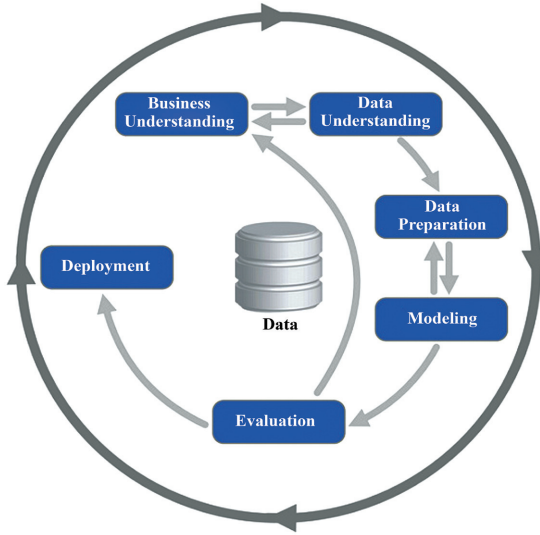
1.4.6. Deployment

The deployment phase consists of first planning a deployment taking into account the results and a strategy of deployment that summarizes the steps to be followed during deployment. Further, a plan is set in place to monitor and maintain the project. A maintenance strategy is developed so as to avoid the wrong use of the results obtained from the use of data mining. Further, a report documenting the data mining project/engagement is produced. This report includes all the deliverables and the results in an organized manner. Also, there is often a meeting that is held at the end of the project so as to present the results to the end user/customer. At the end of the deployment, a review of the entire project that implemented data mining principles is done. The failures and successes are assessed and areas of potential improvement, if any, are documented. The significant experiences that were gained during the progression of the project are included as well. This document can contain common pitfalls, failed approaches, possible tips and tricks regarding the most appropriately suited data mining techniques in similar cases, etc.

The sub-steps of the deployment phase are as follows:

- deployment plan;
- planning of the monitoring and maintenance;
- production of the final report;
- review of the project.

The stages of the CRISP–DM model are represented as shown below:



CRISP–DM Model ("Cross-industry standard process for data mining," 2018).

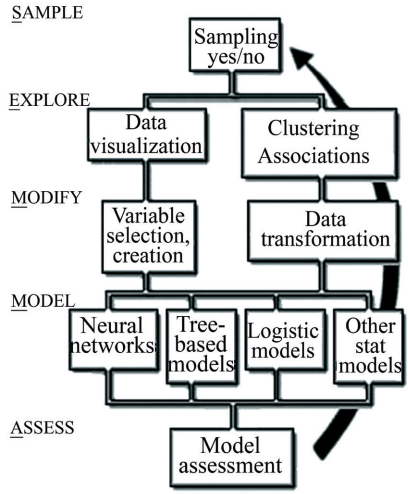
As shown in the diagram above, the CRISP–DM model is composed of 6 steps with arrows indicating the dependency between the phases.

1.5. SEMMA

This is another data mining process model that is a list of sequential steps was developed by the SAS Institute which is one of the largest producers of statistics and business intelligence software ("SEMMA," 2018). The acronym SEMMA stands for Sample, Explore, Modify, Model, Assess ("SEMMA," 2018).

The SEMMA model was introduced to take a data set that is represented statistically so as to easily apply statistical exploration and visualization techniques on your data to predict outcomes and confirm the accuracy of a model. The output of each stage serves as an input to the next stage where it is assessed. In case the output is not enough or requires different parameters, the control is sent back to be explored for additional data refinement. Similar to the CRISP model, this model follows an iterative cycle.

The schema of the SEMMA model is as follows:



("SEMMA – DECISION STATS," 2018).

The steps involved in the SEMMA process model are as given in the following subsections.

1.5.1. Sample

In this step, a part/chunk of a significantly large data set (big enough to be mined) is first extracted. The SAS Institute has preferred to use a strategy for the sampling of data that ensures that the data is now represented in a reliable, statistical format. This sampling is done so as to obtain a statistical depiction of the data set to boost computational performance and maintain costs. The mining of a smaller dataset reduces the processing time by a large fraction. The idea is to find patterns in a relatively smaller data set, which then can be traced to larger datasets as a whole. The key function of this step is to take a portion of a big data set that is big enough to extricate useful information but small enough to manipulate and handle quickly. Moreover, data partitioning is also done during this phase.

1.5.2. Explore

During this step, anomalies, patterns or trends are searched for in the data set so as to better understand the data. Once the sampling is done, the logical next step is the exploration of the data either visually or numerically so as to look for repeating trends, groups, etc. The process of exploration serves to refine the process of discovery of knowledge. In case visualization of the data doesn't produce successful results, additional methods of statistical data

exploration are applied such as factor analysis, clustering and correspondence analysis. If the exploration doesn't yield proper results, the data is redirected back to the sampling step where a new chunk or a smaller chunk is prepared and the cycle continues.

1.5.3. Modify

The Modify step involves the construction of the model. Variables are created, selected and transformed in this step. These variables are the main point of focus for the creation of the model. Here, data is modified by creation, selection, and transformation of variables whose main focus is the model selection process. These variables serve as the decision parameters that are further used to mine the data. Based on the outcomes of the previous step, data may need to be manipulated in a certain way to add more information so as to generate additional variables. Sometimes, it may be required to search and prune the variables by looking for outliers to narrow the list of variables to a select significant few. In short, a selection criterion for the variables is defined and updated if necessary. This model is efficient as it is also iterative which means that once new information is obtained, the data mining models and methods can be changed as well.

1.5.4. Model

The stage of modeling is when we look for a variable combination that successfully predicts an outcome that is desired by the customer. Now, at this point, as the data has been well prepared, we are ready to start the construction of models that are capable of explaining the patterns found within the data. There are several modeling techniques that can be applied to the data such as neural networks, decision trees, logistical models; statistical models such as memory-based reasoning, principal component analysis, and so on. Each model is strong and has its own merits and is suitable only for specific data mining problems depending on the data at hand. There exists software that automatically looks for combinations in the data based on the models.

1.5.5. Assess

The assessment stage is similar to the assessment process followed in the CRISP data model where the worth, merit and reliability of the findings of the mining process are evaluated. This is the final step where the models are assessed and analyzed in order to see how well the models perform. A

proven method of testing a model is its application to a segment of data that was set aside (unused during the construction of the model) during the first stage. If the model is a valid model, then the application of the model will work for this segment of data as well.

The SEMMA model is compatible with the CRISP model as they both are iterative and roughly follow similar steps. This means that once the models have been created and tested, they can be re-deployed in order to gain value in terms of business or research.

1.6. DATA WAREHOUSING OVERVIEW

Before we take a look at various data mining algorithms, models, and techniques, we need to first understand the basic of Data Warehousing.

To begin with, a data warehouse is not a commodity; it is rather an environment. The creation of data warehouses was initiated because earlier corporate data used to be scattered across multiple databases in several formats. Hence, the process of retrieval of information was cumbersome as to obtain the complete information about a particular attribute, it became necessary to interrogate and access all the related databases containing data in different forms, put the small bits of information obtained from each heterogeneous database together and then arrive at the final result set. This entire process was to say the least extensive, cumbersome, inefficient and prone to errors. Additionally, it required a number of manual efforts in terms of analysis. The motivation to solve all of these issues led to the initial thinking behind a data warehouse. An environment capable of bringing together all the data stored in various databases was required. Another requirement of this environment was that one has the possibility of querying this data without actually knowing the actual location of the data. This environment was later developed and came to be known as a data warehouse.

An official definition of Data Warehouse is:

Data warehousing is a technology that aggregates structured data from one or more sources so that it can be compared and analyzed for greater business intelligence (“What is Data Warehousing: Definition | Informatica US,” 2018).

In short, whenever someone requests information, the query is sent to the data warehouse as though it is a single, stand-alone database. But, in reality, it is an environment of databases. The techniques used in the warehousing technology are chosen so as to ensure strategic data management. A few of

the techniques used are Relational Databases, Multidimensional Database Management Systems, Repositories, and so on.

Another way of looking at a Data Warehouse is to think of it as a database with additional features.

A Data Warehouse is a database with the following features:

- the subject-oriented approach that provides a concise view of one/more areas that are selected;
- database that is made up by integrating several databases;
- database that is different than operational databases and provides better performance and storage;
- database that maintains historical data;
- database that provides mostly read-only access.

Some of the general features of a data warehouse are as follows:

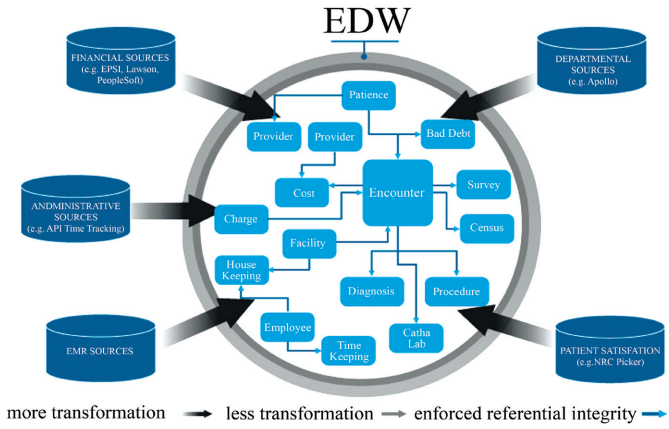
- a data warehouse is built upon a scalable architecture that is capable of handling data expansions in the future.
- it supplies a centralized storage that is capable of storing corporate data and assets.
- a data warehouse has well-defined processes that have been defined and set in place in order to handle loading of operational data.
- it serves as an effective mechanism to provide the completed information requested by the end user by processing the data and convert it into reliable information without needing a lot of technical support for the querying operations.
- the environment of a data warehouse is well managed and secure.

1.6.1. Types of Data Warehouses

Data Warehouses are of the following three types.

1.6.1.1. Enterprise Warehouse

An enterprise data warehouse is a unified consolidated database that stores the business information of an organization and renders it accessible across the organization. This type of warehouse typically serves to cover and include all areas of business for the given company.



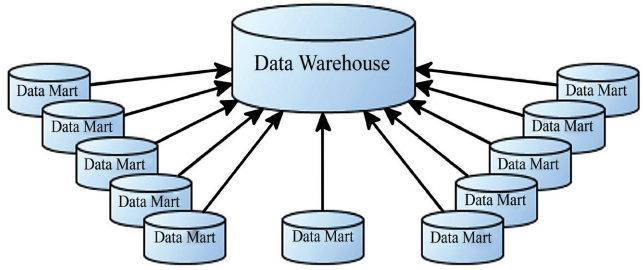
(Barlow, 2017).

1.6.1.2. Data Mart

A data mart is a subset of a traditional warehouse which is generally oriented towards specific business interests or areas. A data mart serves to cover a small subset of the global corporate data that interests a specific group of people. A data mart typically holds only a single subject area (Bonifati et al., 2001).

A typical example of a group of people belonging to a specific business line having interest in a particular data set is the Human Resources team. Another example could be the marketing department.

Generally, a data warehouse is composed of several data marts. This is shown below:

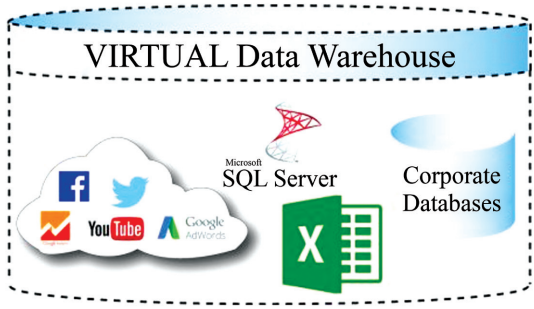


("The Difference Between Data Warehouses and Data Marts – DZone Database," 2018).

1.6.1.3. Virtual Warehouse

A virtual warehouse is a warehouse that provides a set of views of operational databases ("Virtual Warehouse," 2018). By the help of a virtual warehouse, all the data, regardless of its location and regardless of its format, is seen as if it is in a single place and in a consistent or stable format. A virtual warehouse provides access to the data like it is a single entity and it gives us the data in real-time.

Virtual warehouses frequently gather their data from a wide range of sources containing different data formats as well. A virtual warehouse collects the data and displays the collected business data for a specific moment in time, thereby creating a snapshot of the current state/condition of the business data at that particular moment in time.

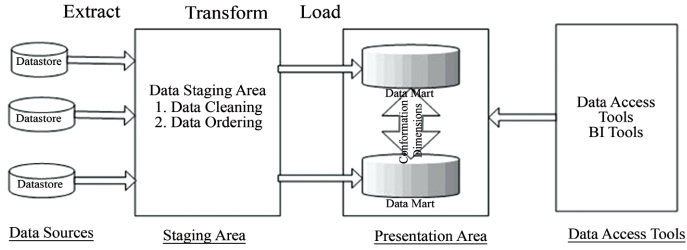


("Why a Virtual Data Warehouse?," 2018).

1.6.2. Data Warehouse Architecture

The data warehouse architecture is a simple layered architecture. Although there exist several different architectures for the data warehousing process, we shall first take a look at a simple layered architecture of a data warehouse. The data is passed from layer to layer and the input from one layer is passed on to the next. The operations performed by the latter steps are based on the output received from the previous step.

The typical architecture of data warehouse consisting of different important components is as shown below:



(Intellipaat, 2018).

The initial and very first layer is called the *Data Source layer* which consists of a combination of several data stores that consist of data stored in several different types of formats. This data may come from corporate relational databases, legacy databases or sometimes from external sources outside the range of the corporation. The data stores can contain data in many forms such as Excel, Text, Relational Databases, and so on. These stores can contain data of several different types and domains.

We proceed to extract the data from the data stores after which the data is put into the data staging area. This process is the Extract process. Here, the data is cleaned and then ordered. No major operations or transformations are performed on the data.

The *Staging area* is divided into two stages namely data cleaning and data ordering stage. Here, in the cleaning step, the processing of data is done and methods used to process the data are applied to the data. The data needs to be first and foremost extracted and then cleansed.

Tasks such as data redundancy removal, filling of data gaps, filtering of bad data, etc. is done at this stage. This is done so as to integrate and merge various heterogeneous data sources into a single common schema. The next step is the Data Ordering step. The staging area step/phase focuses on the application of smart logic to effectuate a smooth transition of the data from a transactional state to an analytical state. This process takes up a decent amount of time for processing of the data obtained from the data stores.

The end stage is the data Presentation layer. This is the final stop for the data. The next operation that is performed on the data is the LOAD operation where the data is now loaded into the presentation area. Here, during the data transformation is done on a major scale.

It is the intended target warehouse where the now cleaned, ordered, integrated and transformed data is deposited in an environment that is multi-

dimensional. At this point, the data is finally available to the end users for querying purposes. Now, the data is available for analysis and query purposes.

Additionally, the data can be made available to the end users or customers in the form of specific data marts.

The three steps that are performed namely Extraction, Transformation and Loading (ETL) are called as ETL tools can merge heterogeneous schema, extract, transform, cleanse, validate, filter and load source data into the warehouse (Bansal, 2014).

After the data presentation stage, the data now becomes available for access via the help of Business Intelligence tools, Data access and Visualization tools. Here, data can be freely accessed so as to generate reports, perform dynamic analysis of information, perform simulations of different business scenarios hypothetically, and so on.

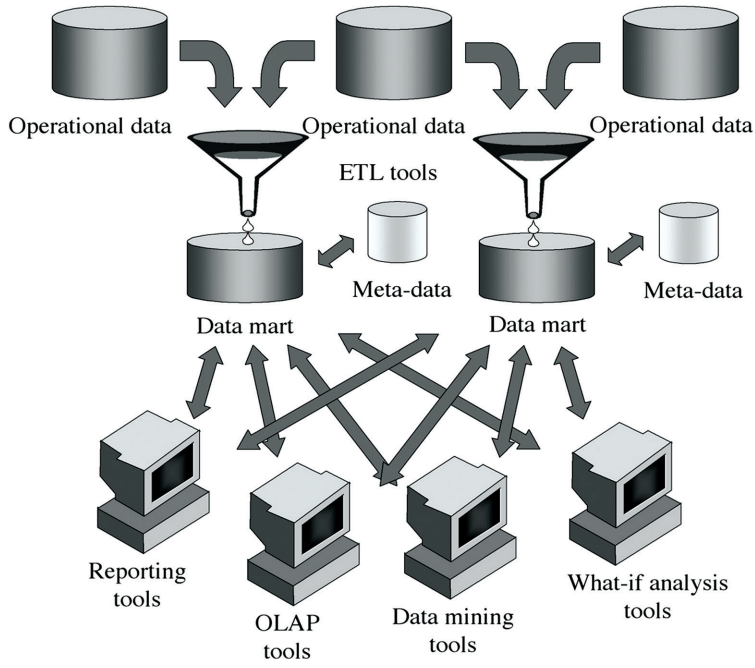
We shall now take look at 3 different architectures of data warehouses. They are as given in the following subsections.

1.6.2.1. Data Mart Bus

The bus architecture is mainly employed when a business need for a specific product arises. The product can be any part of an organization for example 'delivery unit.'

In this case, a single data mart is created for a single process of the business. For instance, first, we develop a data mart for the delivery unit. Then, additional marts are developed based on the approved dimensions of the first data mart that was built. Several marts are joined together with the help of a common bus. The main principle of this architecture is that a set of conformed dimensions (analysis dimensions that preserve the same meaning throughout all the facts that they belong to), derived by a careful analysis of the main process of the enterprise, is adopted and used as a common guideline between marts (Golfarelli and Rizzi, 2009). This serves to make sure to have a logical integration of the data marts and provide an enterprise/organization level view of the data that is stored.

The bus architecture is shown as follows:



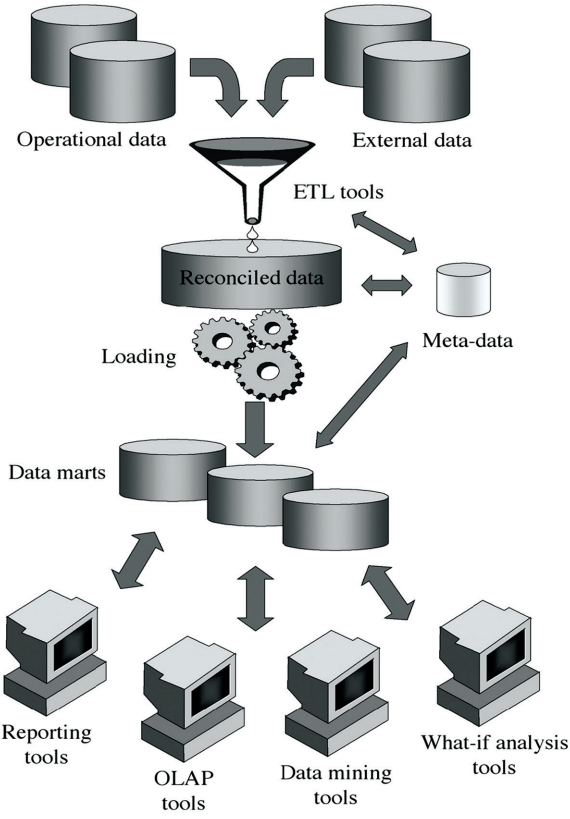
(Golfarelli and Rizzi, 2009).

1.6.2.2. Hub and Spoke

The hub and spoke architecture serves to collect the information from several heterogeneous is an upgraded version of the Data Mart bus architecture with an additional layer called as the reconciled data layer which stores atomic and normalized data. This layer feeds a set of data marts that contain data in multidimensional form. This architecture serves to provide an enterprise-level view of the data and provides an infrastructure that is scalable and maintainable. The infrastructure consists of a hub that is centralized that is capable of accepting requests from numerous applications that are connected via spokes, hence the name hub and spoke. The total integrated data is distributed to each of the data mart from the central warehouse.

A potential drawback is that the data marts cannot communicate with one another.

The hub and spoke architecture is as shown below:

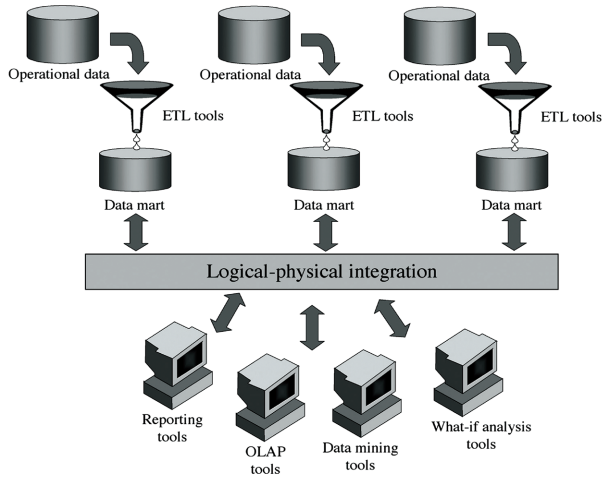


(Golfarelli and Rizzi, 2009).

1.6.2.3. Federal

The Federal data warehouse architecture is a dynamic architecture where already existing warehouses are integrated in a non-evasive way so as to provide a single view across organizations. Each of the warehouses and marts under consideration are virtually or physically integrated with others with the help of different technologies such as distributed querying, meta-data manipulations, etc. Basically, in case of Federal data warehouse, we leave behind traditional decision support systems and proceed to a new level where data between different systems is integrated. A federal architecture can be implemented across geographical regions as well.

The federal architecture is as shown below:



(Golfarelli and Rizzi, 2009).

1.6.2.4. Centralized Data Warehouse

The centralized data warehouse architecture is a specific implementation of the hub and spoke architecture where the reconciled layer and the data marts form a single combined layer and are maintained in a single physical repository. Here, the data marts are not dependent on one another. Similar to the hub and spoke model, the data warehouse maintains atomic level data, summarized data and a logical or enterprise-level view of the data.

1.7. OLAP: ONLINE ANALYTICAL PROCESSING

OLAP or Online Analytical Processing is a technology for discovery of data, a technology that provides limitless reporting functionalities, complex analytical calculations, forecasting and many other features based on multidimensional analysis of data. It permits organizations and the end users to get processed information in a reliable, fast manner with interactive access. OLAP is extremely useful when the end user's requirements are particularly difficult to define and hence allows the user to analyze and explore the data as per his needs based on the model that is multidimensional. This technology offers the end users to start a complex session in a proactive manner where each one of the steps is an output of the previous step. An advantage of this technology is that the design is user-friendly, flexible and easy to use by people from all domains, including people that are not familiar with information technology.

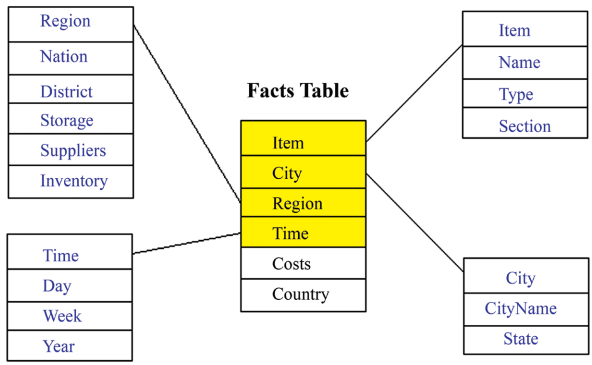
There exist following three types of OLAP models.

1.7.1. Relational OLAP (ROLAP)

This is a relational architecture that makes use of relational or extended database management systems to store and manage all the data stored in the Data warehouse. Additionally, this architecture stores a specific type of middleware that is used to execute and support OLAP queries. It follows a table-oriented organization.

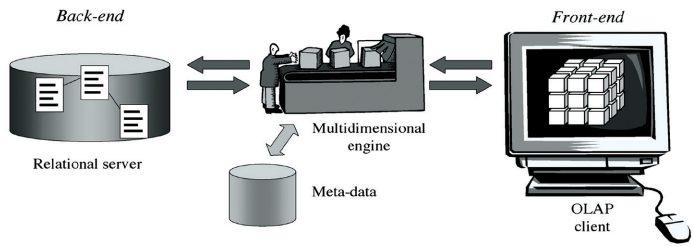
ROLAP makes use of the star schema. A central table called as the facts table is present which stores all the information/facts. Several dimension tables are present that specify the various dimensions that give meaning to the facts table.

It is shown below:



The tables shown in blue are the dimension tables whereas the facts table stores all the records. The facts table is a large table and has a relation with all the dimension tables.

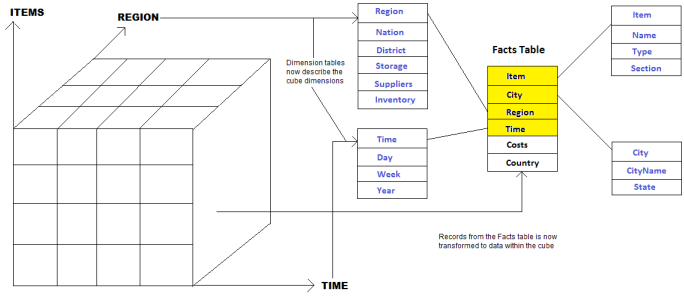
The architecture of ROLAP is as shown below:



(“Similarities and differences between ROLAP, MOLAP and HOLAP,” 2018).

1.7.2. Multidimensional OLAP (MOLAP)

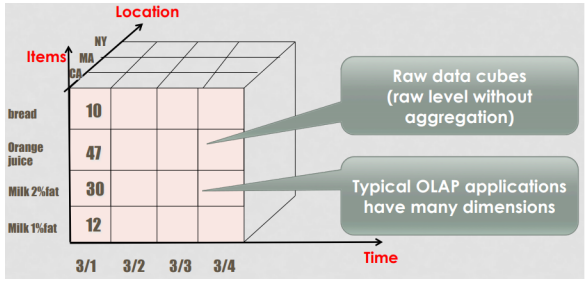
This architecture makes use of array-based data structures along with aggregated and *pre-computed data*. It is an OLAP implementation based on multidimensional database management systems. It offers storage on a multidimensional scale and provides a multidimensional view of the stored data. This implementation offers a better performance that standard OLAP. MOLAP has a cube implementation of the data as compared to the star schema of ROLAP. The ROLAP schema we saw previously is depicted in a cube in the following way:



1.7.3. Hybrid OLAP

The hybrid model is a combination of the ROLAP and the MOLAP models where a relational database concept is used along with pre-computed data. *It combines the strong aspects of the ROLAP and MOLAP models.* From ROLAP it inherits scalability whereas from MOLAP it inherits faster computation thanks to pre-computed data.

The dimensions used in OLAP represent different perspectives that are used to analyze the data. For example, consider a three-dimensional view of grocery store purchases in various cities in USA. The 3d cube would be as shown below:



(Eltabakh, 2018).

We have three dimensions: Location, Items and Time

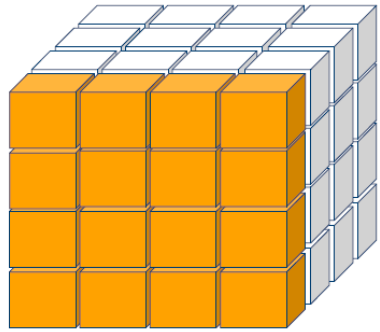
- The Location dimension represents the city of purchase of the products.
- The Items dimension details the items that were purchased.
- The Time dimension stores the date of purchase of the items.

Every single cube represents some raw data which is later queried and aggregated.

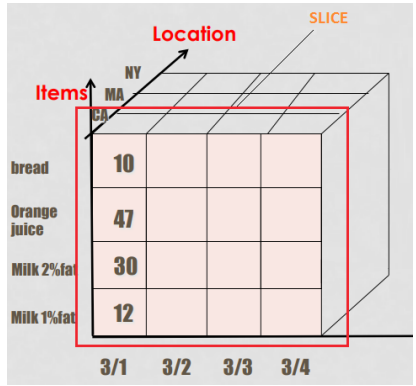
Now that we have seen different OLAP models, we shall now take a look at the various operations performed by the OLAP technology.

The list of OLAP operations is the following:

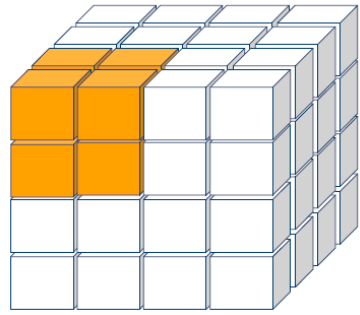
1. *Slice and dice:* The slice operation refers to slicing a section of the cube as follows:



The slice operation is basically a selection of a section of a cube to answer a query. If we consider our example of grocery store purchased items, a slice operation example would be detailing all the items purchased in the city of California.

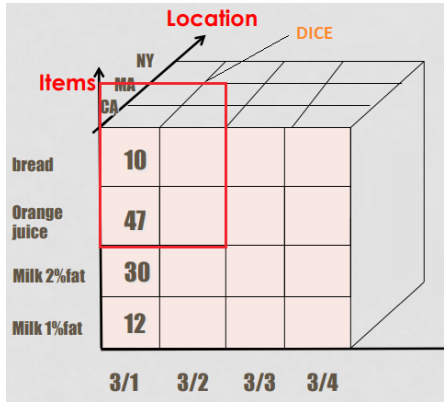


The dice operation is the division of each dimension in the cube. It refers to the product of one or more dimensions for a fixed range. It is shown as follows:



For our earlier example, a dice operation would be retrieving details of particular items (bread and orange) purchased in 2 states (CA and MA) for 2 days.

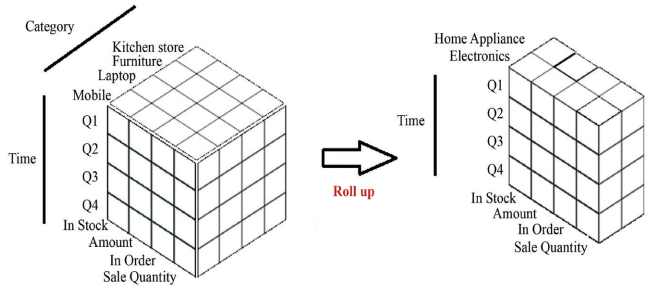
This is shown as follows:



(Eltabakh, 2018).

2. *Roll-Up:* A roll-up operation involves providing a high-level view of the data. It involves computing all the data for one or more dimensions. It is basically aggregated data that is obtained by grouping one or more dimensions together. It refers to an aggregation of one or more dimensions of the cube. It provides a more compact view of the data.

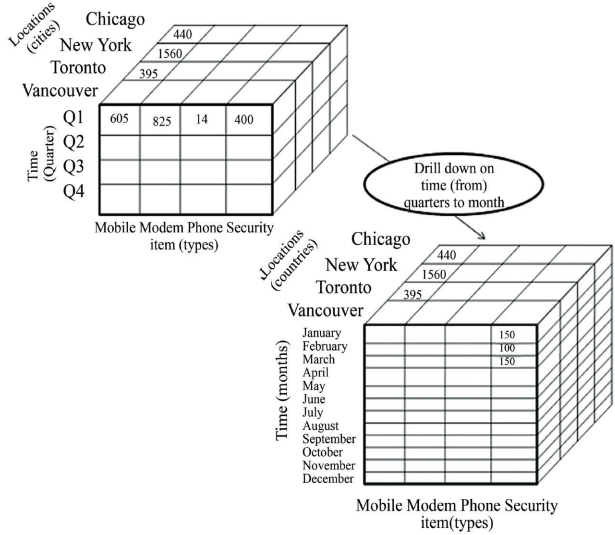
   An example of roll-up is as follows:



("Complete Data Warehousing OLAP Tutorial [2017 Guide]-Cracklogic," 2018).

3. *Drill-Down:* The drill-down operation is where we obtain a detailed in-depth view of the data. We go through all the levels of the data ranging from the most summarized (UP) data down to the most granulated data (DOWN). This operation is exactly the opposite of the Roll-up operation as here we have a granulated view of the data whereas a roll-up operation provides an aggregated view.

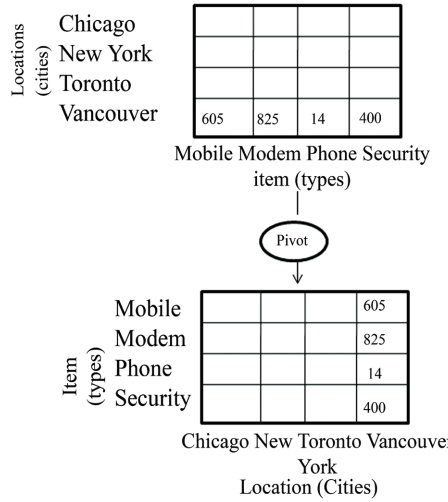
   An example of the drill-down operation is as follows:



("Data Warehousing OLAP," 2018).

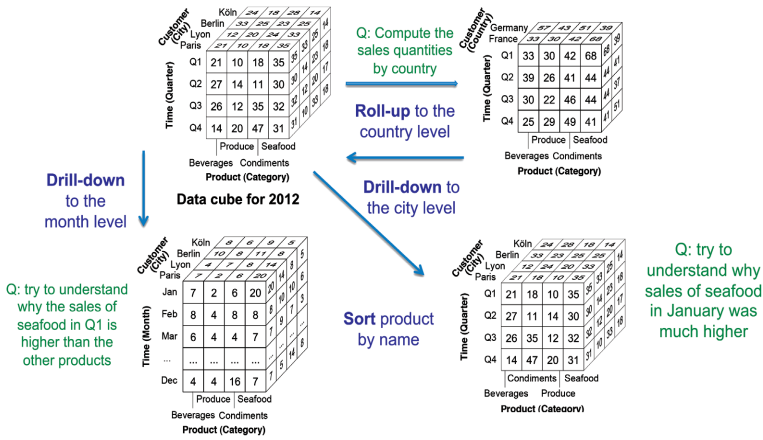
4. *Pivot:* This operation as its name suggests pivots the view/ dimensions so as to change the axe. This view helps change the orientation of the dimensions and offer a pivoted view of the data.

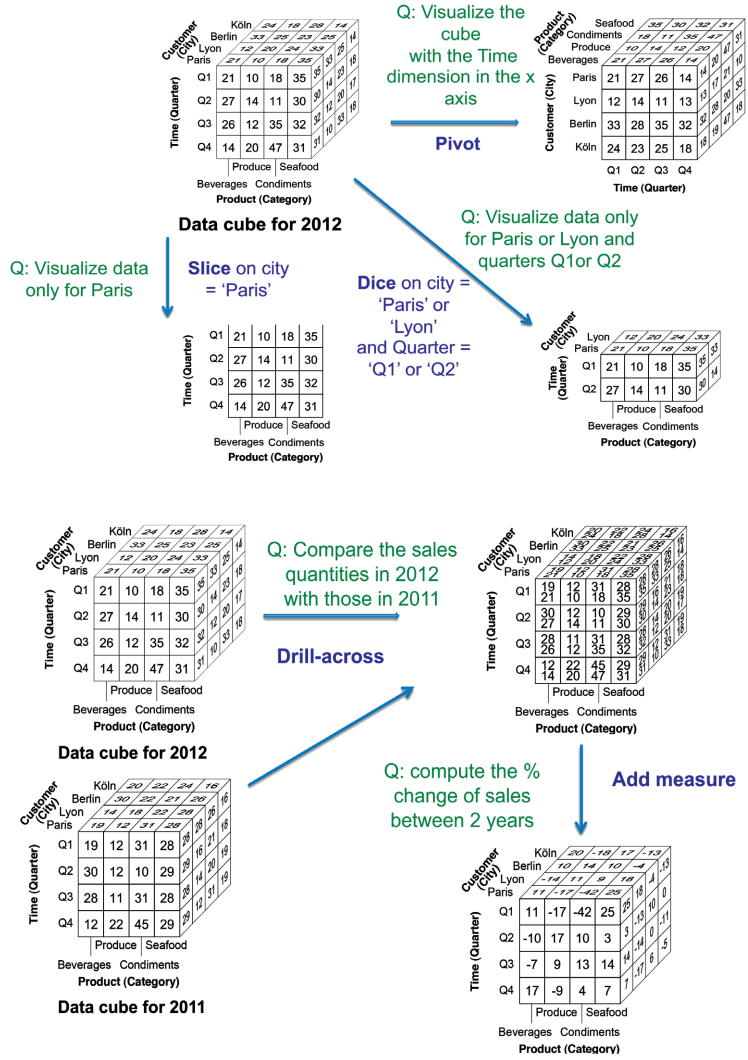
An example of the pivot operation is as follows:

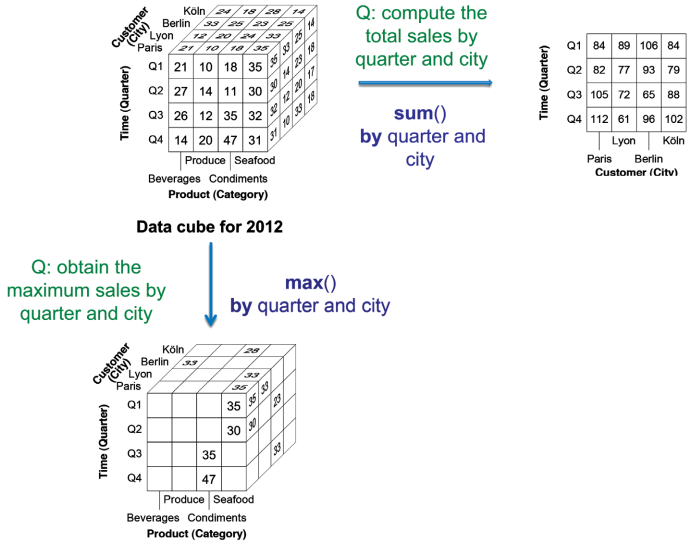


("Data Warehousing OLAP," 2018).

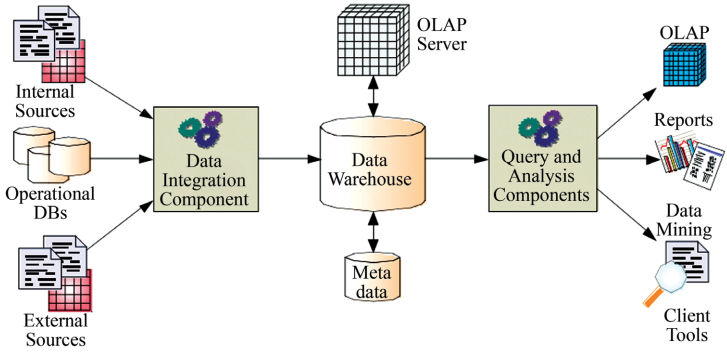
Below are some additional examples of OLAP operations (Galhardas, 2013):




Q: Visualize the cube with the Time dimension in the x axis
Pivot
Data cube for 2012
Q: Visualize data only for Paris
Slice on city = 'Paris'
Dice on city = 'Paris' or 'Lyon' and Quarter = 'Q1' or 'Q2'
Q: Visualize data only for Paris or Lyon and quarters Q1or Q2
Q: Compare the sales quantities in 2012 with those in 2011
Drill-across
Data cube for 2012
Data cube for 2011
Q: compute the % change of sales between 2 years
Add measure
Time (Quarter)
Customer (City)
Product (Category)
Beverages Condiments
Produce Seafood



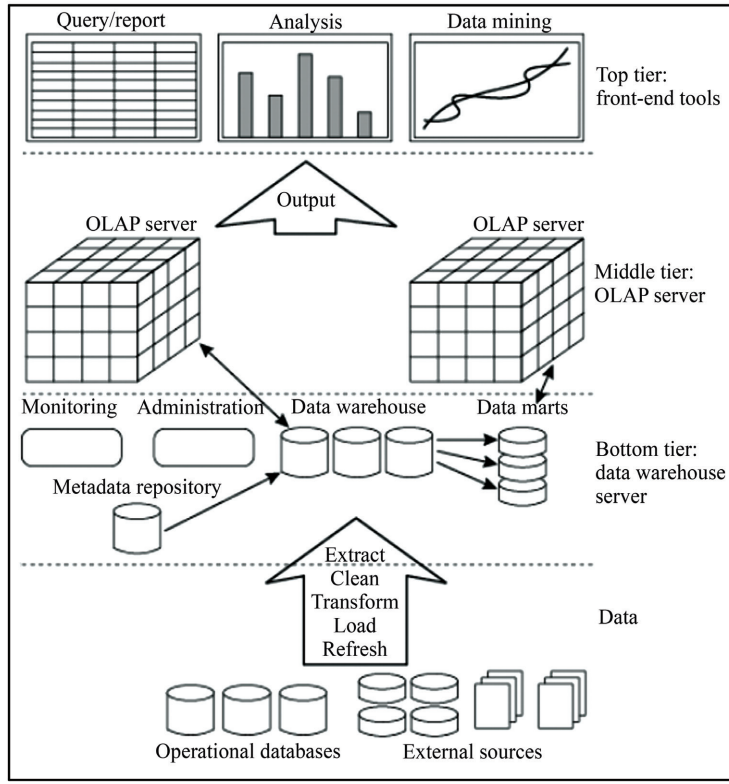
The following schema shows the relationship between data warehousing and OLAP.



(Eltabakh, 2018).

As we can see, a data warehouse can make use of an OLAP server to store the data and run queries on the data. Data from various sources is integrated and stored in the warehouse. The warehouse loads this data into the OLAP and then along with the help of the metadata, queries can be executed on this data to generate reports via client tools, perform data mining, obtain a global view of the data, etc.

The following diagram shows the working of OLAP in a tiered system along with data warehouses:



(Eltabakh, 2018).

At the very bottom layer, we have the data layer where we collect the data from several sources such as operational databases, internal sources and external sources. This data is then transformed with the help of ETL tools and then fed to the data warehouse that is made up of one or several data marts. This is the bottom tier. The next tier consists of the OLAP server that stores the data from the warehouse.

Finally, we have the top tier which is the visualization or presentation layer where the end users can query the server to obtain to information in the form of reports, charts, etc.

1.8. DATA MINING TECHNIQUES

Although OLAP provides the user with a detailed view of the data and what the current state of the data is, it is not capable of predicting what could happen in the future or even why an event is taking place right now. This is where Data mining comes in.

Data mining combines the technique of discovery of data with another technique, namely prediction techniques.

Data mining proposes several methods and techniques that are used to mine useful information from large datasets.

The importance of data mining in various fields has led to the emergence of many implementation techniques and algorithms. The following table shows some of the existing techniques and the associated algorithms.

Technique	**Algorithms**
Classification	ZeroR, OneR, Naive Bayesian, Decision Tree
Regression	Multiple Linear Regression, K Nearest Neighbors, Artificial Neural Network, Support Vector Machine.
Clustering	Agglomerative, Divisive, K Means, Self-Organizing Map
Association Rules	Apriori, SETM, AIS, AprioriTid, AprioriHybrid

Source: Sayad (2018).

We shall now see the data mining techniques in brief.

Some of the common techniques or algorithms are given in the following subsections.

1.8.1. Link Analysis Model

The link analysis model serves to find relationships among individual items rather than characterizing the whole group.

This model implements several mechanisms to find relationships among items. A few of them are: Association Rules, Sequential Pattern detection, Negative Association, Matching Time Sequence Discovery, etc.

1.8.1.1. Association Rules

Association rule mining focuses on discovering frequent or recurrent co-occurring and repeating associations among a collection or a group of items. It is occasionally also called as or known as "Market Basket Analysis," as this was the initial area of application of association rules mining. The principal goal is to look for associations between items that are occurring together more often than expected from a completely random sampling test of all the various possibilities.

Association rule refers to a method meant to locate frequently occurring patterns, associations between items, correlations, or causal structures from data sets found in various kinds of databases such as relational databases,

transactional databases, and other forms of data repositories (Dhanalakshmi and Porkodi, 2017).

Provided a set of items of transactions, this method serves to find rules that help predict the occurrence of a particular item on the basis of other items present in the set of transactions. Association rules mainly consist of methods that help identify interesting, intriguing relationships between different parameters/variables in a database or a data set. This particular data mining technique can play a key role in identifying unique hidden sequences within the items in the data.

The problem statement is defined as follows:

Consider a set of items T, the goal is to find all rules on the item sets in the form x->y such that when x occurs, the event y also occurs with a certain probability.

The general form of an association rule is:

x1, x2, x3, …, xn -> y1, y2, y3, …, ym

This means that when the left-hand side occurs, the right-hand side occurs as well, but with a certain probability. This means that when an event x occurs, there is a likelihood of the event y occurring.

We use the following measures for deriving a particular association rule:

- Support: This value indicates the percentage of support for given rule. This parameter defines the number of occurrences (x) of the rule in a list of n items. It is defined by the variable alpha. For an association rule (A-> B) the support is defined as a union of A and B (AUB) divided by the total number of transactions (T).

  Support (A->B) = (A U B)/ (T)

- Confidence: This parameter is the number of transactions where the union (AUB) appears divided by the number of transactions where A appears

Confidence (A->B) = Support (A->B)/ Support (A)

A very simple and common example that explains the association rule principle is the following:

What is the probability of a customer purchasing milk in the same transaction that he buys bread?

This question can be reformed to ask – If the customer buys bread today at the grocery store, will he also buy milk with it?

Consider the following list of transactions:

Transaction Number	Items Purchased
T1	Bread, Milk, Jam
T2	Bread, Butter, Jam
T3	Bread, Butter, Milk, Jam, Cookies
T4	Bread, Milk, Butter, Jam
T5	Milk, Butter, Cookies

From the above list, we can define the following rules:

a. Bread -> Milk

Support of rule is equal to support of (Bread, Milk) which is 3 out of 5 transactions which is 60%.

The confidence of this rule is equal to the support of (Bread, Milk) divided by the support of Bread which is equal to 75%

This statement means that we can say with 75% confidence that if a customer purchases bread, he will also purchase milk.

b. (Bread, Butter) ->Jam

Support of this rule is equal to the support of (Bread, Butter), Jam) which is 3 out of 5 which is equal to 60%.

The confidence of this rule is equal to the support of ((Bread, Butter), Jam) divided by the support of (Bread, Butter) which is 100%.

This statement infers that we can say with 100% confidence that if a customer buys Bread and Butter, he will definitely buy Jam as well.

Let us take another simple example of transactions as shown in the table below:

Transaction Number	Items Purchased
T1	Beer, Diapers
T2	Beer, Milk, Diapers
T3	Diapers, Milk, Bread
T4	Beer, Diapers, Bread

We derive the following rule:

Diapers -> Beer

The support of this rule is 75%.

The confidence of this rule is 75%.

1.8.1.2. Sequential Patterns

The detection of sequences or sequential patterns is a little similar to association rule mining. On numerous occasions, in case of long-term data that has been stored, the detection of sequences is a useful technique to identify patterns, trends or regularly occurring similar events in the data. It is, in simpler words the discovery of frequent subsequences in a group/ collection of sequences, where each sequence represents events that occur at particular times.

The main job of sequential pattern mining is the analysis of sequential data so as to discover some interesting sequential patterns**.** If we try to explain this in a more precise manner, this technique consists of the discovery of rare and intriguing sub-sequences in a group/set of sequences, where several factors such as its frequency of occurrence, length of the subsequence, the profit that could be gained from mining this particular subsequence, etc. can be used to determine the degree to which the sequence discovered is interesting.

This technique of mining of patterns has a wide range of applications in real life as most of the real-time data is generally encoded in the form of a sequence of separate symbols in many online fields such market-based analysis, webpage analysis, e-learning, bioinformatics, etc.

A typical example is the analysis of customer data over the years. The items purchased frequently by the customer at specific periods during the year can be discovered. When the customer returns to purchase more items, this mined information can be used to automatically suggest some specific items to the customer. These suggestions are based on the frequency and past purchase history of the customer.

Consider the following table that stores the customer purchase information in the form of sequences as follows:

Sequence ID	**Sequence**
1	({a,b}, {d}, {e, f}, {f})
2	({a,d}, {d}, {b}, {a, b, d})
3	({a}, {b}, {e, f}, {f})
4	({b}, {e, f})

The above four sequences are stored in the database. Each *sequence* in the above table represents the items purchased by a particular customer at different times over a long-term. It is an ordered list of set of items that were

purchased together by the client. For example, in the above table, the first sequence states that the customer bought the items ***a*** and ***b*** together, then went on to purchase the item ***d***, then purchased items ***e*** and ***f*** together, and lastly purchased the item ***f***.

In order to perform mining of sequential patterns, we need two values namely a database with stored sequences and a parameter called as the minimum support threshold. This threshold parameter signifies the minimum number of sequences in which a particular pattern must appear or occur to be considered as a frequent sequence.

For instance, if the value of this parameter is set to 2 sequences, the sequential pattern-mining task consists of finding all subsequences that occur in at least 2 of the database sequences.

For the example shown above, several subsequences can be found for the threshold level 2.

Some of these sequences are shown below; where the total count of sequences containing each sequence that was mined (support) is shown as well:

Sequential Pattern	**Support**
({a})	3
({a}, {e, f})	2
({a}, {f})	2
({b})	4
({b}, {e, f})	3
({b}, {f})	3
({a}, {d})	2
({b}, {d})	2
({a}, {e})	2
({b}, {e})	3
({a, b})	2
({d})	2
({e, f})	3
({a}, {b}, {e})	3
({f})	3

For example, the patterns {a} and {b} are quite frequent and have a support of 4 and 3 sequences respectively. These patterns appear in 4 and 3 sequences of the input sequences respectively. The pattern {a} appears in

the sequences 1, 2 and 3, whereas the pattern {b} appears in sequences 1, 2, 3 and 4. These patterns are interesting as they represent some behavior common to several customers.

Although, this particular example extracts frequent sequences on a very small scale, actual sequence extraction is done on thousands of sequences stored in databases. A popular application of this type of mining is text mining where we look for words that most frequently used in the text. In this case, a group of sentences can be seen as a sequence and repeated words that are used in the texts can be mined from this set so as to predict the most used words in texts.

1.8.1.3. Time Series Data Mining

Apart from being used on sequences, the technique of sequential mining of patterns can be applied to time series as well. This is basically data that is related to time. This technique serves to reveal unusual patterns that are hidden within the time series data and maybe predict time series events.

A common example of time series is stock data.

Time series usually represent numeric data and hence this data needs to be processed before mining. In order to mine a time series, it is first converted to a set of sequences by discretizing the series by converting the numbers into symbols. After this, the sequence pattern mining techniques are applied to the time series data to discover interesting sequences in the data.

1.8.1.4. Negative Association Rules

Association rules that we have seen previously consider large sets of data and serves to find associations among different items. In case of negative association rules, the same set of items is considered, but in addition, the absence of items from the transactions is also considered as well. These negative rules of association are extremely prudent in case of market-based analysis in order to identify items that conflict with one another and also find items that complement each other.

For instance, when we mine the following information: ‘Customers who buy Bread also buy Milk.’ here we develop a positive link between the two items.

But, a negative association is the absence of something. An example could be,

‘Customers that buy Corona Beer do not buy wine.’

Another example could be:

'Customers that buy Skimmed Milk do not buy fresh cream.'

Hence, negative association rules do not only provide positive links and relations, but also negative relationships between items that can help modify the marketing strategies so as to boost the sales of products.

The procedure followed is first the positive associations are determined based on the use of the support and confidence parameters. Based on these associations, rules are determined. Then, the statistical correlation between the items that form the rule is determined. If this value is negative it indicates a negative association between the two items.

Generalized Negative Rule Association is defined as a rule that contains the negation of an item. This means that a negative association rule is a rule for which its antecedent or its consequent can be formed by a conjunction of presence or absence (Antonie et al., 2014).

An example of this type of association is X ^ ~ Y ^ ~ Z -> A ^ ~B.

This means: (X) AND (NEGATION OF Y) AND (NEGATION OF Z) implies (A) AND (NEGATION OF B).

1.8.2. Predictive Modeling

This model is also called as the supervised learning model where we make use of observations so as to predict to the outcomes. Predictive modeling makes of stored data, algorithms and machine learning so as to detect the chance/likelihood of a possible event to occur in the future based on past data. The interest behind this type of modeling is to use the existing knowledge of what has already happened to provide the response to the question what will happen? It provides a whole view of what is going on and what will happen in the future.

Some of the classic mechanisms used for predictive analysis are decision trees, regression analysis, neural networks and genetic algorithms. We shall now proceed to look at a few predictive modeling techniques.

1.8.2.1. Classification

The classification technique consists of the predicting a particular outcome based on given inputs ("How Classification Helps Make Big Data Understandable," 2018). Classification is a data mining technique which has the objective to assign an object into one of several categories or classes depending on its other known classes. An algorithm is used which processes

a set containing the attributes and the respective outcome which is called a goal or a predication attribute. The goal is to discover links and associations between the attributes that can help predict the outcome.

We construct "classifiers" that can be later applied to yet unseen data to be able to categorize the data in groups or classes.

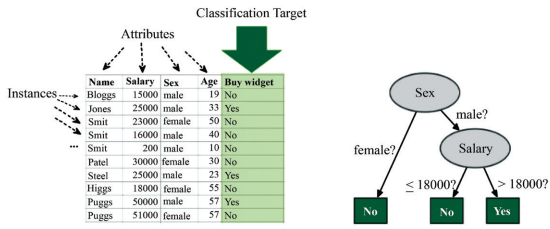
The class that is to be predicted is called the target class. Classification in data mining is similar to normal classification process that we do in our daily lives. An example of classification is when teachers classify their students' educational level according to their attendance, concentration in class, assignments results, etc. We can consider that those attributes are the explanatory attributes that help the teacher predict the educational level, which is the target attribute (Hämäläinen et al., 2010).

More formally, the classification predicts the value of a target class which is of categorical type from other categorical or numerical classes using a classifier. A categorical class is one which has a number of fixed discrete values (i.e.: yes or no, adult or young, weekly or monthly or daily, etc.) while the numerical value is one that can take any value in a finite or infinite range (i.e., age, height, volume, etc.) (Sayas, 2015).

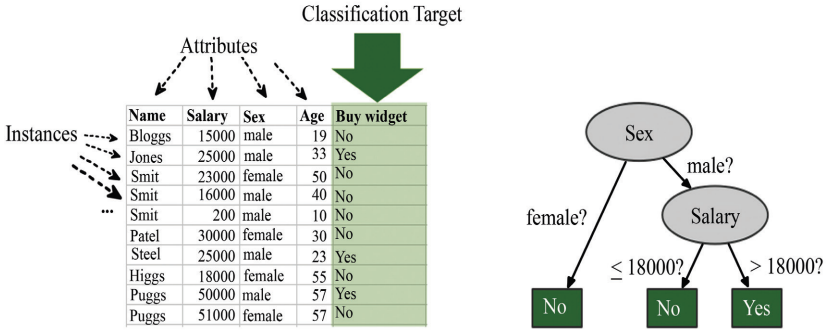
The final target is the creation of a set of classification rules that will answer a question and then maybe make a decision, or predict the behavior (Sayas, 2015).

First, let us understand what a 'classifier' is.

1.8.2.1.1. Classifier



The first step is the learning step, where a portion of the available data consisting of objects with known values for all classes is treated to build the classifier; this set of data is called the training set ("Data Mining Classification & Prediction," 2018). The following figure explains how the classifier is built and a possible outcome of rules that would be used to predict the target class.



In the above figure, the target class is the "Buy widget" column (e.g.: that it is of our interest to predict). By examining the other classes, the rule on the right of the figure is derived. The rule indicates that the person is expected to buy the widget if he is a male with a salary greater than 18,000.

1.8.2.1.2. Testing the Classifier

Another set of data known as the test set is used to test the accuracy of the classifier. This data also provides values for all classes including the target class, but here the classifier uses the explanatory attributes to predict the target class, then the predicted value is compared to the actual value. According to the percentage of correct predictions the classifier can be accepted as satisfactory or rejected ("Data Mining Classification & Prediction," 2018). To formally express the prediction power of a classifier, a confusion matrix (Sayad, 2018) is built from the true and false predictions. The size of the matrix is MxM where M is the number of values of the target class. The following figure shows a confusion matrix for a target class of two values: positive or negative.

Confusion Matrix		**Target**			
		Positive	Negative		
Model	Positive	a	b	*Positive Predictive Value*	a/(a+b)
	Negative	c	d	*Negative Predictive Value*	d/(c+d)
		Sensitivity	*Specificity*	**Accuracy** = (a+d)/(a+b+c+d)	
		a/(a+c)	d/(b+d)		

(Sayad, 2018)

The model is the classifier and the Target is the target class. a is the number of positive values predicted to be positive, b is the number of negative values predicted to be positive, c is the negative values predicted to be positive and d is the negative values predicted to be negative. We are interested in the

Accuracy measure shown in the table to determine the classifier's prediction power.

In order to begin the process of classification, we develop a set of training data containing some attributes and also the likely outcomes. The responsibility of the classification technique is to analyze the set of attributes and then arrive at a conclusion.

Consider the following training set:

Age	Weight	Chances of Diabetes
65	150	Yes
25	75	No
77	130	Yes

Here, we have 2 predictor columns (Age and Weight) that determine the value of the predictor attribute (Chance of Diabetes). In case of a training set, the value of the predictor attribute is known. The algorithm of classification, then tries to understand how the value of the predictor attribute was attained. It determines the relationship between the predictors and the decision taken. The algorithm then proceeds to develop a set of rules of prediction, which are usually nested if then statements.

So, based on the training set, a rule can be developed as follows:

IF AGE > = 65 and WEIGHT > 130 THEN Chance of Diabetes = YES

Another rule that can be developed is as follows:

IF AGE <=25 and WEIGHT < 76 THEN CHANCE OF DIABETES = NO

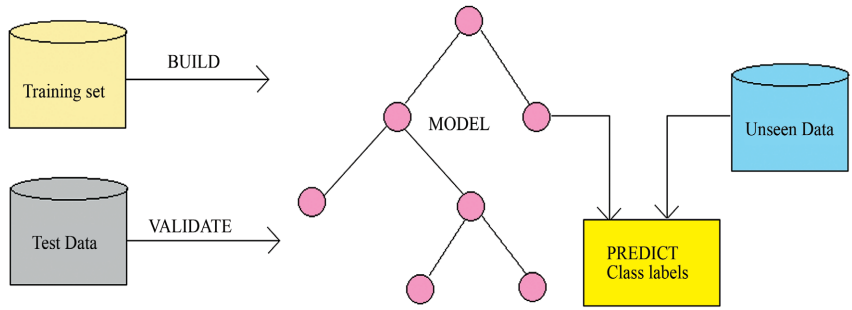
Once the rules are developed, the algorithm is given another set to analyze called as the 'prediction set.' This set is similar to the training set except it lacks the prediction attribute or the decision attribute.

An example of a prediction set is as follows:

Age	Weight	Chances of Diabetes
70	140	
27	79	
80	130	

Here, the predictor columns or data (age and weight) help figure out and estimate the aptness of the classification rules. The rules are then updated repeatedly until the predictions are considered to be correct, effective and useful.

The process of classification can be shown as follows:



1.8.2.1.3. Challenges in Classification Analysis

a) **Data Cleaning**: Due to the large sums of data, values can have some errors and invalid values known as noise or missing values. The noise can be treated using smoothing techniques and the missing values can be replaced by most occurred value ("Data Mining Classification & Prediction," 2018).

b) **Choosing Training/Test sets:** The available data that is used to build the classifier and test it might not be sufficient. A rule of thumb is to use two-thirds of the data for training and one third for testing. In the case study, we will show that a good classifier can be built by less proportion of data and can have an acceptable representation of the whole data.

1.8.2.2. Regression Analysis

Regression analysis is similar to the classifications analysis technique but instead of a predictive attribute as we have seen, a numerical value is predicted.

Although classification analysis and regression analysis are both used to predict events and outcomes, regression analysis focuses on the prediction of a numeric or continuous value whereas classification simply classifies and assigns the data into different categories. The regression analysis technique carries out the discovery of a model or method or function that maps objects into suitable numerical values.

A number of values such as profit, rates, temperature, distance, interest rate, etc., could be easily predicted using regression analysis.

The working of regression analysis is the same as that of classification analysis. In this model, we have a training process where a set of training data with predictor values are stored. In this particular data set, the target values are known. The algorithm of regression analysis will estimate the value of the target as a function of the predictor columns for each record in the data. The relationship between the predictors and the target is then encapsulated in a model which can be further applied to a different data set where the target value is not known.

A model of this type is usually tested by means of statistical computations that calculate the differences between the predicted value and the value that is expected. The data used to develop a regression model is divided into 2 parts one part is for the initial build of the model and the second part is for testing the validity of the model, just like classification analysis.

1.8.2.2.1. Types of Regression Techniques

Some of the techniques used for regression analysis are as follows:

- ***Linear regression:*** This is one of the oldest regression techniques which is used to estimate the relationship between two variables. This algorithm makes use of a mathematical formula of a line (straight line: ($y = mx + b$).) to determine the relationship between two variables. This conveys that given a simple X–Y graph, the relation between X and Y is a straight line with possibly a few anomalies.
- ***Standard multiple regression****:* This is another technique of regression analysis where all predictor variables are considered at the same time. Here, for example, if we have 3 predictor variables, all of them are taken into account for providing the outcome.
- ***Stepwise multiple regression:*** This algorithm is slightly different from the previous one. This algorithm does a selective predicate selection. It analyses all the predicates and decides which of the predicates are best to use to predict the outcome. A stepwise model is followed where the order of importance (so as to obtain the outcome) of the predictor variables is decided and then a relevant subset of the predicates is selected. At each step, one or more predicates are added or removed so as to develop the model. In this type of regression technique, it is highly likely that all predictors may appear in the final regression model.
- ***Hierarchical regressio*n:** This technique is similar to stepwise

regression where the order of importance of the predictor variables is defined. It is a process that is sequential but the variables are fed to the model in a pre-specified manner in advance. The order of insertion of the variables is not built in. This method is used frequently when the end user has domain or field expertise in creation of the regression model.

- ***Setwise regression:*** This technique or algorithm is very similar to the stepwise multiple regression technique. The principal difference is that a set of variables is analyzed instead of single variables.

Consider the following training set that provides the height of children of ages (814) over the years:

Age	Weight	Sex	Family Ethnicity	Height (cm)
9	30	M	Asian	110
12	45	F	African–American	120
11	40	F	Hispanic	108
8	25	M	Asian	114
15	26	M	White	125
12	36	M	Hispanic	101
14	48	M	Hispanic	128

Here, we have four predictor columns (Age, Weight, Sex, Family, and Ethnicity) that determine the value of the predictor attribute (Height). In case of a training set, the value of the predictor attribute is known. The algorithm of regression, then tries to understand how the value of the predictor attribute was attained. It determines the relationship between the predictors and the decision taken. The algorithm then proceeds to develop a set of rules of prediction, which are usually nested if-then statements.

Based on the training set, rules are developed that predict the value of the predictor attribute (height).

Once the rules are developed, the algorithm is given another set to analyze called as the 'prediction set.' This set is similar to the training set except it lacks the prediction attribute or the decision attribute.

An example of a prediction set is as follows:

Age	Weight	Sex	Family Ethnicity	Height (cm)
10	31	M	Asian	

9	25	F	White	
12	46	F	Asian	
9	52	F	Asian	
13	42	M	African–American	

Here, the predictor columns or data help figure out and estimate the aptness of the regression rules. The rules are then updated repeatedly until the predictions are considered to be correct, effective and useful.

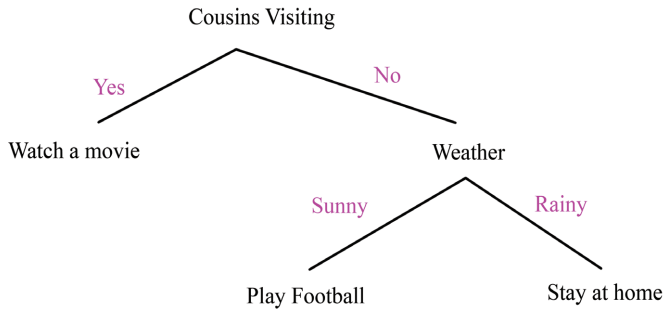
1.8.2.3. Decision Trees

A decision tree is another technique of predictive modeling that makes use of hierarchical trees. It represents a group of rules in a hierarchical manner that enables classification of large groups of items that are not similar into smaller groups of similar items as we go deeper down the tree structure. At each step of the tree, a rule is formed and based on this rule the children are formed and then the leaf nodes, and so on.

Decision trees are used so as to learn from tuples that are labeled. It is similar to a flowchart where each parent node of the tree is a question and the leaf nodes are the answers and/or questions. Each non-leaf node is a test question based on the value of an attribute and each branch is an outcome of this test and each terminal or leaf node is a label that is the result.

Like any tree, the topmost node is the root node or parent node.

A simple example of a decision tree is shown below. Here, the decision that needs to be taken is: "What do we do this Saturday?."



As we can see above, the main root node is the first decider. Based on the answer of this question, the further nodes are built and finally the activity is decided.

This is a very simple example of a decision tree. Real-world decision trees are more complex in nature.

In case of testing against a decision tree, given a tuple, for which the final label (decision at leaf node level) is not yet known, the values of this tuple are first tested against the constructed decision tree and a path is traced from the root node to the leaf node which has the final decision.

This method is a very popular and well-used method of predictive analysis. Additionally, a decision tree easily converts to classification rules. This is because; the use of this model does not require any extensive domain or functional knowledge and people that do not have a background of the domain can use this method too. The visual representation offered by a decision tree is simple and easy to understand and assimilate by people belonging to all domains. Additionally, the decisions can be calculated fast and are generally accurate provided the data at hand is correct.

1.8.2.4. Neural Networks

Neural networks have been around for a long time in the information technology field. They have been in use to mimic the decision-making capabilities and self-learning powers of biological neural networks. A neural network is a system of softwares and structures that basically tries to simulate the working of the human brain. A network like this involves several processors that may also be called as nodes or neurons or elements that work in parallel. Each of the elements has its own database of knowledge along with access to data in its own local memory storage.

By definition, a neural network is a set of connected units (input and output) where each connection has a weight attributed to it (Singh and Chauhan, 2009).

Neural Networks follow a learning process where the weights attributed to each and every connection are adjusted from time to time so as to correctly predict the decision labels as we have seen in case of decision trees. A network is initially 'trained.' In this process, it is fed large quantities of information or data along with rules about the connections between the attributes. Each input is coded and stored. Weights are assigned to the nodes and the output is verified. As initially the output is known, the weights are adjusted so as to arrive at the correct answer. In this iterative manner, the network is built based on the data that is fed to it.

A neural network then computes the data and extracts associations and similarities within the data. They are capable of learning what sets of

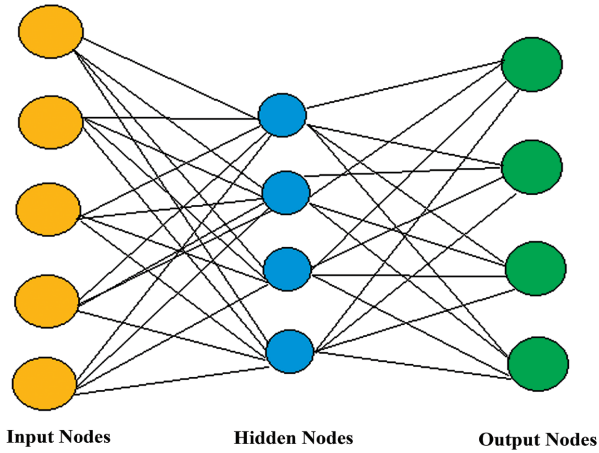
occurrences go together based on observing different patterns over time.

Neural networks are extremely effective in deriving meaning from data that is complex, imprecise and sometimes thought to be random and imperceptible by humans. They are more than capable of detecting patterns and trends that usually go by unnoticed by the human eye. Neural networks have been successfully implemented in several industries where real-world data is in use.

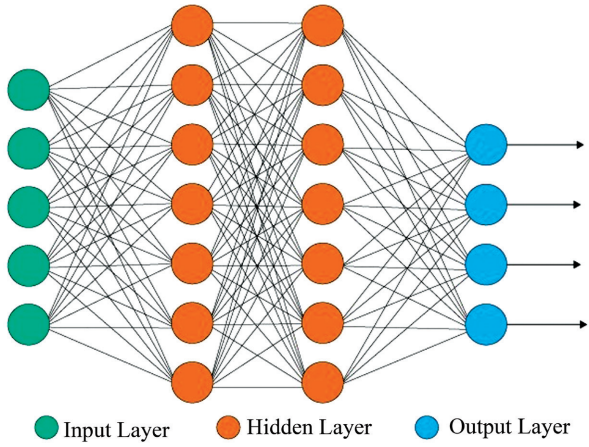
In a neural network, we have input and output nodes which are connected by means of hidden nodes. In total three layers are present: input layer, hidden layer, and output layer. The input layer consists of the nodes that stand for variables present in the data mining problem at hand.

The 2nd layer, the hidden layer contains nodes that are hidden which take the input provided by the input nodes and then pass the inputs on to the next layer. The nodes are passed to the next layer namely the output layer through means of a function that is weighted. The final layer is the output layer which receives weighted nodes from the hidden nodes and this layer contains and represents the variables that are part of the solution along with their weights. The outputs represent the categories that are part of the solution of the problem.

A neural network looks something like the following:



A neural network can have more than one hidden layer as well which is used for a deeper learning process.

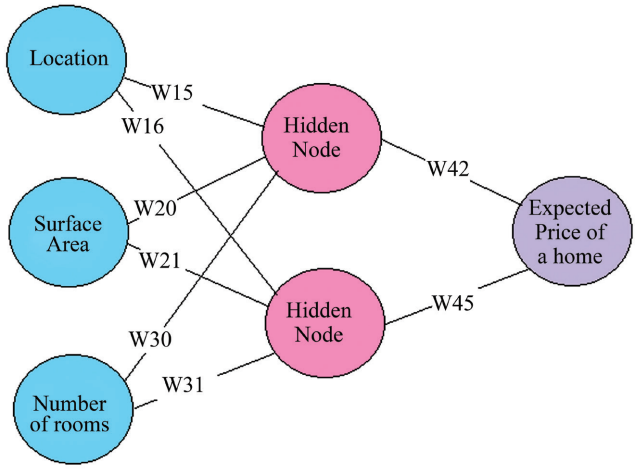


("Neural Networks Projects and Research Topics," 2018).

We shall now take a look at an example of a neural network at work.

Consider the following input parameters that are used to decide the price of a house: Location of the home, the total surface area of the home and the number of rooms.

Below, is a simple neural network that represents the calculation of the expected price of purchasing a home based on its location, surface area and the total number of rooms it possesses.



Although neural networks are very popular, they do have a couple of drawbacks. One of the most known hurdles is that in order to obtain good results the training data needs to be very big so as to form a good network.

Secondly, a neural network is a 'black box.' The logic behind the output is not known to end user. The calculations performed by the hidden nodes are not exposed to the user and the user simply has to rely on the predictions without knowing how the decision was made.

1.8.2.5. Memory-Based Learning

Memory-based learning (MBL) is a method of approximation that dates back to the early 19th century. The concept followed is the same as that of neural networks. Each data is stored in memory and based on this data predictions are made by looking at similar points between the data stored in memory. Once similarities are found, a model is built to fit those findings and then predictions are made.

The foundation of memory-based learning is based on the storage of training data that is analyzed to build a model and then based on the information that was previously learned, a prediction is made.

1.8.3. Segmentation Model

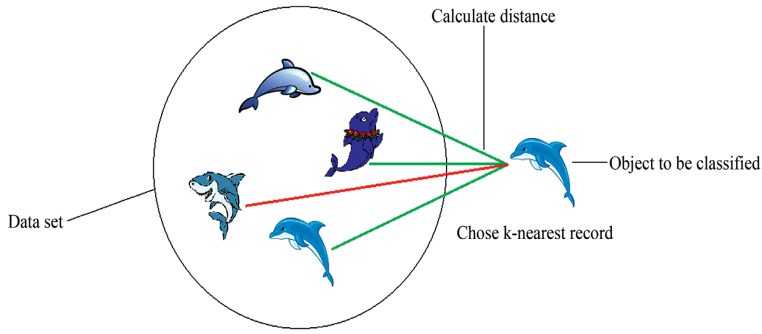
Database segmentation is also called as unsupervised learning where data is partitioned into groups that are similar. Some of the classic mechanisms for data mining using this model are Neural Networks, clustering, Kohonen Maps and K-nearest Neighbors.

1.8.3.1. K-Nearest Neighbour

The K-nearest neighbor algorithm for data is an algorithm that tries to solve the problem at hand based on similar issues encountered in the past. This algorithm is based on the principle that if the issue has already been solved in the past, we can try the same solutions that have worked successfully in the past. This algorithm is a powerful algorithm that is very useful and well employed for pattern detection in many domains. It is a classification algorithm that stores all the available and possible cases and performs the classification operation based on similarity functions. An example of a measure of similarity is possibly the distance calculating function. This algorithm doesn't take in any parameters and is an algorithm that follows the principle of lazy learning.

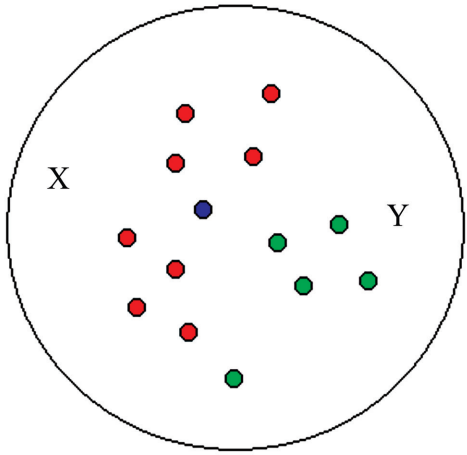
A very simple way to understand this algorithm is the following statement (ethz.ch, 2018):

"Tell me about your neighbors and I will tell you who you are."



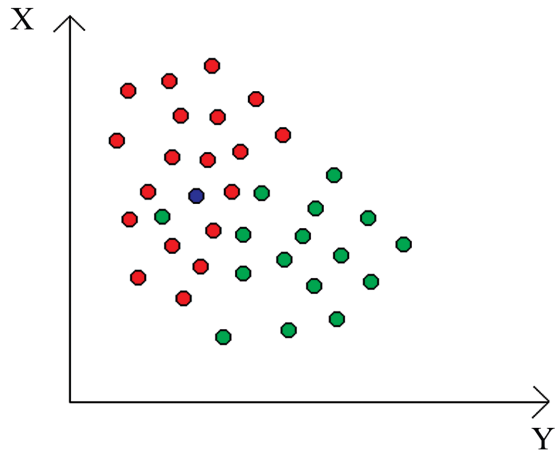
Similar to the above statement, an object is classified based on the presence of its neighbors. Any item that needs to be classified is studied and analyzed in light of its nearest neighbors based on the majority of votes for its neighboring items or objects. After analysis of the object, the object is assigned to the commonest class (classified into) among its K number neighbors based on a distance-based function.

This is shown in the following diagram:



Consider the item in the color blue. We need to classify it based on this algorithm. Here, we have two possible classes – X and Y. So, based on the distance of the blue dot from its neighbors of class X and Y, it is classified to be a member of either class X or class Y. The distances between the blue dot and the objects (dots/items) surrounding this item is measured and then analyzed so as to give a result.

A graph of the items is plotted as follows:

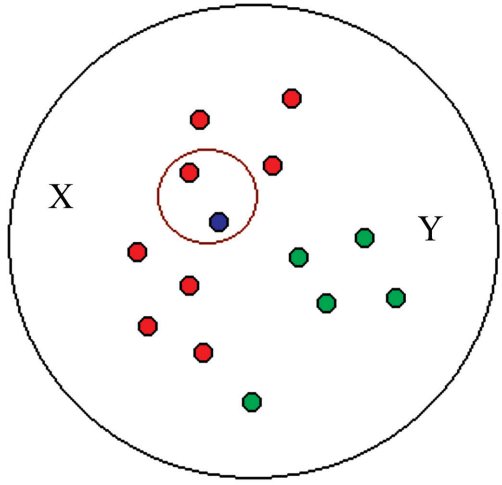


The distance of 'k' items surrounding the item under consideration is measured and based on the answers, the item is classified. The value of k may vary on the data set. Generally, a rule that is followed is that

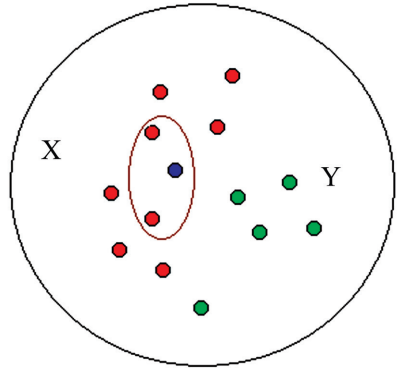
K < SQRT (n)

Where *n* is the total number of objects

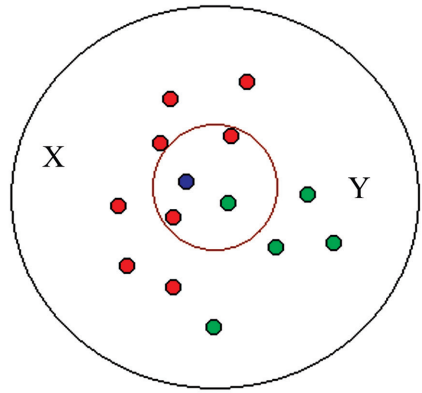
Below are examples of 1 nearest neighbor, 2 nearest neighbors, and 3 nearest neighbors, respectively.



1-nearest neighbour



2-nearest neighbours



3-nearest neighbours

Various formulas are used to measure the distances between the two points. The mainly used formula is the Euclidian distance formula which is as follows:

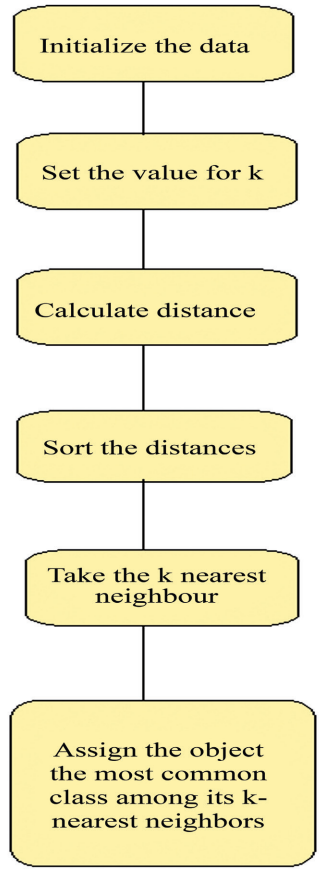
Euclidean distance between two instances ("k-nn – K-Nearest Neighbor Learning Dipanjan Chakraborty," 2018):

$d(x_i, x_j) = \text{sqrt (sum for r = 1 to n } (a_r(x_i) – a_r(x_j))^2)$

Before the algorithm is applied, the following assumptions are made:

- all the items in the data set corresponding to a point in a n-dimensional space.
- each item is represented as a numerical value.
- each of the items in the data set has a class label associated to it.

The steps that are followed for the application of this algorithm on a data set are as follows:



The use of this algorithm is effective as it is simple yet extremely intuitive and also as it can be applied to a wide distribution of data. It provides good accurate results provided the initial data set is large enough so to calculate the distances correctly. The downside to the use of this algorithm is that it takes more time in order to classify new samples or objects. Moreover, the choice of k is always tricky and in order to arrive at accurate results larger data sets are needed.

1.8.3.2. Clustering

Clustering is nothing but the grouping of a set of physical objects into groups/classes of similar objects. Clustering is the process of grouping a set of data objects into several groups or clusters so that objects within this

cluster have high similarity (Berkhin, 2006). Clustering is an activity that we humans have been doing since a long time. Since our childhood, we have learned to spot the difference between trees and dogs or between cats and dogs by clustering them into groups within our subconscious mind.

Given a set of 'n' samples, clustering focuses on finding meaningful partitions of the samples so as to make groups of similar partitions. Each of the subsets that are formed can be considered to be a class or a subset of its own. Clustering is different from classification in the sense that clustering doesn't follow pre-defined rules regarding the constitution of a group or a cluster of objects or items. This method determines to group items similar to one another together, but which may be extremely different from other groups or clusters. A cluster is finally a gather of items that are similar within the boundaries of the same cluster. This statement signifies that the data items are rather akin to each other within the same cluster and they are moderately different or dissimilar or unrelated to the items in other collections/groups or within other clusters.

Classification can further help identify and pinpoint the regions that are dense and sparse so as to further discover correlations and patterns among the data objects present in a cluster. Clustering operation can be used as an effective tool to prepare data for classification by classifying or grouping the data into subsets of similar items.

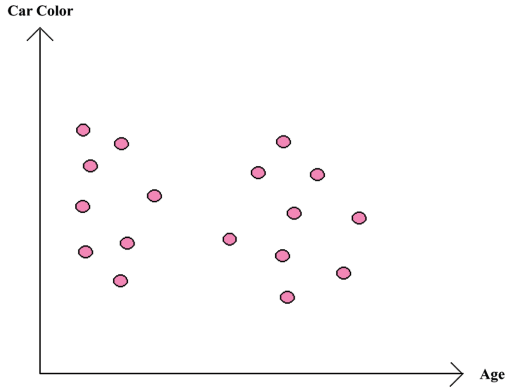
The clustering process involves looking for similarities between objects in the space of data sets. The values of the attributes that describe the objects or items are usually the decision factors along with the measure of distances. It is a process of uncovering collections and clusters in the sample data in such a manner that the level of association between two items is highest if they are a part of the same group and lowest if not.

The clustering algorithm as a tool of data mining has seen many applications in areas of importance such as security, biology, astronomy, business intelligence, etc. It has been used on a large-scale in many domains such as pattern recognition, market analysis, image mining, image processing, etc.

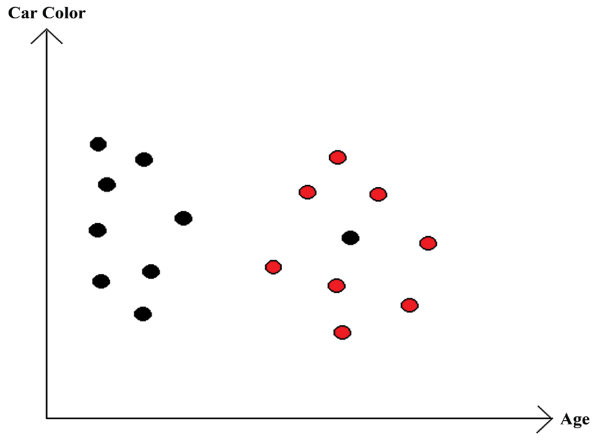
An application of clustering in business is to help businesses detect groups of similar customers and characterize their spending patterns. Based on the items they purchase, they can be divided and put into similar groups. In the field of biology, clustering has been implemented in the process of taxonomy where derivation of specifies into groups is done based on their origins: plant or animal. It is also in use to organize genes into similar categories and

further get a deep insight into the structures of humans. Another application of clustering is in geographical domains where it is used to identify regions that are similar to one another based on several geographical conditions such as land, region and climatic conditions. Further, another market-based implementation of this method is the discovery of types of houses in neighborhoods based on the type of house purchased, its price, sale value, location, etc. A use of clustering in data mining is also the grouping of the large quantities of documents present on the Web that can help increase search times for the documents.

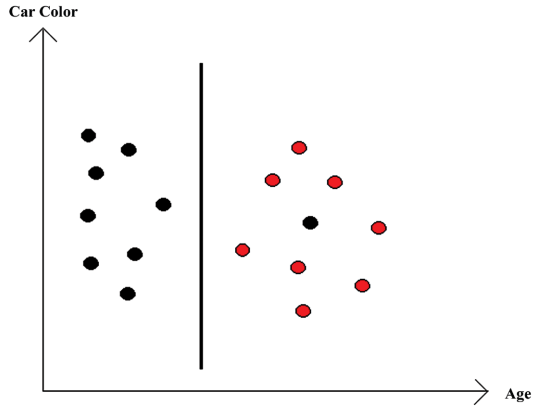
Clustering of items in a sample space is as shown below:



The above graph shows a set of sample points that contain the values of the car color against the age of the customer who purchased it. Now, we add the color values as follows:



As we can see from the graph, the color of car purchased by young customers is mostly black and then as the age increases, the color of the car changes to red. Clustering involves isolating the two color preferences and then separating them into groups. This is shown below:



There are five different methods to implement clustering (Berkhin, 2006). They are discussed in the following subsections.

1.8.3.2.1. Partition-Based Method

In this method, all the sample points in the sample space are divided into k subgroups and a score is assigned to each k partition. Heuristic methods may be applied to enhance the speed of the search algorithm.

1.8.3.2.2. Hierarchical Methods

This method is classified further into two different types namely decisive clustering and agglomerative clustering.

In case of decisive clustering, all the points present in the sample space are assumed to be part of one cluster. At each next step, a split operation is performed to form two clusters that are different from one another. When each point is a cluster this process of splitting ends. Agglomerative clustering follows the reverse of the process of decisive clustering. Here, we begin with each point in the sample space and keep on merging two points until all the points/samples are part of the same cluster.

1.8.3.2.3. Distance-Based Method

In this method of clustering, the notion of vectors is used. Each point in the

sample space is assigned a vector attribute. The distances between the items are calculated based on their vector attributes. The means of calculation are usually Euclidean distance formulas.

1.8.3.2.4. Model-Based Method

The model-based method the notions of normalization and probability are implemented. This method treats each cluster as a normal distribution that is multivariate and then computed the probability of each point belonging to a particular cluster.

1.8.3.2.5. Density-Based Method

This method of clustering depends on density functions applied to the points in the sample data. It calculates the density of the points in neighboring regions so as to discover clusters of different and arbitrary shapes and sizes. The growth of the clusters is done only if sufficient points are found in the neighborhoods which are added to the cluster in a recursive manner.

Clustering is also known as data segmentation because this method divides the data into groups or methods based on their similarities.

As we have seen, classification analysis retrieves salient, relevant information about item under consideration along with its meta-data which is then used to classify different types of data in different classes. This is like clustering as this process performs a segmentation of the data into different partitions or segments called classes. However, unlike clustering, in this process, the knowledge of the different clusters is known beforehand. Hence, in case of classification, clustering can be considered to be a step of pre-processing of data.

1.8.3.3. Kohonen Maps or Self Organizing Maps (SOMs)

Kohonen Maps or Self Organizing Maps are a neural network method that is used for performing clustering or cluster analysis. They are named as Kohonen Maps after their creator Teuvo Kohonen. They are maps that are ordered topologically ("Self-organizing map," 2018). These maps are highly visual maps that help achieve a sophisticated view of data that is not multidimensional, complex and has deep complex relationships between its objects.

The output given by the Kohonen Maps basically highlights the important points of the data and eventually leads to formation of clusters or

groups of similar data. This is an unsupervised learning method that helps give a low dimension or simple (usually 2-d) view of highly complex data. It provides a discretized representation of the data or the input sample data which is called a map. The visualization of the data on a low or small scale is like multidimensional scaling where we reduce the scale of the visible dimensions to render the data more comprehensible.

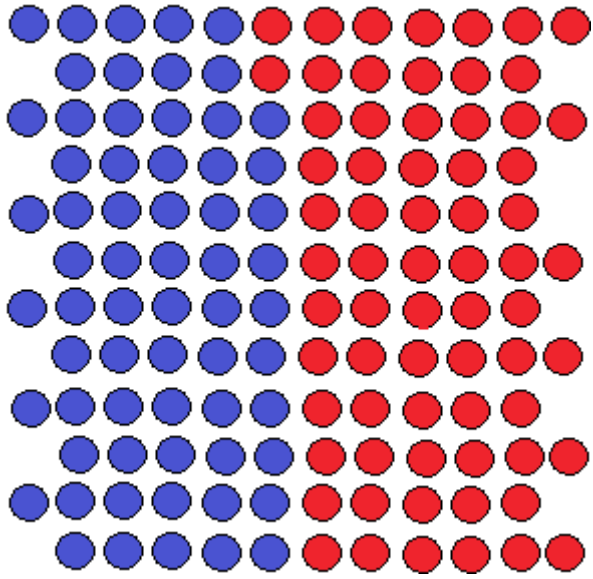
The main intention of using Kohonen Maps is to depict all the sample points initially present in a high dimensional space by different points in a low dimensional target space which is usually (2d or 3d). This organization is done so as to preserve the distance between the points and the relationships between them as much as possible. The use of these maps is practical in case of mappings between points that are not linear.

The method followed by Kohonen Maps is somewhat similar to the k-means method which is a method of vector quantization, originally from signal processing, that is popular for cluster analysis in data mining. *k*-means clustering aims to partition *n* observations into *k* clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster ("Self-organizing map," 2018).

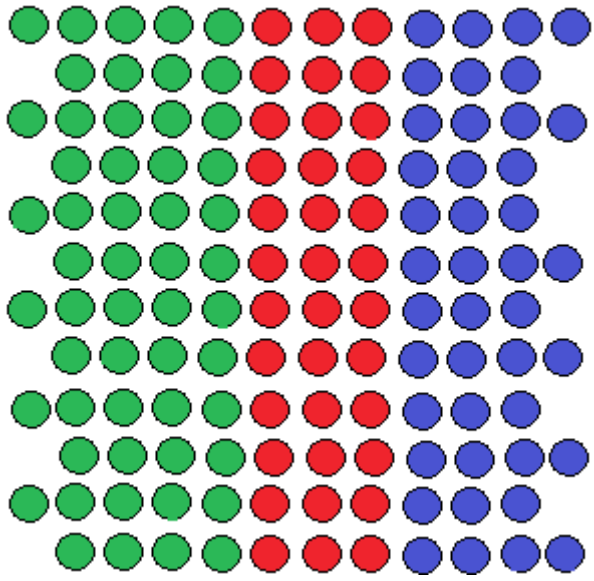
In case of Kohonen Maps, the process of clustering is done by having several groups or units competing for the object at hand. The unit or group that has a weight vector attribute value that is closest to the current object is the winner and it becomes the active group or unit. This process continues with the purpose of moving closer and closer to the input object and so the weights of the unit that wins are adjusted along with the weights of its neighbors. This is continued until a feature map is achieved. The assumption made by this algorithm is that there exists some ordering between the input objects or there is some topology between the input objects and that all the units will ultimately take the form of this structure in space to form a feature map.

The working of Self Organizing maps is similar to the processing of neurons in our brains which is effective in visualizing multidimensional data in 2-D or 3-D form. The approach of a neural network to the method of clustering has powerful academic and theoretic links with the processing system of human brains. This approach has been used for clustering of large quantities of Web documents. In order to render this approach more effective and scalable in case of very large databases as currently the processing times are long and the procedure of using and implementing these maps is complex.

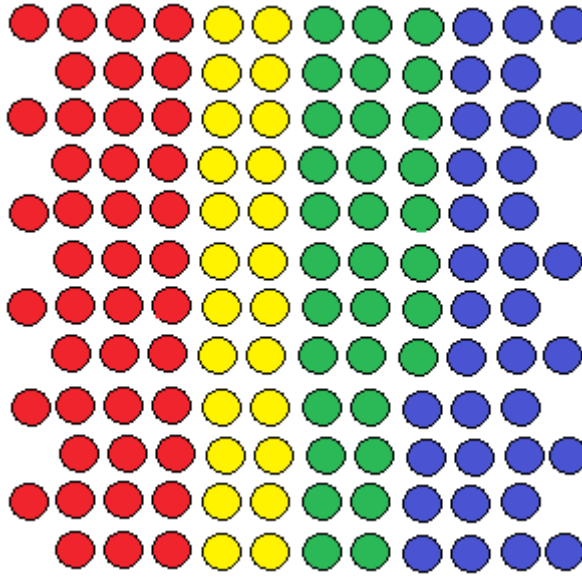
Some examples of clustering using Kohonen maps are shown below:



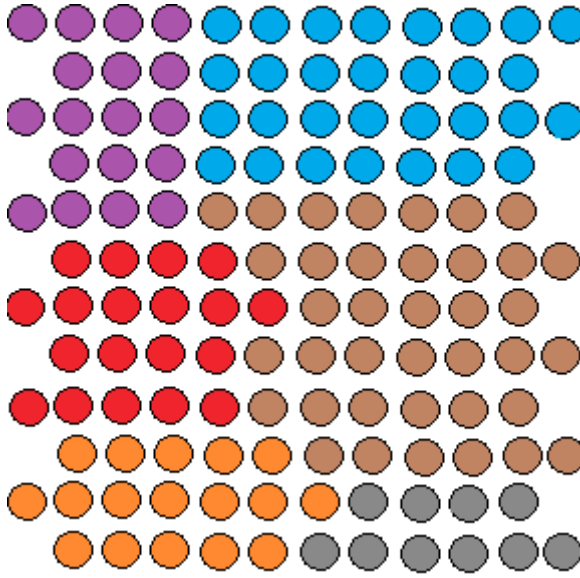
Kohonen Maps with 2 classes



Kohonen Maps with 3 classes



Kohonen Maps with 4 classes



Kohonen Maps with multiple classes

1.8.4. Deviation Detection Model

1.8.4.1. Outlier Detection

In *data mining*, the deviation detection model is the process of identification of items, events or observations which do not conform to an expected pattern or other items in a dataset (Han et al., 2011).

This process is also referred to as outlier detection. This means that a data set is observed to find items in the set that do not match the expected patterns or regular behavior. An outlier is an element that deviates in a quite significant manner than the other items, giving the impression that it was generated by some other mechanism.

This process focuses on detection of anomalies, deviations, exceptions or aberrations from the normal behavior. They can also be called as outliers, noise or novelties as well. The detection of deviations can provide useful information that can be used for further analysis. A deviation or an anomaly is an item or object that is extremely different from the common average of the group or within a combination of data. The items that are deviations are statistically different from the rest of the data and hence they suggest that something out of character has happened and this event needs more scrutiny.

The deviation detection of outlier analysis model is being put to use in many domains such as banking, healthcare, networking, environmental systems, etc. A very common implementation of outlier detection is the detection of fraud in the banking where suspicious behavior and discrepancies in transactions can be flagged. Additionally, in case of networks, outlier detection can help find faults within networks such as sensor networks. The same logic can be applied to the detection of disturbances within the eco-system. Usually, the data that is anomalous and doesn't follow the norm is often excluded from the data set. But, outlier detection can help derive meaning from this data that doesn't follow the norms.

The discovery of outliers can be thought of as good news or bad news depending on what we were looking for. They may depict the events that we are looking for or they may also serve as a basis for finding errors in the sample data set.

Outliers can indicate good news or bad good news in the sense that they represent precisely the opportunities that we are looking for; bad news in that they may well be no more than invalid data. Each outlier needs to be treated differently based on the information that is provided.

For example, X usually goes to the bank on Monday and withdraws some money, after which he does some groceries. But, one Monday, he withdrew some money and then went to watch a movie. This is an event that is an anomaly. Using the same logic on complex data such random wayward events can be detected and analyzed further.

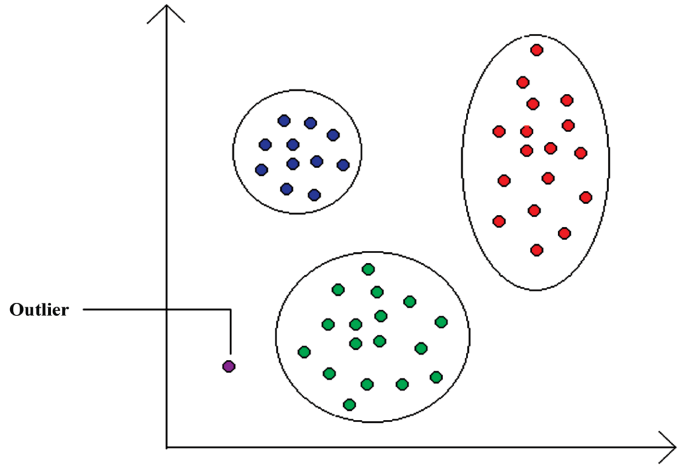
An example of an anomaly is possibly erroneous data that may have been wrongly fed into the database. For example, consider the parameter or attribute age of a person. If the age that is entered is a negative value such as -45, it is clear that the value is erroneous. This erroneous value could be attributed to human error. By the use of outlier detection, such errors could be found very rapidly and then corrected in an effective way. Either this value can be corrected or later dropped from the data set.

Sometimes there are cases of false outliers where the issue is sometimes not directly related to the errors in the data. Sometimes, the data present is correct, but the meta-data has been updated. In this case, the changes are not yet perceptible in the data mining data and this data may present as an outlier, when actually it is not. The correction is simple in such cases; the meta-data is updated and the issue is resolved quickly.

Another example of an outlier is a histogram. Consider a histogram that shows the earnings of a group of people. The histogram shows that students and senior citizens have low incomes. In such a histogram, an outlier may be a person that is currently employed and earns more. This outlier can be considered to be a positive outlier as it is normal that a person that is working has a higher income. This data may appear as an outlier maybe because the data set doesn't consist of sufficient data and contains data about students and senior citizens.

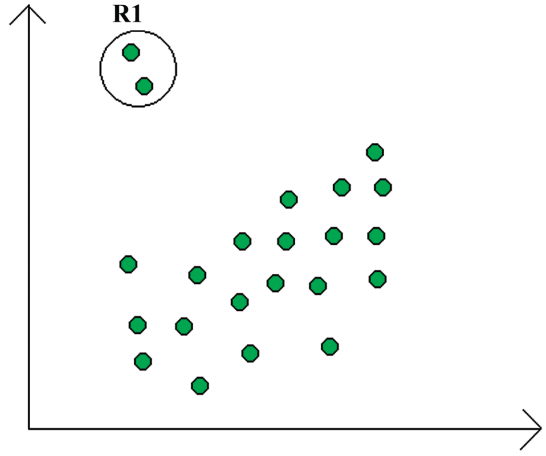
Summarizing the detection of outliers, it depends on how one treats the deviations based on its nature and the information that it gives. The deviation must be analyzed and then treated appropriately.

Some examples of outliers are as follows:



In the above image, the dot that is not present in the circle is an outlier.

Another example of an outlier is as shown below:



In the above image, the objects present within the region R1 are outliers.

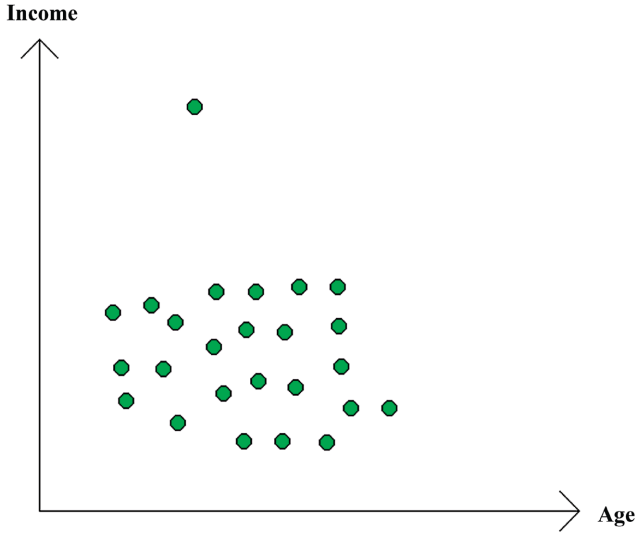
There exist three different types of outliers namely global, contextual and collective. We shall take a brief look at the three types of outliers (Han et al., 2011).

1.8.4.1.1. Global Outliers

An outlier is called a global outlier if this outlier is considerably different from the rest of the data set. These types of outliers are also called as point

anomalies and are very simple to detect. This outlier is completely different from the entire global data set. In terms of computer analogies, a global outlier is sometimes like a global variable in programming which is visible instantly within the scope of the application.

Consider a graph that shows the income of a group of people between the ages of 24–34 years.



In the above graph, the range of the incomes is more or less the same for this age group, with a single exception where the income of one person is extremely high than the global average. This is an example of a global outlier.

1.8.4.1.2. Contextual Outliers

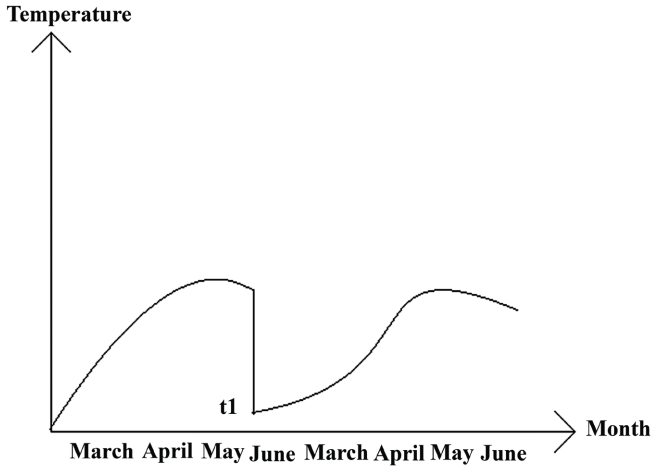
A contextual outlier is an object that deviates in behavior within a specific context of the object. These outliers are subject to conditions related to context and hence are often known as conditional outliers. They depend largely on the current context that is selected. It is important to note that the same value may not be considered as an outlier or an anomaly if we look at it within another context. Hence, as part of the input definition or problem definition, the context needs to be correctly specified so as to find contextual outliers correctly.

Consider the temperature of different cities in a country. For example, the average temperature stays between 2 and 15 degrees in spring. Here, if

the temperature rises to 35 degrees during spring, it is an anomaly related to the context, which is the 'season' in this case.

However, if we consider the same value of 35 degrees and analyze it within a different context such as for the season 'summer' this value is no longer an outlier. Hence, the context plays a key role in the determination and detection of contextual anomalies.

This example is as shown below:



As we can see, the temperature is more or less the same throughout the months, but someday around June, it decreases abruptly. This here is a contextual outlier as this is not the norm for a context 'temperature.'

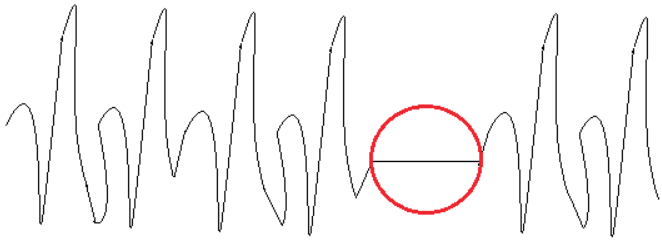
1.8.4.1.3. Collective Outliers

A set of items within the complete data set that is different as compared to the entire data set, it is called as a collective outlier. Here, the values of the data items individually are not anomalous but the value of a collection of data points is a deviation from the entire data set in the global sense. Individually the items may not necessarily be anomalies, but as a collective group, their occurrence in a collective manner constitutes an outlier.

For instance, consider a small business of a bakery. Daily, this bakery delivers around 50 orders. Sometimes, delays cannot be foreseen and the order is delivered late or not delivered at all. Statistically, delays and interruptions are considered to be a part of the business, although they can be treated as outliers. But if on one day, all 50 orders are not delivered, it is considered to be a collective outlier. In this case, if we look at an individual order that

is not delivered, it can be classified as a global outlier or may not at all be an outlier as the chances of this happening (1 order not being delivered) are possible. But, if we look at the whole group, we see an outlier as all of the orders were not delivered on that particular day.

An example of a collective outlier is shown below by means of an echocardiogram:



In the above echocardiogram, the area in red is a collective anomaly. Although the individual sample points are not considered to be deviant, the entire group of sample points signifies a possible deviation from the norm in a collective manner.

1.8.4.2. Genetic Algorithms

Genetic Algorithms are based on Darwin's Theory of Evolution. These algorithms can be used for both classifications as well as outlier detection. The basis of genetic algorithms is the same as Darwin's theory – 'Survival of the fittest' (Larose, 2005).

These algorithms strive to duplicate the natural selection process that has been proven by Darwin and apply the same to business problems. The theory of 'survival of the fittest' is based on the natural evolution process of mankind where living organisms were first created by means of reproduction, they evolved and mutated and only those offsprings who were fit enough managed to survive and evolve in the environment. In reality, in nature, the environmental constraints and stress factors force the different types of species and the different types of individuals within the species to compete for resources and struggle to survive so as to produce the offspring that is the fittest.

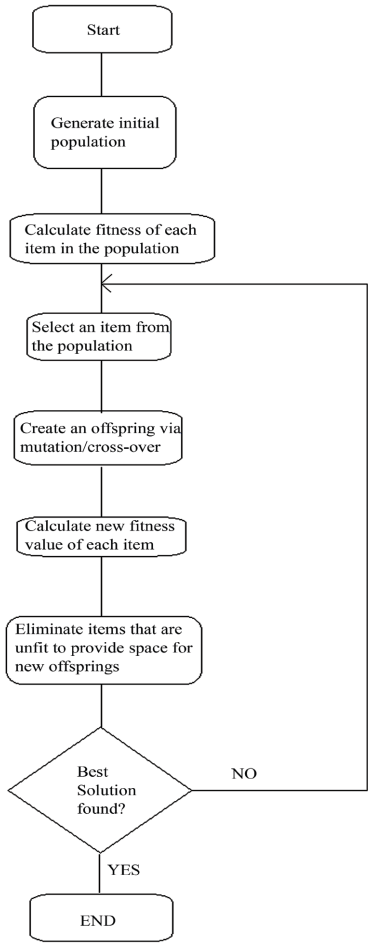
When the same theory is applied to business processes or problems, we provide several possible solutions for the problem at hand, mutate the solutions, and finally, only the strongest solution survives the set of tests we put them through.

The generic process or steps followed by a genetic algorithm is as follows (Larose, 2005):

- **Data Selection:** The population is chosen which consists of n number of solutions for a given problem at hand.
- **Fitness Calculation:** The fitness of each solution is calculated by means of a fitness function so as to determine the initial strength of each solution.
- **Cross Over:** In this step, the next generation of solutions is created by means of a crossover function. Crossover functions involving swapping parts or rules of two different solutions to form a new solution. The crossover function is applied to all of the n solutions.
- **Mutation:** This step involves the evolution of the solutions by the application of a mutation function. Mutation often involves inversion of the solution's rules.
- **Discard Weak Solutions:** The fitness of the new solution is recalculated and the weak solutions are eliminated based on the threshold that is set for the fitness value.

Based on the resulting solutions obtained, either the process is stopped by the selection of the strongest or a new iteration of the process is done where another generation is created and tested again for its fitness.

The algorithm can be explained in a simpler form via a flowchart as follows:



We shall now see an example of the application of a genetic algorithm on five solutions.

Consider a marketing strategy for a grocery store. If the store wants to decide where to invest their money, they perhaps would like to classify their customers based on their purchasing behavior. The same concept can be used to detect outliers such as products that are never sold.

Each item purchased has a value assigned to it and maybe a group of items that are purchased together have higher value.

Consider the following five solutions, with random values assigned to the products.

20	15
25	20
35	30
10	05

Solution 1

34	29
21	17
36	30
39	33

Solution 2

35	27
43	38
28	22
37	30

Solution 3

45	41
51	47
37	29
22	13

Solution 4

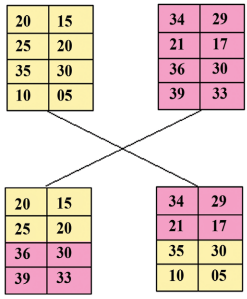
22	16
33	29
51	43
42	37

Solution 5

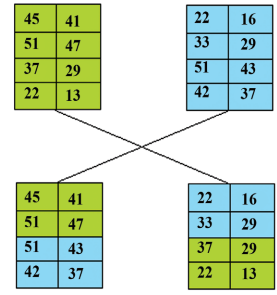
The above sample solution set shows a combination of the values assigned to groups of products that are to be grouped together for promotional purposes.

In this solution, we first calculate the fitness of each solution based on a fitness function. In this case, the fitness can be calculated based on the sale values of the product combinations.

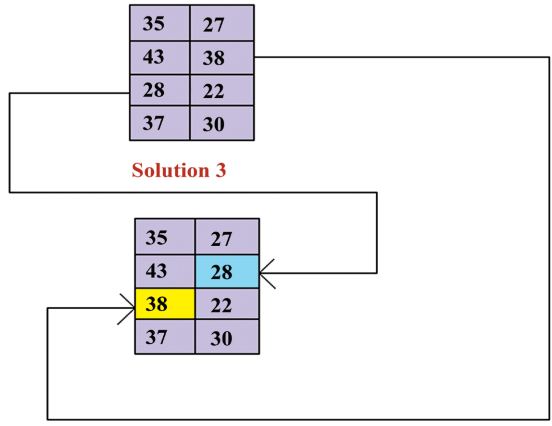
Once the fitness values have been calculated, crossover and mutation functions are performed on the first generation of solutions to produce the 2nd generation. This is shown below:



Cross-over between solutions 1 and 2 to create 2 children

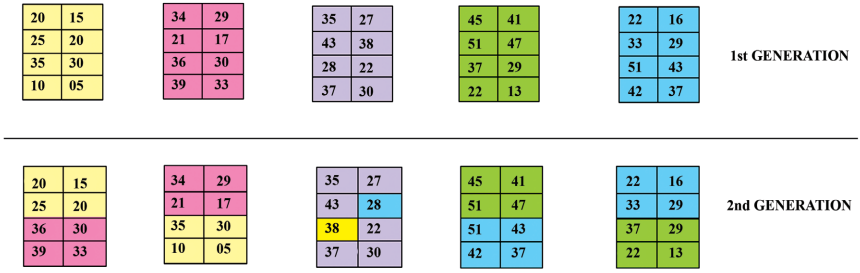


Cross-over between solutions 4 and 5 to create 2 children



Mutation operation by swapping two values

The 2nd generation of evolved mutated solutions is as follows:



1.8.4.3. Frequent Item Set Mining

This technique serves to mine frequently occurring sets of items in a large group of items. This is a common problem that exists in case of market-based applications. The process of identification of groups of items or products, characteristics and similar symptoms, which frequently occur together in the given data set or database is one of the most basic tasks in the field of Data Mining.

Initially, the motivation behind the search for frequent item sets originated from the need to analyze data from supermarkets so as to analyze customer behavior based on their purchase history of the products (Agrawal et al., 1993). The operation of finding frequent sets of products details how often products are purchased together and helps gain an insight into the shopping behavior of the customers. The information obtained from this process

can later be used for creating profiles of customer, developing marketing strategies, etc. so as to boost sales and make more profits.

Consider a set of items I where T is a set of transactions over I items stored in a database D. Each transaction over an item is associated with a transaction id such as each transaction has a unique id.

Here, in the process of the frequent item set mining, we make use of an important measure namely support.

Support: This value indicates the percentage of support for given rule. This parameter defines the number of occurrences (x) of the rule in a list of n items. A particular transaction T is said to support a dataset X if X is a subset of I. The cover of the set X includes all the transactions in the database D that support the transaction T. The support of a set X in D is equal to the total number of number of transactions in the cover of the set X in database D. The frequency percentage of a data set X in the database D is the probability of the occurrence of X in a transaction or in simpler words, the support of X divided by the total number of transactions in the database.

A set X is said to be a frequent set if it is able to support a minimum support threshold that has been decided previously. The frequency F of a transaction is then denoted by the following formula:

F = {X c I / Support (X, D) >= Minimum Support}

We shall take a look at an example of the use of frequent item set mining now.

Consider the following table containing sets of items (I) stored in a database D purchased by customers along with their transaction ids (tid):

Tid	Item Set (I)
10	{milk, bread, butter}
20	{milk, bread, wine, butter}
30	{milk, bread}
40	{bread, butter}

Here, I = {milk, bread, wine, butter}.

The table below shows all the sets in the database D with a minimum support that is equal to 1.

Item Set	**Set of Ids**	**Support**	**Frequency**
{}	{10, 20, 30, 40}	4	100%
{milk}	{10, 20, 30}	3	75%

{bread}	{10, 20, 30, 40, 50}	4	100%
{butter}	{10, 20, 40}	3	75%
{wine}	{20}	1	25%
{milk, bread}	{30}	1	25%
{milk, wine}	{20}	1	25%
{milk, butter}	{10, 20}	2	50%
{bread, butter}	{40}	2	50%
{milk, bread, butter}	{10}	1	25%
{milk, bread, butter, wine}	{20}	1	25%

If we are given a minimum threshold, then this list can be further refined to obtain results that have a support value equal or greater to this threshold.

If we apply a minimum support threshold of 2 to the given data set, we get the following results:

Item Set	**Set of Ids**	**Support**	**Frequency**
{}	{10, 20, 30, 40}	4	100%
{milk}	{10, 20, 30}	3	75%
{bread}	{10, 20, 30, 40, 50}	4	100%
{butter}	{10, 20, 40}	3	75%
{milk, butter}	{10, 20}	2	50%
{bread, butter}	{40}	2	50%

As shown in this table, only those entries with a support that is greater than 2 are displayed.

1.8.4.4. Classification and Regression Tree (CART)

The classification and regression tree analysis method is a model that uses classification as well as regression techniques for analysis of data. It is a tree-based technique that is somewhat different from the traditional methods of data analysis and mining. The cart methodology introduced by Breiman et al. (1984) as an elegant solution for representing data in the form of trees.

The idea behind this method is to build decision trees using partitioning, statistics and regression analysis. This method uses the concept of a learning set which is a set with data and classes that are pre-assigned for all the items within the data set. A recursive portioning algorithm is used to build a decision tree in a methodical manner and with a statistical approach. In this

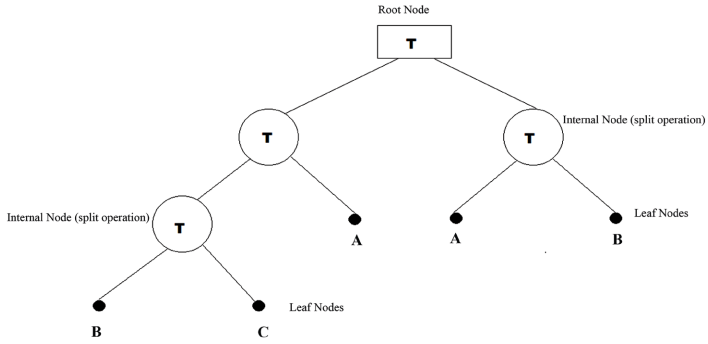
approach, a decision tree is built by either a splitting operation on each node of the tree into two child nodes.

The algorithm starts with a root node and a learning set L on top of the root node. A node can either be a subset or a terminal or a leaf node. A parent or non-terminal node can either be split into two daughter nodes or be left alone as it is now in its rawest form. The process of splitting continues until the resulting leaf or child nodes are deemed to be the 'purest.' Only univariate values for leaf nodes are considered in this algorithm. Each split operation is directly dependent on the value of a predictor variable. The condition (defined by the predictor value) is either satisfied or not satisfied by the value of the node. This means that all the values in the learning set L are at a point or a node where the node under consideration either satisfies the condition or not. It also means that if the condition is satisfied by one parent node, all the subtrees and nodes that fall below this node also satisfy the predicator condition that is applied to this node.

Each node that is the leaf or terminal node is left alone from the splitting process and assigned a class label value. If the terminal node doesn't have a class label or it is unknown (not present in the learning set), it is assigned the class label of its parent node. It is equally possible that there exist several terminal nodes with the same class label.

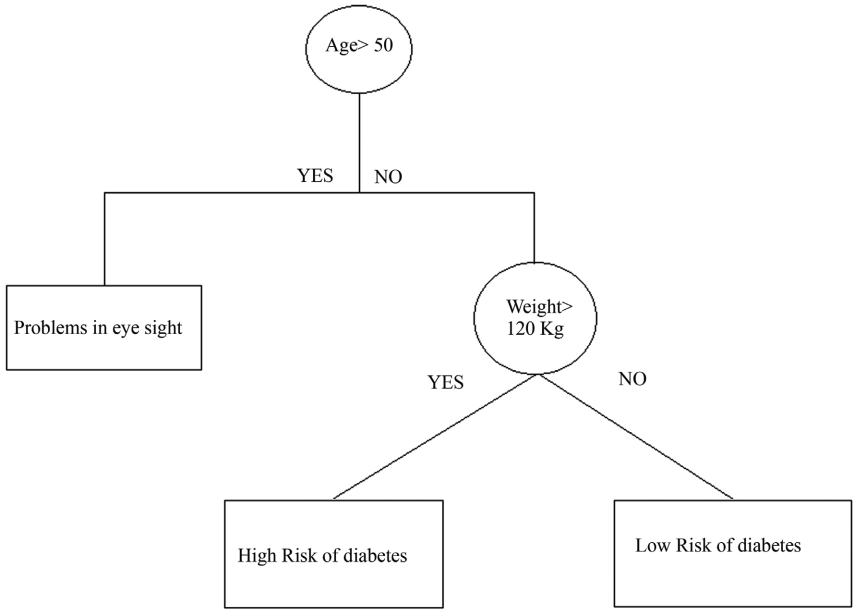
In order to construct a model that has a structure similar to a tree using the process of recursive partitioning, the CART method decides the best possible split of the set L (learning set) based on parameters such as identification of the predictor variable, determination of a rule for splitting, determination of terminal nodes and the allocation of a class to the leaf nodes. The step of assigning the nodes is comparatively easier than the selection of the predicate for splitting so as to create a decision tree.

The tree looks something as follows:



The process of the CART is identical mathematically to a few regression techniques, but it varies a great deal in terms of representation of data which makes it easy to interpret and read the data by people who do not share a statistical background. This particular algorithm combines the complex analysis procedures of statistical regression mechanisms with the simplicity of representation and depiction of data offered by decision trees.

Another graphical representation of a decision tree using the CART algorithm is shown below:



In the above example, a single input value is denoted by each root node. Additionally, this node also has a split point on this variable value. Here, the assumption that is made is that the variable is numeric and it can be split. The terminal nodes of the above tree contain an output variable that is usually a class that makes a prediction.

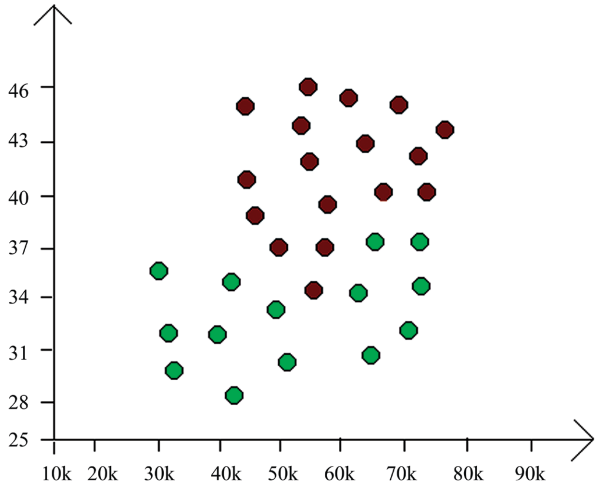
The above example is a simple binary decision tree that is given a dataset with two inputs of age and weight and based on these inputs; it predicts the possible health risks for a person.

We now take a look at another example that will help us better understand the working the CART algorithm.

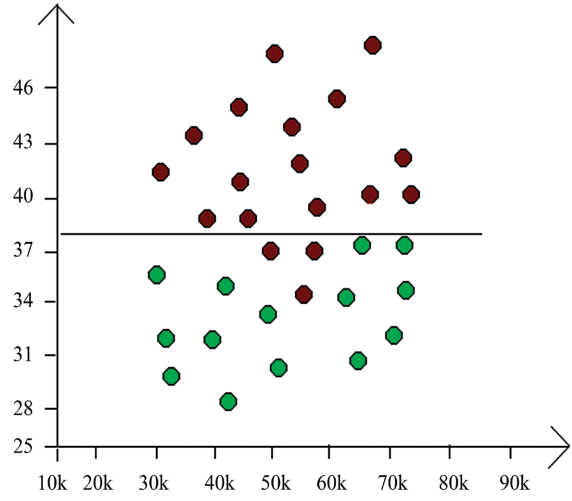
We consider a graph that depicts information regarding the investments done by people based on their age and salary. The range of the ages is

between 25–45 years. The salaries are measured in dollars.

The graph is as follows:



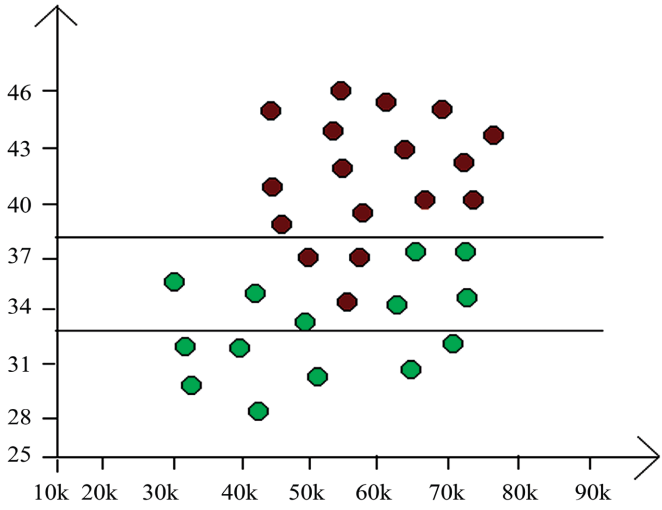
The CART algorithm analyzes the data and performs the function of finding a variable that allows the splitting of the data into homogeneous groups. One logical split of the data is shown below:



The criterion for splitting the data is age in this case. Here, we split the data at the age of 38 and above. The rule that is formed at this point of splitting is as follows:

Age > 38

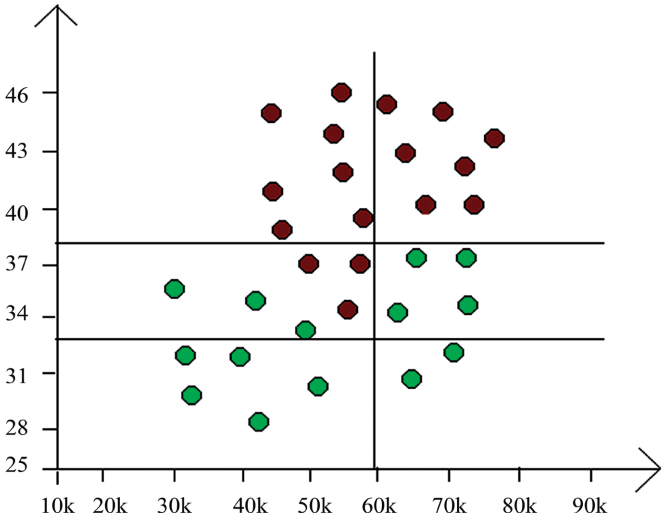
We further split the data at two different age values to further discover more rules.

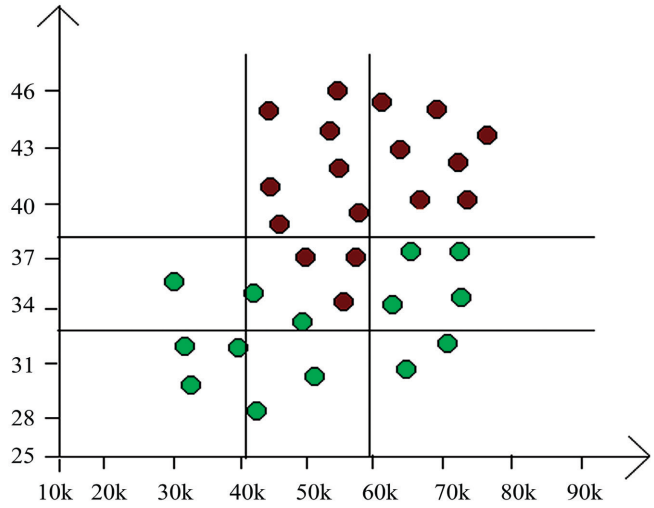


Here, the rule that we find is:

Age <= 33

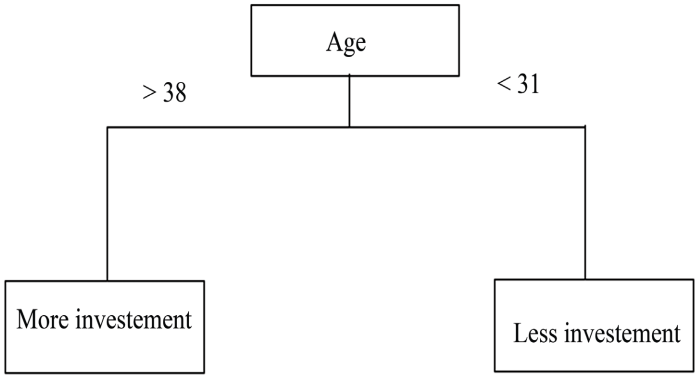
Now, we further split the income to get more homogeneous groups as follows:

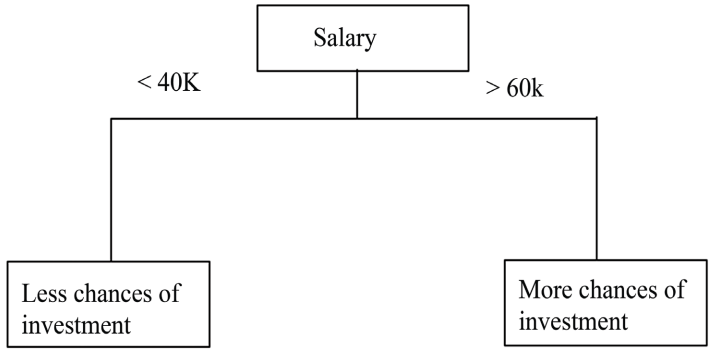


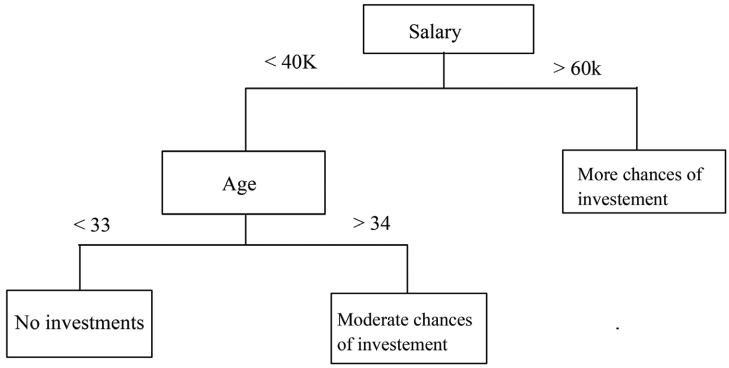


The CART method, keeps on splitting the data until each terminal node has a minimum number of records that are specified by a minimum split criteria. The minimum split criteria is then later used to prune the tree and eliminate subtrees that do not result in homogeneous class values.

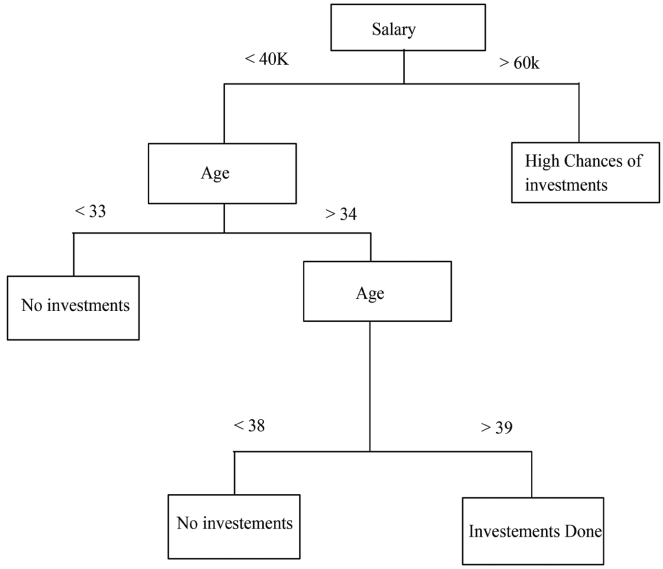
Based on the splitting that we have done, we get a decision tree as follows:







The tree can further be developed in a more detailed manner as follows:

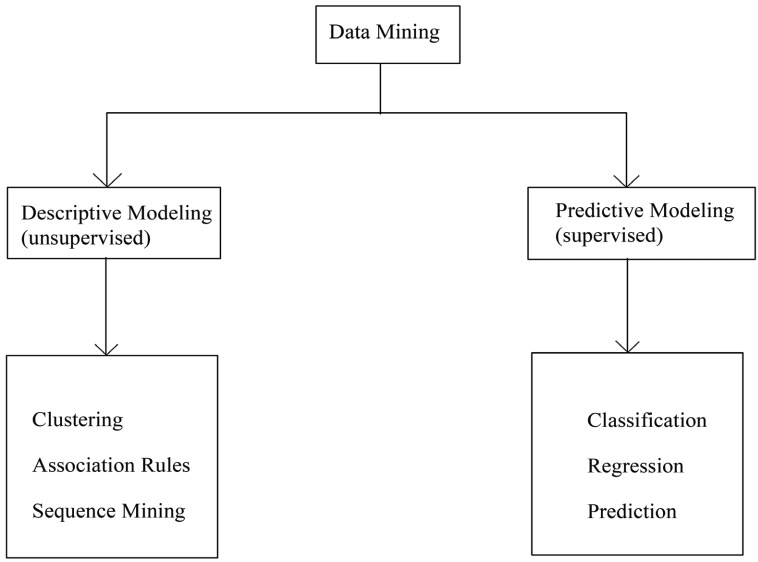


1.8.4.5. Visualization

The technique of visualization in data mining was introduced for more effective presentation of data that if formed or processed. Many studies show that the human brain is remembers more visually than lexically.

The process of visualization converts simple raw data consisting of characters or numbers to a solid image. The image is a static form and is a graphical representation. Visualization includes techniques such as tree map, scatter plot matrix, parallel coordinates, and spatial visualization (Rodge, 2016).

Based on the type of learning, data mining techniques can be classified as follows:



2
CHAPTER

APPLICATIONS OF DATA MINING IN MANAGEMENT

CONTENTS

Data Mining has several applications in the management of many industries where it has been successfully employed to perform managerial tasks and take decisions.

We shall take a look at some of the applications of data mining in different sectors.

2.1. TELECOMMUNICATIONS

The industry of telecommunications was one of the first industries to adopt the strategy of data mining. The interest in data mining arose due to the presence of large quantities of telecommunications data that was not being analyzed to its fullest extent. A lot of data is generated by this industry in the form of call details data, hardware and software related data, customer subscription data, cell data, billing data, etc. As this data quantity is quite extensive, a manual analysis of this data is highly impossible and inefficient. The necessity of handling such large volumes of information led to the creation of automated knowledge-based expert systems that helped analyze this data. Although the automated systems helped consolidate the data, the final decision regarding the information analyzed still rests with 'experts' in the telecommunications domain that can detect the value of the information found, if any. Often, the experts don't have the required in depth knowledge of the entire domain and hence an information bottleneck is introduced in this process.

Hence, the rise of data mining gave this industry an opportunity to better analyze their own data. The data stored by this industry is very interesting in a data-mining context. One of the major concerns was the sheer volume of the telecommunications data. Billions of records are handled and stored by this industry. Secondly, the data that is available is very raw and often it cannot be used directly for processing by a data mining methodology. Data mining can be implemented in various ways in this industry such as failure detection, fraud detection, customer retention, etc. Some of these applications of data mining within the telecommunications industry are discussed in the upcoming sections.

2.1.1. Marketing/Customer Profiling

The telecommunications industry maintains detailed records of the customer data such as general customer data, call detail data, subscription data and so

on. Significantly, call stream data is stored which is then used for billing purposes. This stack of data helps identify the call behavior patterns of a customer. By analyzing the customer call data, profiles of customers can be generated and several marketing strategies can be developed so as to ensure customer loyalty and attract new customers. Additionally, forecasts can be made based on the information recovered with the help of data mining.

One of the well-known applications was done by MCI's promotion named "Friends and Family," which was introduced in the United States in the year 1991. The marketing team performed an analysis of customer data and observed that particular sub-groups within a particular network had more activity than others, thereby revealing an opportunity of adding calling circles. What the company did was that they offered a promotion that reduced calling fees within a particular circle of people. This circle was the 'Friends and Family' circle. This particular promotion emerged originated when market researchers at MCI noticed small subgraphs in the call-graph of network activity – thereby suggesting the possibility of adding entire calling circles rather than the costly approach of adding individual subscribers (Weiss, 2009). This promotion relied mainly on its customers to bring along their own list of family and friends. This particular implementation of data mining where associations within network data was mined was one of the early uses of data mining within the telecommunications industry. This marketing strategy was later terminated around the year 1997.

One important output of data mining is that the companies, with the help of data mining can build solid profiles of customer based on their phone usage and their regular patterns. The profiles created with the help of data mining can then be used for marketing purposes to retain the customer. The profiles can also be used to better understand customers and forecast potential risks in the retention of the customer.

2.1.2. Customer Churn

The issue of *Customer Churn* is a very serious issue that is present in the telecommunications domain. This term means that customers constantly keep leaving one telecommunication provider to switch to another one. This is similar to customer attrition where customer quit their service providers for another. The telecommunications industry faced a major customer churn when companies started offering incentives such as a bonus of 100 dollars to switch telephone carriers. This incentive led to customers changing phone carriers in order to earn the profits. The telecommunication companies

needed to find a way to determine which customers would quit and which ones were more likely to stay. This issue of customers randomly changing phone carriers was increasing losses in revenue as well increased costs in order to attract new potential customers.

Data mining was the perfect solution to analyze the data and predict which customer is most likely to leave. Data mining typically uses billing data of clients, subscription information (contract details, features, expiration date, etc.), customer information etc. to form predictive models. Based on the constructed model, the companies can take action regarding the clients if they wish to.

For example, a company may offer the client incentives such as free phone, or more minutes on their international roaming data etc. so as to retain the clients. A successful implementation of data mining was the use of a neural network to estimate the probability of cancellation of a customer contract at a given time t in the future (Mani et al., 1999).

2.1.3. Fraud Detection

In the telecommunications industry, fraud is a prime issue that is present since a long time. There exist elements that tend to use the services offered by companies in an abusive manner such as carrier identity theft, piggybacking, not paying the phone companies, etc. Another part of the issue is that a legitimate customer can have his carrier identity stolen and framed for fraud as well. This situation is delicate as there is a threat of customer attrition if the customer is wrongfully framed for fraudulent behavior. The client may choose to leave the carrier and switch companies if he loses his trust in his existing phone carrier. Additionally, if a carrier has proven to be a target of fraudulent behavior, potential new clients are deterred from subscribing to this carrier as they fear the security of their line and they may think that they will suffer the same consequences. Also, as there exist multiple carriers, if other carriers have not been a victim of fraud and if they offer the same services at the same costs, the customers would trust other carriers more.

Fraud is prevalent in the telecommunications domain as tracking of the fraudster is extremely difficult and expensive. Additionally, it is a time consuming task which needs a lot of resources to be dedicated to it which is not ideal for telecommunication companies in terms of finance revenue. Furthermore, in order to commit fraud in this industry, a fraudster doesn't necessarily need a lot of technically sophisticated equipment. Typically there are two possible victims in case of a fraud- the telephone carrier or a

customer. The carrier loses its resources that are now being used by a criminal who didn't intend to pay for the services he used. Secondly, a client could be a victim of fraud attack in which case his integrity is put under question and he is investigated to validate the credibility of the scenario. If the customer is indeed a victim, the false accusation of fraud will compromise the trust that the client has put into the company. Moreover, the fraudster may use the customer's line and the customer may be wrongfully billed which further shakes the trust a customer puts in the phone carrier. Often, the motivation for committing fraud is monetary gain.

There exist several types of frauds. Some of them are as follows:

Subscription Fraud: This type of fraud is quite common where a customer subscribes to a particular service (e.g., Internet access) without having an intention to pay for the services he used.

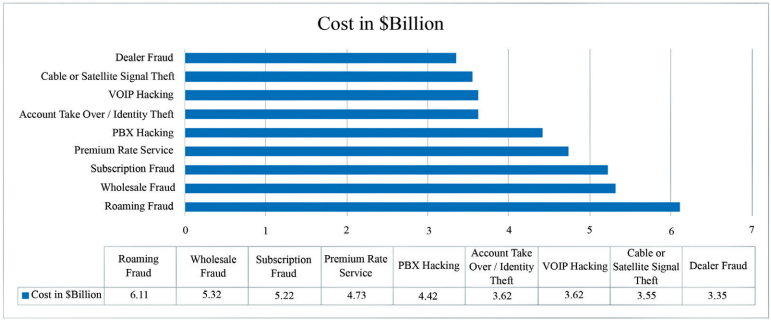
Technical Fraud: This fraud is becoming increasingly common and is mostly related to hacking of the telecommunications systems. A malicious person may exploit the technical weakness of the system for his own gain. The hacking can be performed by means of trial and errors or by means of snooping softwares and so on. This usually happens in case of services that are new and not tested in depth for vulnerabilities. Superimposed Fraud: This type of fraud activity is also called as identity theft where a malicious person gets access to a client account illegally. The fraudster gets unauthorized and complete access to a customer's account and all his personal data. In the telecommunication industry, this generally involves cloning of the SIM card where a copy of the customer's SIM card is illegally made and cloned into another SIM card. The criminal, with a malicious intent further uses this SIM card and performs illegal operations using the customer's name and masking his own identity. The unsavory person superimposes the customer and pretends to be that customer and performs wrongful activities that are naturally billed to a legitimate client.

Although this method of fraud is common nowadays, the telecommunications industry has its hands full with it as such kind of fraud detection is hard and it is difficult to sort out the real victims from the fraudsters. This is one of the most reported types of fraud within the telecommunications industry. Usually criminals and malicious people would profit largely from the use of a cloned SIM card and a stolen 'telecommunication' identity. Social Engineering Fraud: This is a type of fraud that is growing at an alarming rate. Malicious people, not wanting to invest their time in performing hacking using the traditional methods

(trial error, softwares, etc.) as they are time consuming, prefer to find out detailed information about the phone carriers by using their soft skills. They charm the people that possess the information that they want into divulging secrets unknowingly. The information needed by the person with malicious intent could be with an employee of the company or with someone within their support center. In this case, the intended victim/target will not have the slightest idea that he is being used to extract sensitive information.

Sometimes, social engineering reaches a new level when a malicious person manages to convince an employee to commit fraud on his behalf.

Internal Fraud: This type of fraud is when an employee of the carrier company commits fraud by using information sensitive to the company for his own benefit. The employee may use the information he has about the company to attain financial profits, sell the information to competitors, perform espionage etc. The fraud may be couple with other types of frauds such as social engineering, technical fraud, etc. A worldwide survey conducted by the Communications Fraud Control Association performed in 2013 reported that the incidents of fraud are on the rise ("Fraud in the telecommunications industry – part 1 – Smart-IPX," 2018). The following diagram shows the different methods of fraud and the loss in revenue caused by frauds:



	Roaming Fraud	Wholesale Fraud	Subscription Fraud	Premium Rate Service	PBX Hacking	Account Take Over / Identity Theft	VOIP Hacking	Cable or Satellite Signal Theft	Dealer Fraud
Cost in $Billion	6.11	5.32	5.22	4.73	4.42	3.62	3.62	3.55	3.35

("Fraud in the Telecommunications Industry – Part 1 – Smart-IPX," 2018)

The above table clearly shows that one of the most common frauds is roaming fraud, followed by wholesale fraud and then subscription fraud. The cost of handling and dealing with fraud is staggering and although the carriers have increased security measures to prevent fraud, criminals continue to find ways around the system and find weaknesses to exploit.

The techniques of data mining have been a strong instrument of help

in changing the detection of anomalous behavior. It allows the detection of activity with a specific degree of confidence, working effortlessly with large volumes of data.

Data mining techniques make use of existing records stored by the companies regarding fraudulent activities. Based on the data present or stored by the company, different classifiers are used to build models or classes of anomalous behavior. Another data mining method that is effective in detection of fraud is outlier or anomaly detection which helps find deviations in the customer behavior and help isolate possible fraudulent transactions within the telephone data.

The use of data mining for fraud detection has been present since a long time. This was first used in case of credit card fraud detection, where deviant behavior of a customer with his credit card was analyzed so as to detect fraud. Before being used in the telecommunications industry, the fraud detection methods were used prevalently in the credit card industry. The same principle has been highly useful in case of telecommunications fraud in order to detect potential fraudulent attacks. Earlier, in the last 15 years in the telecommunications domain, the systems of fraud detection were simply based on certain threshold levels that were pre-defined. Now, with the changing nature of this industry coupled with the advancement in the technology used in this domain, a simple threshold is not sufficient to detect frauds. Now, using data mining, fraud detection models can be built to detect potential risk of fraud and statistically predict the probability of a fraudulent activity. Two methods of data mining that come to mind are association rule mining and predictive modeling. In case of association rule mining, rules regarding irregularity in customer behavior can be established based on the customer's call records. An example of a rule can be the following:

'If a customer A calls a country B more than 20 times in an hour, then flag the customer.'

Another example of a rule could be something like this:

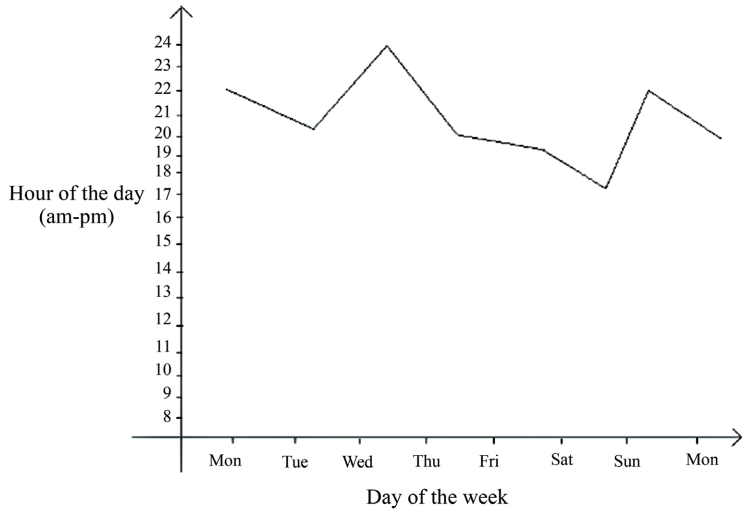
'If a customer uses more than X terabytes of his internet data to download data (videos, audio, etc.) from unsafe sites, flag the customer.'

Once, the rules are applied, the resulting flagged customers' data can be analyzed closely so as to eliminate false positives as this can be possible in some situations due to the vastness of the data. This led to the development of several customized monitoring techniques that were developed by leading telecommunication providers such as AT&T, Verizon, etc. where they used the call history data of a customer to form a threshold/baseline for each

and every customer which was then used to compare every new call the customer made. So, for instance, if a customer regularly called country A 5 times a day, no alerts were generated. But, if he suddenly started calling country B 15 times a day, an alert may be generated or the customer account may be flagged.

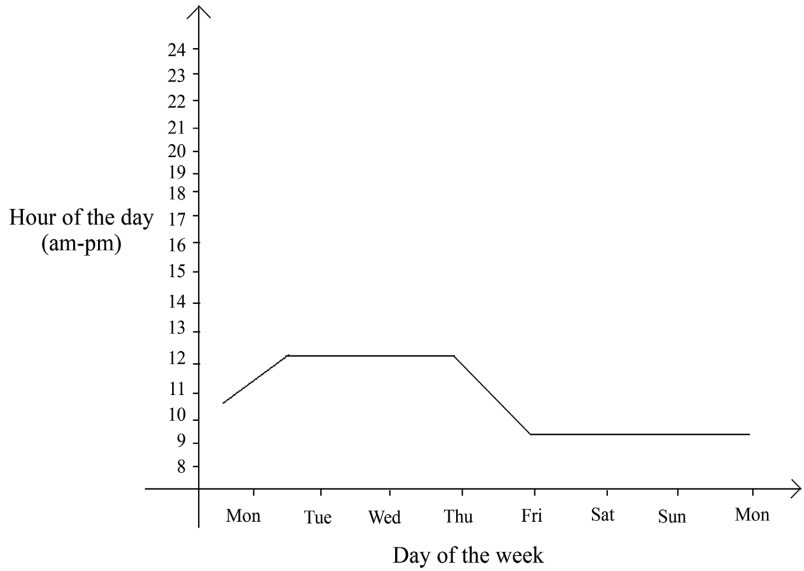
Further, in case of using anomaly detection as the data mining technique, odd behavior by a customer can be detected by developing anomaly detection models. Here, all records of a customer are analyzed to form clusters and the odd clusters, or deviations are further analyzed to detect deviation in the normal client behavior.

Consider the following example that depicts call data of a customer (random) shown by the graph below:



The above data represents the regular call data of a customer. We see that the customer makes regular calls to country X at different times of the day through the week. Now, if the customer suddenly changes his pattern and starts calling the same country at completely different times than usual, it can be considered to be an anomaly.

An example of such anomalous behavior is shown below:



As we can see from the above graph, the customer behavior has significantly changed which may be considered to be suspicious. The data presented in this example is sample test data; actual customer data records are more detailed in nature.

The applications of data mining in the telecommunications industry are vast as this industry generates too much data that is not really understood by the companies and is a potential goldmine for their growth. It is a market that is growing and ripe for data mining research.

2.1.4. Fault Isolation in the Network

The task of monitoring and maintenance within a telecommunications network is a highly important task that needs constant work and efforts. This is because these networks are massive and highly complex with several hardware and software configurations. Although several of the networks are somewhat capable of a small range of self-diagnosis, the networks keep growing more and complex each day. The networks, on a collective basis generate millions of alarms, warnings and statuses each month which renders the maintenance strenuous. As the complexity of the alarms generated kept on increasing, there rose a need for the development of complex systems to handle the management of the network (Weiss et al., 1998).

To manage and monitor a network effectively, all the alarms need to be analyzed in an automated manner so as to detect faults in the network in a

timely fashion before they could degrade the performance of the network. The handling of faults must be done in a proactive manner to ensure that the network remains reliable. As a network potentially stores large volumes of data, a single critical fault may give rise to the execution of several alarms that may seem unrelated and the detection of an isolated fault that caused the domino effect may be extremely hard. Data mining can help simply this process to a great extent by generating classifier classes or using a specialized algorithm for mining the data of telecommunication networks. It can help discover that actually multiple alarms are caused by a single fault.

One of the most common tools implementing data mining that is used for the analysis of telecommunication networks is the Telecommunication Alarm Sequence Analyzer (TASA) that helps with the mining task for alarms and helps find correlations between them (Klemettinen et al., 1999). This particular tool serves to find patterns within the alarm generation in the network data that are repeating and gives their statistical probability using a data-mining algorithm that is specialized for telecommunication network data. This data is then further given to networks specialists who later use this information to build rules around their alarm data that help detect faults in a real-time scenario. For example, TASA may discover the following information about a particular alarm:

'If an alarm type X and alarm type Y occurs within 10 seconds of one another, then a alarm type Z is generated no later than 30 after the generation of alarms X and Y with a probability of 0.5%. Another data mining tool that was used to predict faults in telecommunication networks was to identify problems within circuits (Sasisekharan, et al., 1996). Usually raw data such as phone line information is represented using a time series, which needed to be classified and preprocessed before mining. This approach tried to aggregate data to a level that is semantic using an encoding scheme. This helped represent the time series in a conventional method so that regular standard data mining techniques can be applied to such data.

Yet another tool that was used in the telecommunication networks was to predict switch failures using a genetic algorithm that mined alarm data for temporal and sequential patterns (Weiss & Hirsh, 1998).

Although, data mining is serving the telecommunication industry well, there is a slight potential issue regarding its implementation in the telecommunications industry. In the United States, the public is getting more and more concerned regarding the privacy of their data. It was noticed that telecommunication companies were willing to share customer data readily

with government agencies such as the National Security Agency (NSA) for data mining (Krikke, 2006). This rising concern can result in restrictions being imposed on the data shared for the data mining purposes and thereby legally limiting its use.

The applications of data mining on telecommunications data are possible as the data is extracted usually from the billing stream. But, as the industry changes methods and switches from circuit-switched networks to packet-switched networks, the section of billing is likely to be impervious to phone usage of the customer. This could lead to a loss of data as packet-switched networks are capable of generating greater magnitudes of data as compared to circuit switched data. Hence, eventually, the industry may decide not to store this data anymore and just handle it temporarily. This could pose a serious threat to the data mining research possibilities in the telecommunications industry. But, there are a few scenarios that can change this view. They are as follows:

New Architectures of Networks

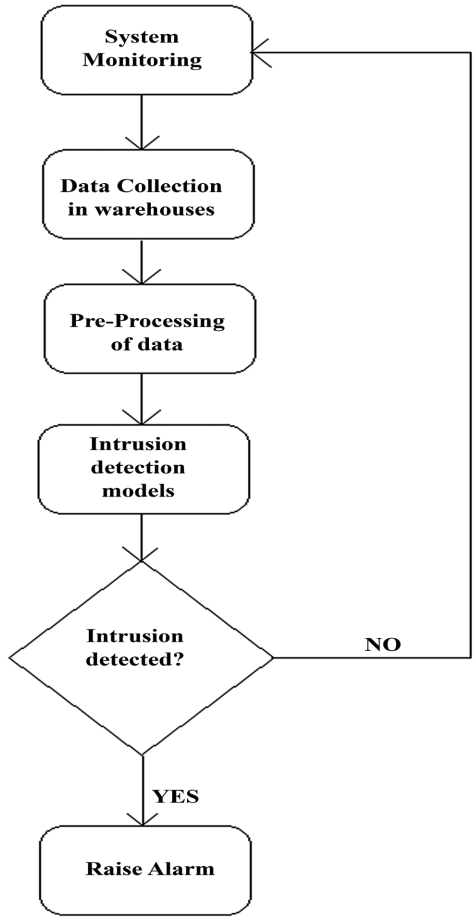
The telecommunications network is a domain that constantly introduces new, next-generation, advanced networks. This signifies that these newly developed networks must answer all the demands in infrastructure and be robust, secure and reliable at the same time. For instance, the capacity (network capacity such as internet strength, etc.) will be needed to be assigned dynamically. This demands that the potential load on a network can be predicted by the company which can be easily done with the help of data mining which can analyze the load and then make predictions based on the existing network load data.

Another potential application of data mining is to ensure the safety of networks by securing then against intrusions. Nowadays, criminals are getting smarter and more sophisticated in their attacks on cellular networks and hence, the telecommunications industry must ensure that their network is robust enough to prevent intrusion of their networks. This field of detection of intrusion will always be open for research by data mining as security is a growing concern worldwide.

An intrusion is basically any act that could compromise a network. It is basically any act that will weaken or breach the confidentiality and integrity of the network. Usually, defense measures include proper authentication of users, firewalls, prevention of errors in programming of the network and protection of customer information using encoding. Data mining can play

an important part in the process of intrusion detection by use of the data mining method of anomaly or outlier detection where anomalies in the regular customer patterns are focused on. A model is created that analyzes the customer data and adds a special level within the data mining model that focuses primarily on odd behavior in the customer's regular activity. It aids in separating the regular behavior and patterns of a customer from the unusual and odd behavior, thereby separating and forming a clear distinction between daily common network activity and odd behavior. Additionally, the data-mining algorithm may be helpful in localizing data that is relevant to anomaly detection.

A typical workflow followed by an intrusion detection system is as follows:



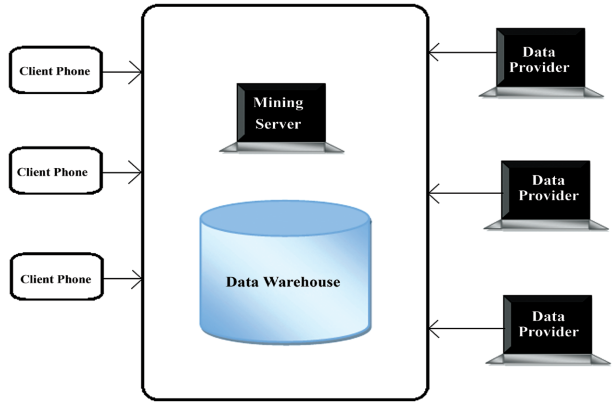
The above flowchart depicts the steps followed in case of intrusions. The system is under constant monitoring using a monitoring system that collects data regarding the network activity. The system stores this data in databases and data warehouses. This information is then processed and refined for the purpose of data mining and then fed to an intrusion detection model that implements data mining algorithms for outlier detection. If it detects an intrusion, it raises an alarm and notifies the interested parties immediately. If no intrusion is detected, the system continues its monitoring activity as usual.

Mobility and Micro Billing

In some parts of the world such as Japan and Europe, the technology of networks is highly advanced. The networks there permit the use of regular mobile phone handsets to purchase items, goods, pay bills, buy items from vending machines, parking meters and so on. All these activities generate vast amounts of data that is 'billable' and hence such data needs to be analyzed. All the data generated by these activities may be stored and hence can be analyzed for various scenarios such as customer profiling, fraud detection, etc.

Mobile Services

In the mobile industry, customer service is of prime importance. It is interesting to know what kind of services to provide the customers so as to retain them and attract new customers. One parameter that is key in attaining new clients is the ease of use of mobile phones and its applications and services. Customers prefer an interface that is easy to use and not too complex and encumbered with several different services all at once that is not easy to navigate. Also, when phone companies wish to introduce new features, it would help them immensely to know the best possible way to introduce new features and services without displeasing the customers. Predictions regarding possible phone interfaces would be helpful to the phone companies as well. Data mining can help in both these cases and possibly several other ways. Data mining can propose solutions that are adaptive that help mobiles companies make better and beneficial business decisions.



The client data and the provider data from different providers are stored in large warehouses. The data stored in the warehouses is processed by a data-mining server running algorithms on the data in order to find the best possible ways to entice customers. Based on customer history, the mining proposes each client the best incentive (free data, promotions, etc.) based on their preferences.

The data mining technique can be used for several possible customer services such as promotions, profiling, performing surveys to predict which customers may purchase additional services etc.

Security

National security is a key concern for all nations. Usually, depending on a countries' government, the laws may dictate that the telecommunications industry hold on to customer call records for a period of a few years. With the world witnessing several terrorist attacks in the recent years, it is of urgent need that the culprits are brought to justice. The telecommunications industry can help in the search for the culprits by providing their data to the authorities for analysis. This is where data mining comes in. The government agencies can use the telecommunications data and mine it for criminal activity. With the help of data mining, using the telecommunications data (phone records) the government can search, track and potentially identify terrorists and their cells and potentially stop imminent threats to the security of their nation.

For example, as per federal regulations in the United States, the companies belonging to telecommunication industry are legally bound to

maintain customer call records for a period of two years. With help of this information, the Department of Homeland Security or the National Security Agency can identify potential threats against the United States.

2.2. FINANCE INDUSTRY

The finance and banking industry is becoming more and more computerized. This domain is one of the most commonly solicited domains by people world over. As banks store people's hard earned money, including that of big enterprises, banks have to handle an obscene amount of data. With the advent of computerized systems and the digitization of this domain, more and more information is being generated every minute. This industry has a constant flow of transactions per day and hence a constant influx of transactional data. The large quantities of data stored by banks are probably just lying untouched, unused in their warehouses and not being put to good use and are probably deleted every few years. The financial experts are unable to mine information and possible patterns, associations between financial data as the sheer volume of data generated is quite large and the task of analysis of this data is daunting for us humans. There exist several correlations, patterns and experts or managers due to the size of the data do not spot associations in financial data and these that they are dealing with. Hence, data mining comes into play.

The financial data stored by the banks can be a potential treasure for solving perpetual business issue in the banking and finance domain such as fraud, customer retention, market predictions and so on. The information mined from the financial data can be put to better use such as customer targeting, acquiring new customers, portfolio management, potential investments etc. The possibility of identifying the right information from the right data can result in a loss or gain of millions of dollars which is a substantial amount.

A few of the possible applications of data mining in the financial industry are as follows:

- Targeted marketing in order to increase customer loyalty: Classification and clustering of customers for targeted marketing.
- Prediction of loan payment: Credit rating and risk estimation and analysis
- Prediction of customer behavior for launching new services and products

- Prediction of company's performance
- Detection of financial distress/prediction of bankruptcy financial ratios
- Identification of fraudulent and malicious behavior
- Security of financial data
- Support of real-time decision making

2.3. BANKRUPTCY PREDICTION

The banking industry suffers from bankruptcy time and again due to the instability of the markets. Hence, to be able to predict possible bankruptcy has now become one of the most used applications of data mining techniques on financial data warehouses. Bankruptcy in major financial corporations can wreck havoc in many ways – economical, social, management, investments, etc. If a corporation goes bankrupt, they lose millions of dollars, they lose the confidence of their clients and investors, they lose their financial standing, they lose their ability to generate revenue and profits and they may also be forced to let go of their employee due to their financial insolvability. This has serious social and economical implications. Hence, it is a worthwhile area of research for data mining.

The use of data mining to predict bankruptcy originates to the work of Altman in the year 1986 where he argued that the failure of a corporation is actually a long-term process and that the data stored by the corporation in terms of financial statements and bills must include some warning symptoms or factors that can predict bankruptcy. He applied multiple discriminant analysis techniques to develop a working model for predicting the bankruptcy of a company (Altman, 1986).

In the years that followed, several researchers developed many other models to predict bankruptcy using statistical models. In the years that followed, sometime around late 1990's and the early 2000's, researchers started actively looking at data mining techniques to build bankruptcy prediction models.

Lin and Mclean (2001) in their paper described four different models of data mining to predict possible bankruptcy.

They made use of the following four methods: Discriminant Analysis, Logical Regression, Decision trees and Neural networks. They proposed an algorithm that was hybrid as well. Their test sample data consisted of the financial data belonging to around 1133 companies in the United

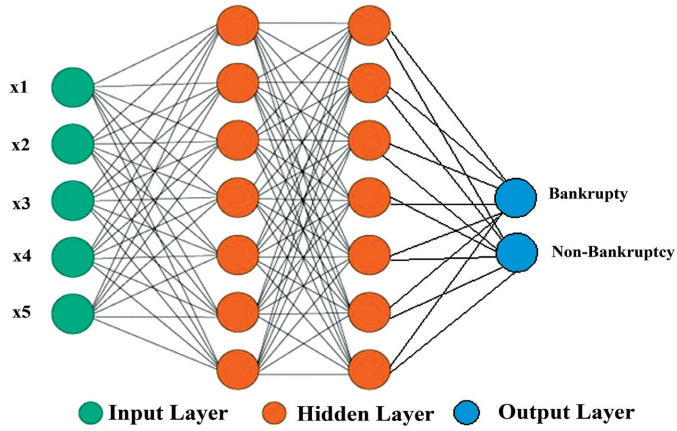
Kingdom of which 690 companies were not bankrupt and 106 were faced bankruptcy. This data was used as a training set against their hybrid model which generated 37 financial ratios that were used as input variables. They implemented feature selection methods in their model that reduced the number of input variables to 4 by means of human judgment. The testing set that used consisted of 289 companies that were afloat and 48 companies that went bankrupt. The results yielded better performance for the hybrid model as compared to the 4 individual methods implemented.

Another model based on data mining techniques was proposed by Tung et al. in 2002 which combined Neural Networks and Fuzzy logic systems. It was called as 'Generic Self-Organizing Fuzzy Neural Network or GenSoFNN.' This model posed the advantage that it could represent complex domain problems in a simple linguistic model (if then else statements) rather than complicated mathematical and statistical models that can only be understood by the domain experts. Their algorithm was made up of simple IF-THEN fuzzy logic rules that have the capacity of self-adjusting the parameters of the fuzzy logic rules that are developed using self-learning techniques that have been acquired from the neural network technique. This possibility gives this model an edge over other models as it presents a linguistic approach rather than a complex statistical and mathematical approach which makes it easy for regular researches (not necessarily proficient in the statistical and mathematical domains) to understand and implement the model to mine data in order to predict bankruptcy. This model makes use of clustering methods to avoid noisy data by increasing the model's capacity to tolerate noise called as discrete incremental clustering (DIC).

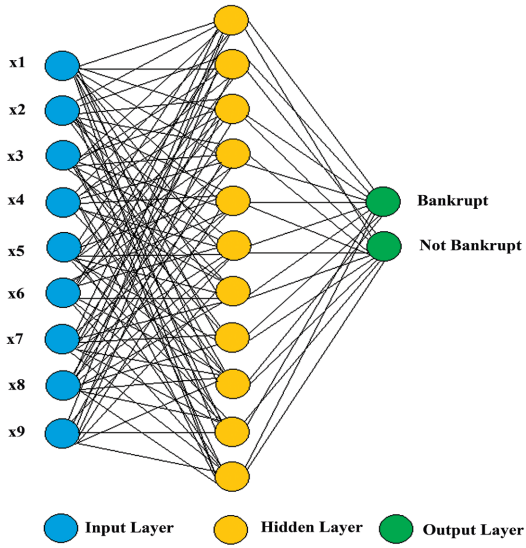
The model took in nine specific variables that were financial as their model input. Previous studies had revealed the importance of these nine specific financial variables and their contribution in making the predictions more accurate. The sample or training set used by Tung et al. was made up of 548 banks that were bankrupt and 2555 banks that weren't. Of this data, around 20% of the data was used as the training data set while the rest (80%) was used as the testing set. The data was processed to include equal number of bankrupt and non-bankrupt banks. The model was applied to the processed data and it generated around 50 sets of IF THEN rules of the fuzzy nature that illustrate the connections, interactions and correlations between the nine specific input variables and the influence they have on the financial outcome of the organizations under consideration and analysis. The models reinforce the initial stated importance of these nine variables on the financial health and well being of an organization.

The results reported a performance of 93% on a testing data set of the last financial year and a performance of 85% on a set from 2 years ago and 75% accuracy for a data set of 3 years ago. This clearly states that as the data gets older, the accuracy of the algorithm keeps on decreasing.

A possible implementation of a neural network for bankruptcy analysis can be seen below:



Here, only five inputs are used, but if we consider the algorithm implemented by Tung et al.,2002 the neural network that generates the fuzzy logic rules would look something more as shown below:



Another model that was developed for bankruptcy detection was one proposed by Shin and Lee in 2002. This model was based on genetic algorithms. The authors argued that although neural networks produce useful rules that further help in data mining, there is a major drawback to the rules derived from neural networks which is the comprehensibility of the rules. They were of the opinion that Genetic algorithms can produce rules that are easy to understand and implement.

They implemented genetic algorithms on one or more financial variables of the organization so as to predict thresholds for the variables, the value of them being classified as dangerous or not based on the calculated threshold. The model implemented using genetic algorithms made use of a rule formation consisting of five conditions combined by the AND logical operator; each of which referred to a variable from the nine existing financial ratios.

One of the rules generated by a model consisting of two financial variables namely quick ratio and debt ratio is as follows (Shin & Lee, 2002):

If DEBT RATIO > 1.50 and QUICK RATIO < 0.35

THEN **DANGEROUS**

Another slightly more complex rule that can be generated using this model is as follows (Shin & Lee, 2002):

IF VAR1 is GREATER THAN OR EQUAL TO C1,

AND

IF VAR2 is GREATER THAN OR EQUAL TO C2,

AND

IF VAR3 is GREATER THAN OR EQUAL TO C3,

AND

IF VAR4 is GREATER THAN OR EQUAL TO C4,

AND

IF VAR5 is GREATER THAN OR EQUAL TO C5

THEN **DANGEROUS**

In the above conditional rule, the values C1–C5 stand for threshold values of the variables VAR1–VAR5 respectively. The threshold values are found via the genetic search algorithm and their range is from 0 to 1. Here, the model selects five out of the nine financial ratios and is also allowed to chose one more financial variable as sometimes the rule structure is non-

linear. So, the addition of another ratio makes the results more linear.

This model was then applied on a data set that consists of 264 bankrupt and non-bankrupt organizations respectively. The model was given the input of nine financial ratios and 90% of the input data set was used as the training set. The remaining 10% of the data was used for the test set that was used to validate the model.

The following five rules were generated by this model (Shin & Lee, 2002):

Rule number	Description
Rule l	IF Net income to stockholder's equity is less than 0.426[a] AND Liquidity ratio is less than 0.847 AND Current liability to total assets is less than 0.520 AND Stockholders' equity to total assets is less than 0.595 AND Financial expenses to sales is less than 0.665, THEN Dangerous
Rule 2	IF Net income to stockholder's equity is less than 0520 AND Quick ratio is less than 0.697 AND Stockholders' equity to total assets is less than 0.590 AND Financial expenses to sales is less than 0.503, THEN Dangerous
Rule 3	IF Net income to stockholders equity is less than 0.426 AND Liquidity ratio is less than 0.560 AND Retained earnings to total assets is greater than or equal to 0.082 AND Stockholders' equity to total assets is less than 0.590 AND Financial expenses to sales is less than 0.590, THEN Dangerous
Rule 4	IF Net income to stockholder's equity is less than 0.560 AND Quick ratio is less than 0.697 AND Retained earnings to total assets greater than or equal to 0.130 AND Stockholders' equity to total assets is less than 0.577 AND Financial expenses to sales is less than 0.515, THEN Dangerous
Rule 5	IF Net income to stockholder's equity is less than 0.560 AND Quick ratio is less than 0.697 AND Retained earnings to total assets is greater than or equal to 0.082 AND Stockholders' equity to total assets is less than 0.590 AND Financial expenses to sales is less than 0.520, THEN Dangerous

The results obtained by this model presented an accuracy of about 80%.

2.4. CREDIT RISK ANALYSIS

As we discussed in the previous section, bankruptcy is a growing concern in the banking industry. This means that there exist many companies or individuals that are under financial duress. Hence, the practice of analyzing the risk and potential credibility of clients and companies came into effect. This practice is known as assigning a 'credit rating.' This task of assigning a credit rating to particular clients keeps on getting more demanding as the creditors offer competitive rates and better services.

Hence, before loaning money to a customer or an organization, a credit check is performed on the potential client who has requested a loan. The process of analysis of credit risk is long and complicated and must factor in a lot of variables such as age, income, debts, investments, market rates, etc. As there are many variables, this process sometimes doesn't yield good results if performed in the regular manual manner. There is where data mining techniques have been an immense help to the process of credit risk analysis.

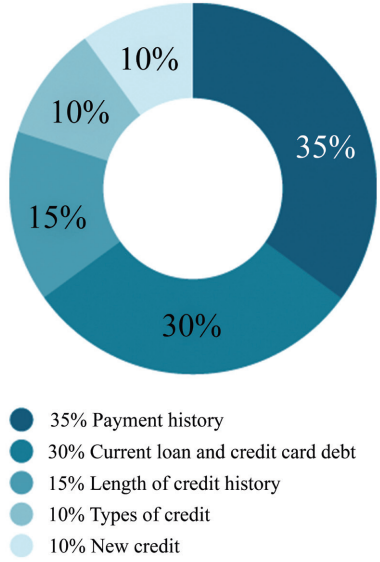
The possibility of a client defaulting on a loan is a sensitive issue that is faced by the banking industry. The assessment of risk associated to giving a loan to a customer is a critical element in such cases, mostly for banks and

financial organizations. The determination of the key variables that play a part in assessing the risk associated with the credit of client is a major part of the data mining process. It is at this stage that the crucial parameters that can determine the monetary risk are discovered. A very well-known method that is used to determine the possibility of an applicant asking for credit to default on his financial obligation is credit scoring. The method of credit scoring basically takes a look at the financials of the applicant and decides if he is likely to default on his payments by assigning the applicant a credit score. This method needs to be precise and efficient as the correct judgment of the credibility of a client or an application allows the bank or financial organization to increase the amount of credit that they have granted and at the same time, minimize their losses by investing in clients having a 'good credit.' A credit score is very similar to the GPA system in schools where a number is calculated and your success is measured as compared to others at the school. A credit score is basically the same thing – a number that grades your potential as a credit-worthy person.

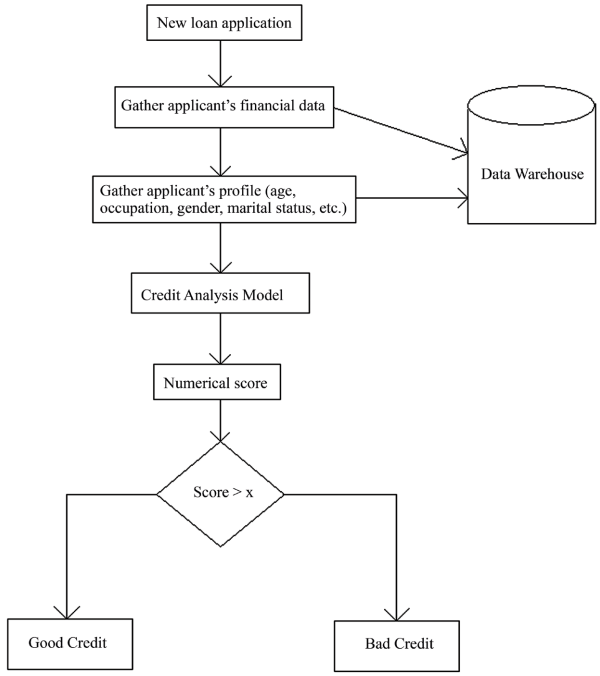
Hence, many models that scored 'credit' came into existence. Initially, these models were mainly statistical models that are used world-over to predict the risk of an individual or an organization. These models are generally multivariate; taking in several parameters. Mostly, the parameters taken as input were the main financial and economic indicators of the applicant. In case of an individual these parameters are simple financial ones such as income, age, marital status, etc. whereas in case of organizations they are the financial and economical values such as statements, credit, debit, etc. These models took in these parameters as input and then assigned weights to each one of these indicators which signifies the importance of each parameter in the calculation of the credit rating and prediction of a client defaulting. The final result is often a numerical value or an index that denotes the worthiness of a client in terms of credit which basically measures the probability that a client is going to default on his loan.

One of the earliest models for scoring of credit was developed in the early 1900's by authors Fisher and Durand in the year 1930. The main functionality offered by a credit-scoring model is to generate and classify the applicants into two distinct classes namely: 'good credit' and 'bad credit.' As the name suggests, the 'good credit' class is the good class and the applicants that fall under this class are highly likely to reimburse the loan they took whereas the class 'bad credit' consists of applicants that will not be able to make the payments on their loans and hence they should be denied credit as the probability of them defaulting on their financial obligations

is very high. The classification of a client as having 'good' or 'bad' credit is usually dependent on the basic characteristics of the applicant such as age, income, occupation, education, marital status etc. and some crucial key factors. A key factor that is used to assign a credit score is the payment history of an applicant. Things such as previous loan payments, the types of loans undertaken, etc. are looked into. The model usually checks if the applicant pays all his bills on time over a period of a few years. Secondly, the model checks how much money the applicant uses as compared to the money he earns. Basically it is a ratio of the money used and the total money present in the applicant's account. The money lenders tend to think that the closer the ratio is to 1 (applicant spends and earns the same amount), their chances of defaulting on a loan are higher. Another crucial factor that is considered is the type of credit an applicant has. If a client has different types of credit such as car loan, student loans, a mortgage, the creditors tend to believe that the applicant can successfully different types of credits and is able to pay them off. FICO (Fair Isacc Corporation), a credit rating tool, bases the credit score on five different factors that are shown below ("How to Calculate Your Credit Score – Get Credit Report – Wells Fargo," 2018):



Based on the above factors, the process of credit scoring can be outlined as follows:



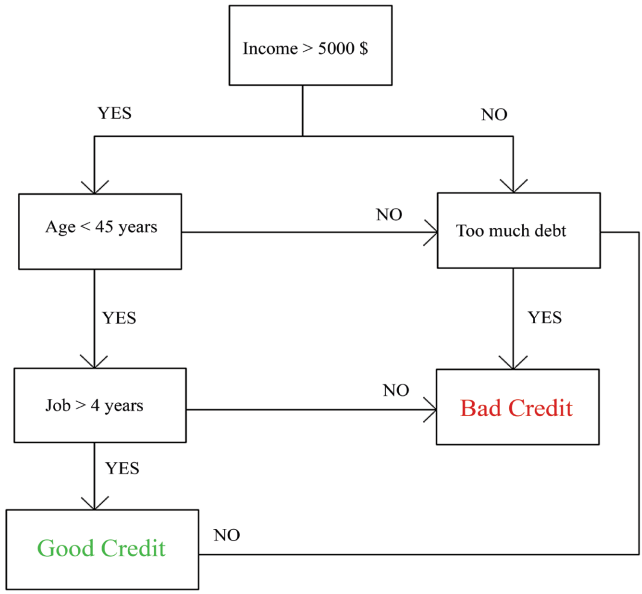
These statistical models can be applied to small to medium businesses as well as they can be considered to be as groups of customers.

A technique for determining credit rating was introduced by Huang et al. (2003) that made use of Support Vector Machines (SVMs), a technique based on machine learning. To test this model, the use of two data sets was made; one had 74 firms that were Korean and another that had 265 firms. Two data sets were used; one containing 74 Korean firms and the other containing 265 US firms. A 5 rating categories were defined for the two sets. Two models for each of the data sets of each country were built; each having a different input vector. Support vector machines and a neural network with back propagation were implemented and used to predict the credit rating. It was found that the vector machines performed better.

Another data mining technique that was used to perform evaluation of credit risk was decision trees in 2003 (Mues et al., 2003) where visualization was possible with the help of decision trees. The advantage of decision trees or diagrams was that they avoided repetition of sub-trees that were similar. To test model, two data sets were used by the authors. The classification of the data was done with the help of a neural network. Then, to extract the rules, Neurorule and Trepan algorithms were applied. Decision graphs were

built with the help of Entropy-based decision graphs which represented the rules in a tree structure. It was discovered that the performance of the Neurorule and Trepan was very good as compared to other methods of rules extraction. The last step involves graphical representation of the rules using decision diagrams.

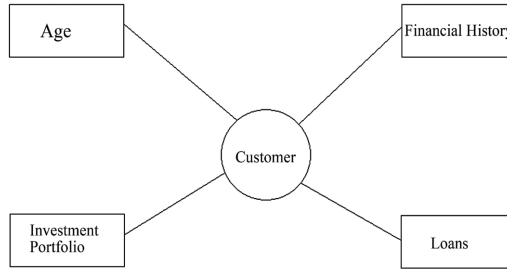
A sample decision tree along with a classification implementation that calculates credit rating is shown below:



2.5. TARGETED MARKETING

Data mining can be used to run marketing campaigns so as to increase sales and introduce new products. Nowadays, banks have a wide range of customers each one of which is looking for something specific. Plus, customers today have a lot of options in terms of choice of banks as there exist over thousands of banks that provide a lot of incentives. Hence, customers always have their pick of banking options where they can take their business to. Hence, the bank managers must be constantly aware of the needs of their customers as if they are unable to offer the customer what they want, the customer can simply relocate to another banking firm. If the bank loses a customer, they lose potential profits and investments and also they need to invest a lot of resources into attracting new customers.

Each customer has a lot of data associated to him.



The banking sector, like the telecommunications sector has a lot of data about the customers such as their financial history, their purchase history, transactional data, investments data, etc. Here, is where data mining comes into play. This stack of data stored by the banks helps identify the call behavior patterns of a customer using mining techniques. By analyzing the customer data, profiles of customers can be generated and several marketing strategies can be developed so as to ensure customer loyalty and attract new customers. Additionally, forecasts can be made based on the information recovered about the customers with the help of data mining.

Earlier, the data mining techniques were more oriented towards the extraction of quantitative and statistical data from the data warehouses. These conventional techniques were useful in acquiring interesting interpretations of the data and getting insights into the processes behind this data. Although a statistical and quantitative representation of the data was provided, it was later analyzed by human analysts. Although these methods eventually led to the discovery of knowledge, it is still susceptible to human errors. The human analysts can miss out on some useful information that could help improve the business processes and profits. But the basic idea behind these methods is the same; data is fed to a system as input, and this input is then mined to give new information or the output. This process can be described as turning information stored in the background (input) to output.

Data mining can help in better ways as compared to existing methods. The techniques used by data mining reduce the human effort by a significant percentage. Additionally the data mining techniques can offer different ways to consolidate the customer data.

There are two major applications of data mining in terms of marketing: Customer retention and customer attraction. Both these applications include deep analysis of existing customer banking data.

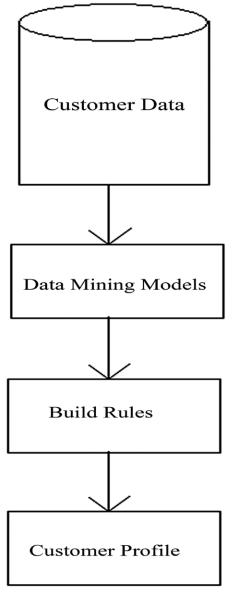
Several data mining methods such as classification, regression analysis, association rules, CART, etc. can be successfully applied to the customer data to build profiles on the customers and offer them incentives and services to entice them to continue working with the same banking firm. The use of these techniques can help the banks look for existing customers that may be interested in their new products and services and then maybe offer loyal customers some rewards and incentives.

Additionally, data mining can help analyze and sift through the purchasing history of customers and then offer them incentives that are specifically tailored to the needs of each and every customer. This serves to retain the loyalty of the existing customers and assure the customers that all their banking needs are being satisfied by their bank. Targeted marketing is the preferred way of retaining the customers as well as attracting more customers as mass campaigns do not generate good positive responses and are usually not worth the time and effort that goes into their production. Targeted marketing is necessary in today's banking world as the churn of customers is a rising concern in this domain; similar to the telecommunications sector. Just like the mobile industry, customers are offered rewards to switch their banking firm.

For example, in France, students are offered 100 euros for opening an account with a bank X let's say. The same tactic is employed by a bank Y which offers 120 euros to students to open an account with them. So, what the students do is that they open an account with both the banks to gain incentives and after a respectable amount of time; they close their account with one of the two banks. So, the bank is at a loss in terms of financial resources as well as customers which is not in the best interest of the bank. The loss of a customer can be potentially very expensive if we take a look at the bigger picture. Here, data mining can be a lifesaver that can help predict the customer churn and help banks decide which customer is going to leave them soon by converting the data in useful gems of knowledge.

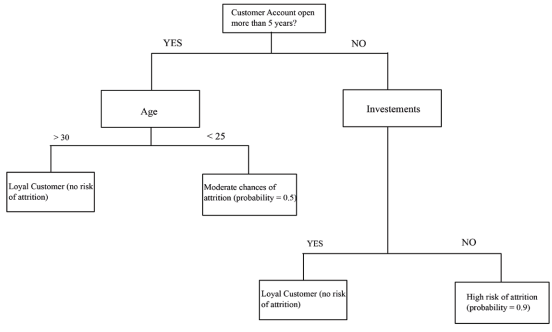
Further, based on the knowledge acquired from the data of their existing customer the banks can determine what the existing customers want and then offer the same to potential new customers. Along the same lines, new customers can be targeted for services and products. With the help of data mining, the banks are now capable of taking more business decisions that are customer-centric.

The process of building a profile is simple and is shown below:



Customer retention usually involves measuring the retention rate, identification of root causes of retention and the development of an action plan to improve the situation. The process of measurement usually can be a significant step in the process of data mining where the classification of attrition can be done and rules can be extracted based on the data of the customers who quit. There exist several implementations of data mining for marketing purposes in the financial sector (Ling & Li, 1998; Hu, 2005; Li et al, 2010).

We shall now take a look at classification of customers using a decision tree.



The above decision tree shows a simple example of rules based on the client data that predict the probability of attrition based on parameters such as age, period of time the customer has been with the bank and the investments the client has made so far.

The use of association rule mining can be built to rules pertaining to the possibility of a customer churn.

Some of the rules can be as follows:

Rule 1:

IF AGE < 25

AND

IF NUMBER OF YEARS WITH BANK < 4

AND

IF NUMBER OF TRANSACTION PER MONTH < 10

AND

IF INVESTMENTS =0

THEN CHURN = 1

Rule 2:

IF AGE > 30

AND

IF NUMBER OF YEARS WITH BANK > 5

AND

IF NUMBER OF TRANSACTION PER MONTH > 100

AND

IF INVESTMENTS > OR EQUAL TO 2

THEN CHURN = 0

Rule 3:

IF AGE < 25

AND

IF NUMBER OF TRANSACTION PER MONTH < 10

AND

IF INVESTMENTS = 0
THEN CHURN = 1

Rule 4:
IF AGE > 30
AND
IF NUMBER OF YEARS WITH BANK < 2
THEN CHURN = 1

Rule 5:
IF AGE > 50
AND
IF NUMBER OF YEARS WITH BANK > 10
THEN CHURN = 0

Rule 6:
IF AGE > 45
AND
AND
IF INVESTMENTS = 5
THEN CHURN = 0

2.6. COMPANY PERFORMANCE PREDICTION

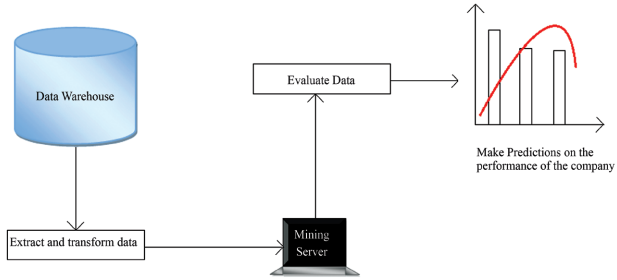
The banking industry is under constant stress to generate revenue. Hence, they often engage in estimations of the company's growth and future based on the current state or the current time period. This usually is some sort of trimestral review where the performance of the last 3 months is used to give an estimation of the performance of the upcoming 3 to 6 months. The same procedure is performed to calculate and estimate the performance of the firm for the entire year.

Usually these estimations are done with the help of the statistical analysis tools that provide some information regarding the data. But, the data that is generated by these tools is later analyzed by financial experts that further

extrapolate this data, apply formulas on it and then make predictions on the performance of the company. This process is subject to human errors. The experts might have missed some information that is present in the resulting set or it is possible that the statistical analysis didn't yield detailed results. Hence, data mining is considered in such cases.

Data mining can be implemented successfully with the help of several data mining techniques such as association rules, sequential rules, etc.

The process followed is as shown below:



The process involves accessing the data stored by the banks within the data warehouses. This data is then pre-processed and fed to the data processor. The processor that processes the data extracts the data and transforms it so that is ready for processing by the data mining methods. The processed data is fed to the data-mining server which applies the data mining algorithms on this data. The data mining techniques applied on the data generate results that can later be evaluated and presented in a visual manner using charts and graphs. The results basically predict parameters such as the growth rate, estimated profits etc.

For example, simple predictions generated by applying prediction models can be as follows:

Based on the current growth of the company, it can be said with a probability of 0.8 that the company will earn a profit of 15% at the end of the fiscal year.

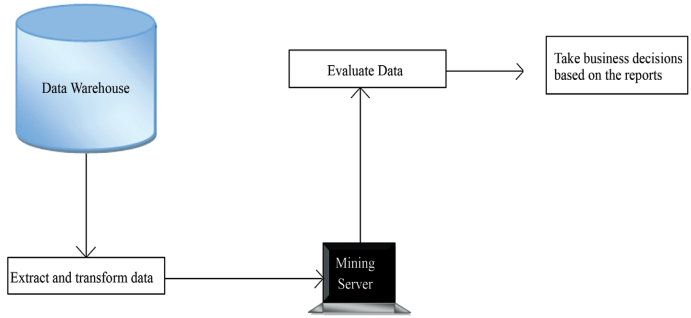
Based on the current growth of the company, it can be said with a probability of 0.7 that the company will earn a profit of 12% 15 months from now.

Based on the current sales rates of product X, it can be said with a probability of 0.9 that the product will not generate revenue in the upcoming fiscal year.

Based on the current sales rates of product Y, it can be said with a probability of 1.0 that the product will gain a profit of 25% in the next 6 months.

Many such rules can be generated with the help of data mining. The approach above shows the use of the results in a graphical context. But, the data mining techniques can yield results that are useful for other purposes as well. In case banking companies wish to take some business decisions, what they can do is that they can analyze the data, what they can do is that they can analyze the results of the data mining process and then based on the results, they can take decisions that help them generate more revenue and earn more profits.

This is shown below:



As we can see from the above image, the process followed is the same as before, but instead of representing the results visually, the data is further analyzed by the managers and then based on their internal analysis, business decisions are made regarding the firm.

For example, consider the following rule:

The investment I is in debt of 10% and based on the current data, it is estimated that the percentage of debt will rise by a percentage of 25% in the next fiscal year.

Now, if a banking firm is already suffering from losses, they would definitely like to minimize all possible avenues of risk and potential loss. Hence, if they receive a rule that says their investment is not going to pan out and is probably going to fail, they might reconsider their investment I. They may decide to un-invest and cut their losses before they could possibly lose millions of dollars and fall into bankruptcy. So, with the help of rules like this, the managers can make a decision to stop their investment I and maybe invest a part of the amount in another sector.

2.7. BANKING FRAUD DETECTION

Fraud in banking is a serious cause of concern for the banking industry and will always remain a cause of worry for most banks. Banks lose millions of dollars on a yearly basis due to fraudulent activities and criminal behavior. The identification of malicious and fraudulent behavior is necessary for this industry in order to react early and avoid loss of considerable amounts of money and resources. The banking industry, due to the rise in such wrongful criminal activities over the last few years, has started looking for effective ways to combat the issue of fraud and malicious behavior.

One of the biggest issues faced by the banks is the dearth of data that consists of genuine transactions and fraudulent ones and the inability of banks to see the difference between a suspicious transaction and a regular transaction. Usually fraudsters tend to perform transactions that seem to be genuine but actually are not, but the issue is that there are very subtle differences between the genuine and the fake ones and hence the difference between them is very hard to spot unless and until you know what you are looking for. The process of fraud detection primarily involves making the distinction between real and fake transactions and classifying them into lists or classes. These classes can then be used as a baseline for new incoming transactions.

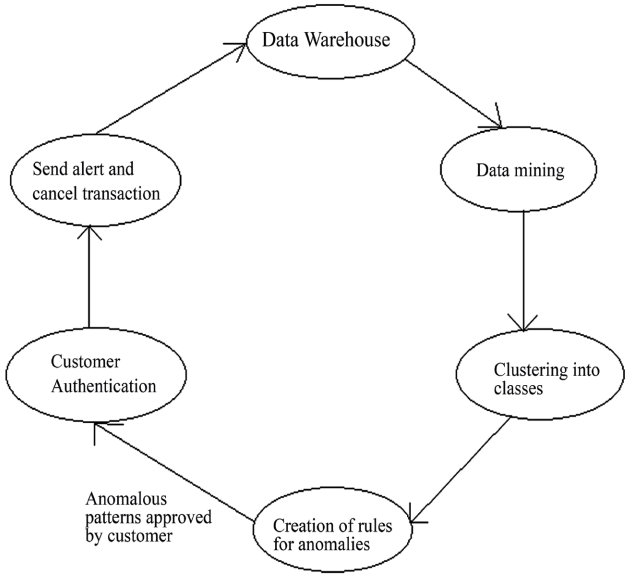
Data mining classification, anomaly detection and association techniques are a perfect fit for the identification of fraudulent behavior. The implementation of data mining techniques on the transactions data of customers helps the recovery of useful knowledge from historical data. The data mining techniques help discover and excavate links and correlations between financial variables to find activities that are suspicious and possibly pose a threat of fraud to the bank. These techniques help shed the light on the transactional activity and help segregate the data into types: fraudulent and non-fraudulent. This process of segregation can further help detection of crimes such as credit card fraud, money laundering, discretionary accruals, etc. We shall take a look at some of these.

One of the well-known applications of data mining is the detection of credit card fraud. Credit card fraud is quite prevalent in most countries. Worldwide, reports estimate that losses due to credit card fraud in the year 2016 were over 24.71 billion dollars. The United States' suffers significantly from this ailment. The reports found that 47% of the credit card fraud in the world happens in the United States of America ("23 Frightening Credit Card Fraud Statistics," 2018). Fraudsters use many methods such as identity

theft, credit card theft, breaching or hacking to obtain a customer's credit card information and they use this information to exploit existing customer accounts to create new lines of credit. They try posing as the person (the victim of credit card theft) and try to get more credit elsewhere. Additionally, a person's credit card information is at risk of breach by unwanted elements as well.

In order to combat this rising concern, banks now look to data mining for possible solutions. One method that was proposed was the use of clustering to study the customer transactions in order to detect outliers in their behavior (Dheepa & Dhanapal, 2009). What this solution implemented was the calculation of the probability density of the past behavior of the user to model and predict the current behavior of the customer so as to see anomalous behavior. Patterns in the customer's transactions can be mined and if any deviations are localized, alerts are sent out.

The process of detection of credit card fraud is as shown below:



The customer's transactional data is stored in data warehouses which is then pre-processed and fed to the data mining module. This module contains the data mining techniques used to generate reports or detect anomalies. This information is stored and when someone (fraudster/ genuine customer) tries to authenticate himself, these rules are applied on the transaction and if an anomaly is detected an alert is sent out and the transaction is canceled.

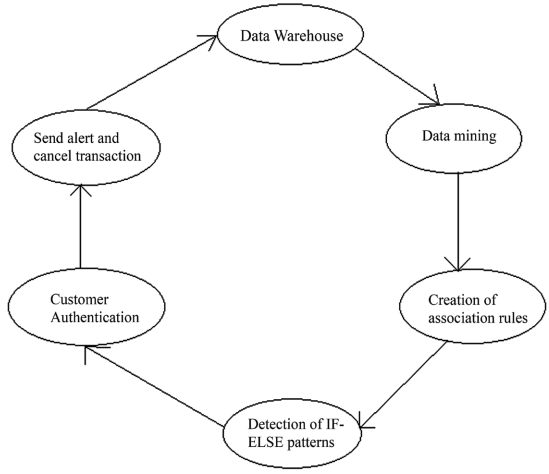
The process in order to detect frauds is described clearly in the above image. Data is fed into the data-mining module that performs the classification of the transactions. Based on the classification, rules are generated to present to the customer and the customer approves the rules that pertain to anomalous behavior.

For example, a rule could be something like this:

'IF customer X tries to withdraw money more than a threshold Y dollars, and tries to authenticate with the wrong pin a total of 4 times, it is not the customer's usual behavior.

In such cases an alert is sent out to the bank and the bank cancels the transaction.

The above method using clustering techniques can be implemented with the use of association mining rules as well as shown below:



Another type of fraud that usually goes undetected is financial statement fraud. The financial statement of a company is one of the most important documents of a company that reflects the status of a company financially ("Financial statement," 2018). This is the document that is shown to the creditors, stakeholders, auditors and the management as well. This statement decides further plans of action and investment for banks. The analysis of this particular document helps the investors and the participants of capital market in taking decisions about their potential investments in the bank as this document represents the financial health of the bank. There exist a set of rules called as Generally Accepted Accounting Principles ("Accounting Principles | Explanation | Accounting Coach," 2018) that define rules that

state accounting principles that are standard for banks. Any deviation from these standardized principles can result in a fraud. The presence of deviations doesn't necessarily suggest fraud but these deviations should be a part of the policies of the company.

The location of fraud in the financial statements is an extremely difficult task. This is because the nature of these statements is different and complex and the tell-tale warnings signs of fraud are usually not visible directly. Even if there exist a few warning signs, it doesn't guarantee the presence of fraud.

The opposite statement is true as well; just because there exist no warning signs of fraud, it doesn't mean that fraud is not present. Sometimes the financial statements can be fraudulent even though, on surface they appear to be in accordance with the standard GAAP rules. Financial statement fraud poses serious threats to the development of banks in this sector and the financial markets as well.

Companies or banks in the finance industry are being challenged with management issues and are put under pressure due to competition in the markets and uncertainties in the world economy. There exist a vast array of financial competitors and possible investments option in the market. Companies in this sector are always on the lookout for new investors, clients and possibilities of growth. During the course of this process; companies need to overcome difficulties, keep up with world laws and keep bettering themselves to appear to be more financially powerful in times of growing changes worldwide. The business challenges linked to investments in the financial industry are enormous and not every firm can bear the brunt of the constantly changing economy and the instability of the emerging markets as well. In some cases, the pressure to gain crosses civil and legal boundaries and some firms feel the need to manipulate their financial statements in order to gain new investors or to escape taxes, etc. This can be done in several ways such as falsifying the company revenues, moving funds to overseas accounts, providing false data in the reports, etc. Due to the competitive nature of this high-risk environment, the cases of financial statement fraud are on the rise. Each reported case of fraud related to the financial health of a firm is a grave issue for creditors, investors, stakeholders and eventually society as well. Hence, the development of efficient models capable of detecting financial statement frauds need to be put into place.

Generally banks take decisions on the basis of the financial statements that are generated by their customers. The statements may be falsified and may contain profits or sales or assets that may be over or understated. This

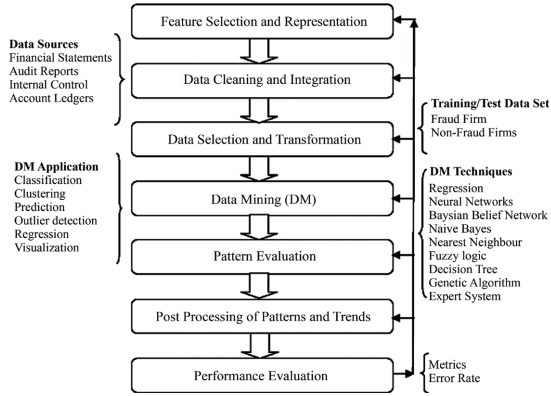
includes liabilities as well. Although all the financial statements are vetted and audited, sometimes, frauds in these documents are difficult to pinpoint because on the surface, they seems to be legitimate documents with no evidence of fraud. Frauds in financial statements are very hard to detect by regular auditing processes as they can be hidden beneath layers of legitimate statements. Data mining plays an important role in the process of detection of fraud. Several data mining techniques can be applied to detect fraud in financial statements such as Predictive mining, association rule mining, text mining, etc.

Financial statement fraud detection is an area that has a wide scope for data mining research. Classification techniques founded on neural networks, regression and decision tree are used for classifying fraudulent ratios in the statements from the non-fraudulent data (Sharma & Panigrahi, 2012).

The K-mining clustering approach has been applied on financial data of customers to choose the best possible investment options for a customer based on their profile (Ingle & Meshram, 2012).

Another approach that has been successfully implemented is the use of text mining to detect fraudulent financial statements. This approach makes use of auditor's footnotes on the financial statements to find indicators of fraud in the textual part of the document.

A simple architecture of a data-mining model applied for the detection of fraud is shown below:



("Google," 2018).

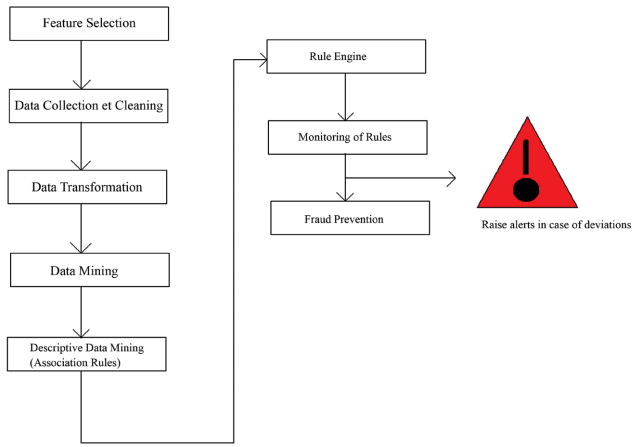
The steps followed by this framework include selection of features and indicator relevant to data mining problem at hand. In this case, it is fraud detection. The data is obtained from data warehouses and is then pre-

processed and integrated. The data is now ready to be processed. Post this step, the data is fed into the selection and transformation tool or module that extracts, loads and then transforms the data and makes it ready for the upcoming mining process. The data-mining module takes in the data, applies mining techniques such as Classification, clustering, etc. on the data to generate patterns and trends. The patterns and trends that are generated are then evaluated using performance metrics such as error rate, percentages, etc.

The above framework is a generic one that can be applied using any of the data mining techniques.

Data mining can also be used for purposes of fraud detection. In this case, the process changes slightly.

This is shown below (Pulakkazhy & Balan, 2013):



2.8. INVESTMENT BANKING

Investment is a term that is well known in the financial domain. Any person in this domain will emphasize the importance of investing prudently. It is basically a task of putting in money towards an asset that can potentially generate profits at a later date. The process of investment is more or less like a gamble that could go either way. It is similar to betting on a card that is said to have the highest value. But, in the financial industry, the values of items keep on changing and hence it is a risky process. The whole aim of investing in something is to reap profits in the future. Investment services are often provided by banks to their customers. Hence, it would be of great use to

the banks to see if they could narrow down the list of possible investment options that are the most promising. This would help them gain more and more profits and also retain their customers if they provide good investment services to them. Data mining can be effectively used to predict the best possible investments based on their profile information such as income, solubility, age, etc. Many data mining techniques such K-means, prediction models, regression models, etc. have successfully been implemented in this domain to predict or provide a list of possible investment options that will yield good returns.

Another operation that is followed by banking firms and other types of firms worldwide is the investment in stocks. Based on the current stock prices, individuals and companies invest money in the stocks of a company that is doing well. They purchase the stocks at a low price and then when the stock prices rise, they sell their stock for a higher rate, thereby making a profit. This is an investment option that is practiced by banks, organizations, individuals, groups on a daily basis. Many firms tend to invest in companies that they think will go on to have better stock prices, but from time to time they are mistaken. When the stock market drops a few points or in case of economical crashes or hits to the market, these firms then end up losing a lot of their money due to bad investments.

Hence, the banking domain has its eyes on data mining as a possible solution to this problem. Data mining has the potential to analyze extremely large quantities of data and the stock market has data in surplus which is very good for data mining. Data mining can perform an analysis of the stock market and generate rules for potential investments. These rules can be helpful to the banks to look for predictive trends so they can invest wisely to be sure that they will gain more profits. Data mining can find associations between companies and dependencies that help form safe investment rules.

For example, consider a sample rule generated as follows:

If the stock prices of company X keeps on falling for 6 months, THEN it will never go up.

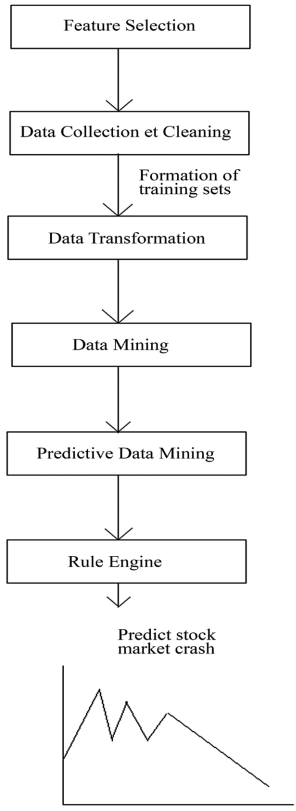
The above statement means that if the price of the stock of company X keeps falling continuously for a period of 6 months, then the stock prices of this company will never rise for a long time which means that anyone who invests in this company will be at a constant monetary loss.

Another rule that can be generated is as follows:

IF the stock prices of company Y rise and the stock price of company Z rise continuously for 3 months, the stock prices of company A rise by 25%.

This rule establishes a correlation between three different elements which gives the investors the possibility to take a decision to invest in the company A. In addition to predicting possible investment options, data mining can also be used to predict possible crashes of the market by studying the existing data and forming a training set.

This process is shown below:



Many implementations of data mining to predict stock prices exist right now as the particular capability of predicting the cost/price of assets from their historical data can help tremendously in increasing the returns on investment.

One technique of data mining that makes predictions using neural networks has been proposed by Das and Uddin in the year 2013.

The technique of data mining can also be applied on time series financial data (Tak-Chung, 2011).

2.9. ONLINE SECURITY IN DATA MINING

The banking industry is growing at high speed. With the advent of the World Wide Web and the internet, banks were forced to keep up with the changing times and introduce their products and services in an online market. With the rise of the digital age, banks were forced to introduce online banking or e-banking. This gives customers access to banking services at all times of the day, 24 hours a day. The customers now have access to services such as money transfers, billing options, online payment options on a regular basis via the internet. This has led to people turning to the online services on a great scale. People now purchase online regularly as they are not physically needed to be present to purchase items.

But as the rise of customers in the online market, the number of malicious people with bad intentions is on the rise as well. The number of cases of fraud keeps on rising. Although we have discussed banking fraud in the previous sections, here we shall take a brief look at the online aspect of banking. As the rise of unseemly people on the internet is evident, customers are constantly being targeted by various types of cyber-attacks. One of the most common frauds is identity theft. In case of online identity theft, malicious people can obtain access to personal information about the client such as their account number, credit card details in an illegal manner. This can be done alone by use of various hacking methods such as phishing, packet sniffing, DOS attacks, etc. Once the attacker has personal information of a customer, he can later use this information to continuously create new credit cards, new accounts, make unlimited transactions and so on. This leads to banks losing their money, customers losing their money and its puts the customer through a great deal of paperwork to prove that his account was really hacked. The customer's credit rating is ruined due to this and the customer will be rightfully displeased with the inability of his bank to protect him from this situation. Hence, banks have now started taking a customer's online data extremely seriously.

Another common form of online attack is 'phishing' where a fraudster poses as the bank or a financial institution and asks the customer to share his private and confidential information. This is done with the help of phishing websites that bear striking resemblance to the original bank's website which is enough to trick customers into entering their private information such as username, password, bank account number, credit card numbers etc. willingly on this site. The fraudsters then go on to use this information to make charges to the credit card, max out the card, withdraw money from their

accounts and so on. This is another major type of online attack. Customers are usually sent emails from sites that seem to be that of their own bank and requesting their personal information. These fraudster provide pop-ups that contains links to web pages that store the data entered (username, password, account numbers) and steal it. Hence, banks now have started sending out regular warnings to their customers directing them not to give out personal information in response to such mails.

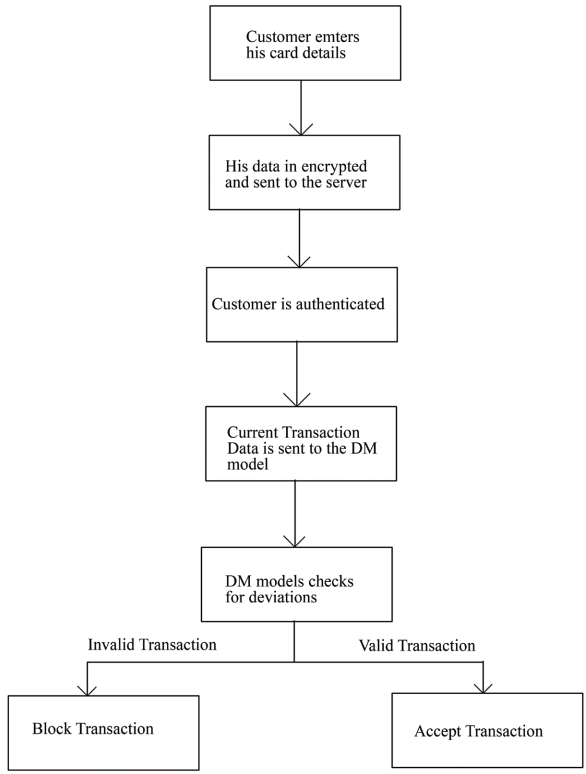
Another fairly common scam is the bank loan offer scam. Here, people receive emails that are 'supposedly' from their banks offering them loans at shockingly low rates for a loan. These emails claim that the bank has already 'pre-approved' the loan and in order to secure it, you need to transfer money in an account as soon as possible. Sometimes, fraudsters can be really ingenious by putting on fake deadlines on these such as 'this offer is valid only for next 48 hours. Hurry!' They demand a mandatory processing fee associated with the loan and many customers, in order to get a better deal, end up paying the amount and losing their money by this scam.

Another well-known scam is the 'lottery scam' where banking clients are sent an email saying that they have won the lottery of a million dollars. This email claims that the recipient has won the lottery and in order to claim his earnings he needs to pay a small processing fee. People are usually skeptical of this type of e-mail, but fraudsters take it to the next level by adding statements like 'The bank has added you as a part of a worldwide lottery competition. All the customers have been added.' This email combined with a phishing approach (look and feel of the bank's website) is able to convince clients that the email is not fraudulent. Hence, sometimes clients end up getting duped by this scam.

Many other scams such as the hitman scam, Nigerian Prince Scam, the greeting card scam, software update scam, etc. are possible. Consequently, banks want to shield themselves and their customers from these possibilities. This is where crypto security comes into play.

Banks have started protecting their customers by using strong encrypting and authentication methods. But, at times, this is not enough to protect hacking of the customer's data. Hence, banks are now employing a combination of encryption, double authentication and data mining to protect its customers from online attacks as well. In this case, the process followed is similar to that of regular fraud detection except that the data is now encrypted and data mining is used to flag odd behavior in the customer's purchasing patterns.

The end-to-end process of a use case of online activity protection is shown below:



An implementation of security mechanisms using cryptography, steganography and data mining has been implemented by Devadiga et al. (2017).

2.10. RETAIL INDUSTRY – MARKETING AND SALES

The retail industry is an industry that has been in existence since a long time. This industry is a highly competitive industry with new businesses being developed on a daily basis. These businesses face a lot of pressure to offer the customers what they want. The retail market is always under a severe competitive pressure. In order to survive in the cutthroat market where businesses keep failing, businesses need to ensure that they are making profits and are able to stay afloat. More retail owners end up using mass marketing that is not directed on a large-scale in order to boost their

businesses. They introduce mass campaigns containing catalogs, pamphlets, banner ads, intrusive announcements on speakers in public areas, door to door campaigns and so on. These methods are proven to be ineffective leaving the customers irate. Additionally, the response they received to their campaigns not positive which led them to lose lots of revenue invested in the process of marketing.

Hence, they needed a new alternative approach to marketing where customers were targeted on an individual basis or on a group basis. The retail industry did not realize until recent times that they have a large quantity of data at their disposition. This industry collects a ton of data about the customers such as their shopping history, sales information, goods records, global purchase records etc. The quantity of data that they are collecting is growing at an exponential rate, which is owed to the rising ease of use and popularity of e-business or e-commerce. Customers and users all over the world are using the online retail industry more and more at an unimaginable rate thereby generating immense quantities of data. This data is just being stored and never used, gathering dust in the data warehouses. This data can be a rich source of research for potential data mining applications. Data mining is a very interesting field and is making a lot of advancements in terms of its techniques.

Hence, the motivation to avoid monetary losses on marketing led to use of data mining as an effective marketing tool. Many modeling techniques are offered by data mining that help improve sales and customer retention. One such technique is market-based analysis where the market data or user purchase data is analyzed and based on this data theories are formed. The principle used being a market-based analysis is that if a customer buys a group of items together, he is likely to purchase another group of items as well. This principle may help the retailer understand the purchase behavior of the customer or buyer. The information returned by the data-mining model can help the retailer know the buyer's wishes and needs and change the way he organizes his store as per the data he received from data mining.

Retail data mining can help immensely with the identification of customer behavior patterns, their shopping trends; help improve the quality of customer service and so on. The work of data mining is not just limited to discovering trends, it can be further implemented to enhance the quality of products and retain customers by improving their satisfaction level.

For example, if the data mining reports say that when customers buy bread, they tend to buy milk and butter as well. If this is the case, a

supermarket or grocery store can change the layout of the store as per the needs of their customer. This can be done for several items and a layout that works well for all customers can be chosen. This information can vary across geographical regions as well. If is one region buying bread equals buying milk and butter then in another region buying bread may be equal to buying juice and eggs. This is where data mining shows its true potential. Based on geographic data, data mining can find rules pertaining to geographical regions as well.

For example different rules such as the following can be obtained by data mining:

IF a customer buys BEER he also buys CHIPS in region X

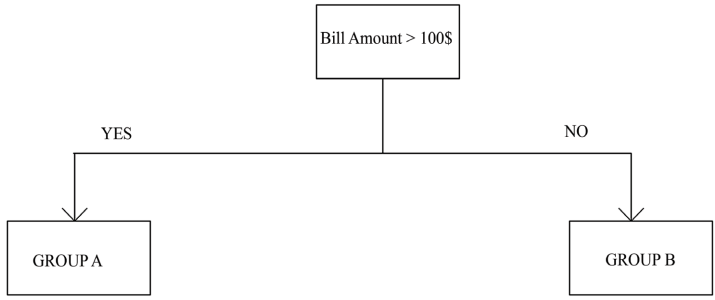
IF a customer buys BEER he also buys cheese and crackers in region Y

Additionally by use of differential analysis several comparisons between different store branches, between different customers, between different regions can be done and presented to the management for further marketing purposes. This can help the management of a retail owner make business decisions to enhance sales and attract more customer while retaining the existing customers.

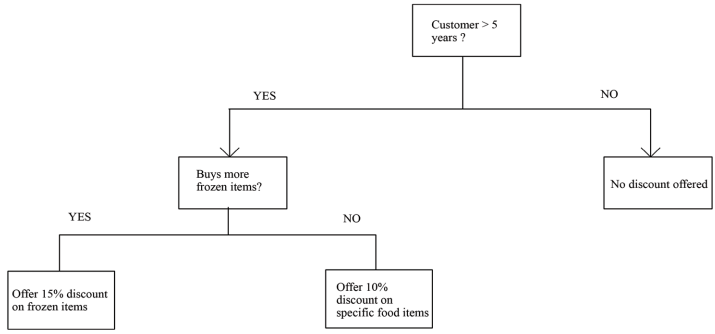
Data mining serves to uncover hidden patterns within historical data of small business and help them better understand its customers, thus enabling small businesses and retail owners in planning, launching new marketing campaigns with minimum cost and promptly.

Another common application of data mining in the retail industry is targeted customer marketing which is typically implemented by grocery stores or supermarkets. The method followed by these retailer businesses is that they offer their customers free loyalty cards. These loyalty cards give the customers access to discounted prices that are not available to other customers that do not possess the loyalty card. Behind the scenes, data mining is implemented to track the customer's purchases and see what he is buying, what items are bought frequently and how much he spends on those items. Based on this data, the supermarkets then offer the customers special discounts that are targeted at individual customers based on their buying habits. They can also offer coupons for a selected set of items that are frequently bought by the customer. They are able to decide when to put certain products on sale and when to offer these products on full price. Additionally they can discover the products that are not sold much and offer promotions based on the products that have some sort of link with this particular product so as to boost sales.

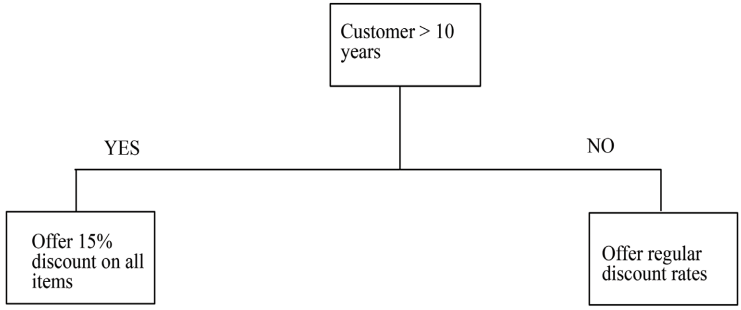
Targeted marketing in groups is also possible with the help of data mining techniques. In this case, customers are classified into different groups based on their purchase history. Consider the following tree that classifies customers into groups based on their buying habits:



Another example is shown below:



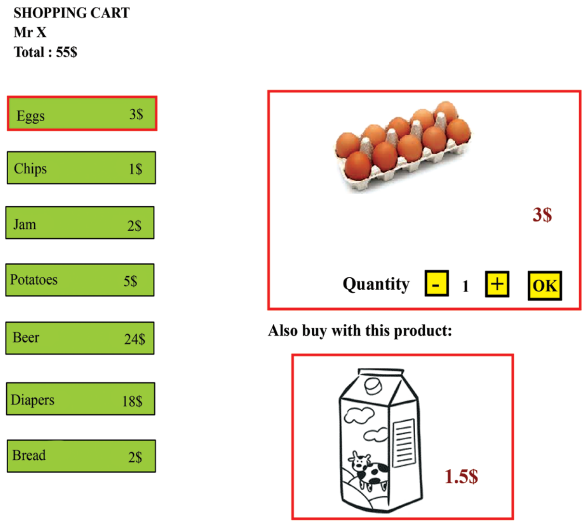
Another example of a simple classification of customers based on their loyalty is shown below:



The same principle that is shown above can be implemented in case of online shopping. Many retailers are now providing customers the

possibility of doing their grocery shopping online due to rise in popularity of e-commerce. In this case, when a customer adds products to his shopping cart suggestions are made to the customer based on his shopping history.

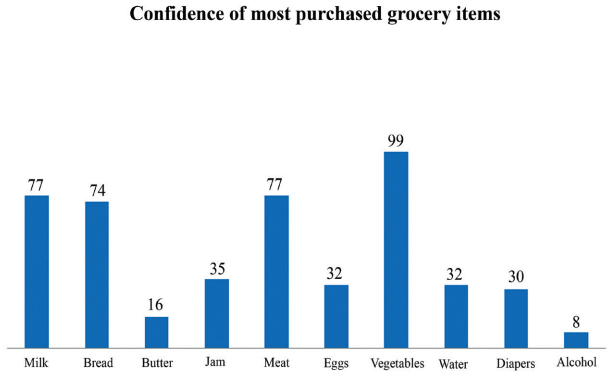
This is shown below:



Additionally, the history of the customer data can be used to predict possible confidence values of products that were purchased.

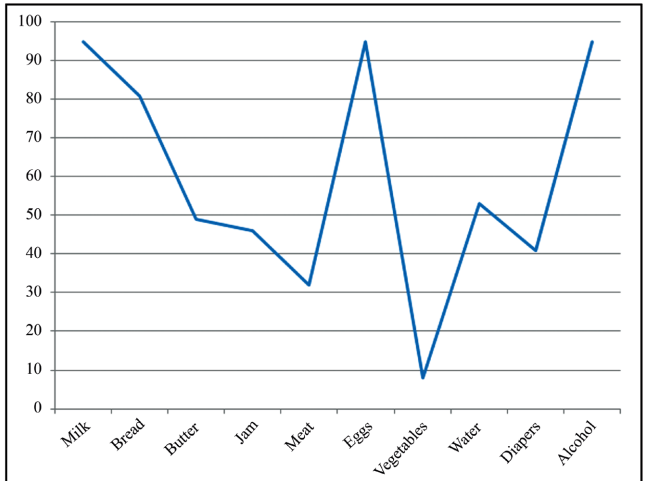
This type of analysis of customer shopping data is also called as market basket analysis.

The data mining analysis can be used by retailers to predict the most bought items and then offer discounts on bulk selling of these items. This is shown below:



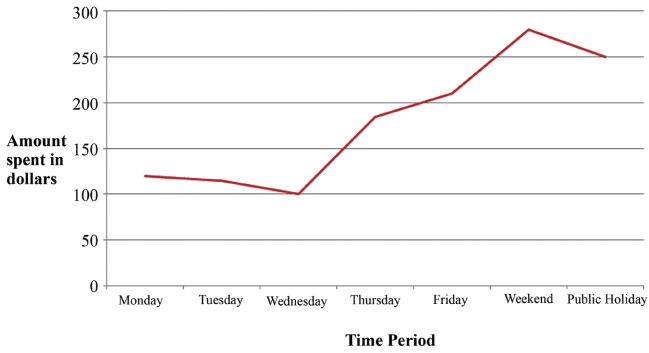
The above example presents a list of the 10 most frequently purchased items in a grocery store. Based on the global data of the store over the years, the store is now able to calculate the confidence level associated with each product. Based on the total support of each item, its confidence level is calculated and a tabular representation is done. This information can then be passed over to the owners of the company who can then make better business decisions in order to improve their business quality, goods consumption, customer retention and attract new customers as well.

The above information can also be represented in a graphical view as shown below:



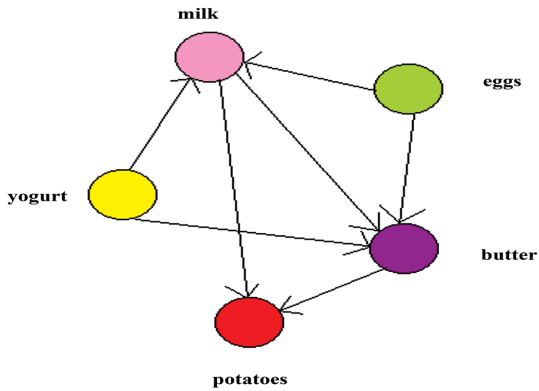
Another type of information that can be seen is the amount spent by people on specific days.

This is as shown below:



The above graph shows a graph of the amount spent by customers on an average against the time when they spent that amount such as particular days of the week, holidays and the weekend. The above graph clearly shows customers prefer shopping on the weekends and public holidays where they shop more and spend more money. This information can be used by the retail owners to develop strategies that help them make more profits. They can use this data to prepare the store for busy days and put into place some measures to assure the customers great customer service. Additionally, they can offer promotions for the products that customers buy frequently to boost sales.

Data mining can also be used to make visual representations of the rules that were formed by applying data mining techniques. This is shown below:



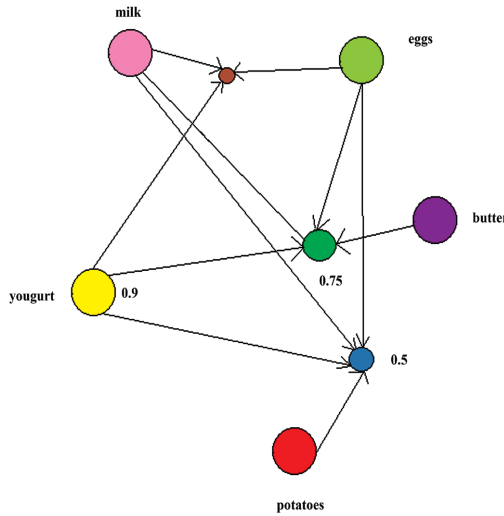
{yougurt, eggs} => {milk}

{milk, butter} => {potatoes}

{yougurt, milk, eggs} => {butter}

The above diagram shows a small graph of three association rules that have been extracted based on customer purchase data. As shown above, each of the rules are depicted visually and are easy to comprehend. This can be done on a large-scale as well where several rules can be viewed in a visual manner.

In addition to just viewing the rules, the visualization can be modified to reflect the confidence levels of each rule too. This can be done as shown below:



{yougurt, eggs} => {milk}

{milk, butter} => {potatoes}

{yougurt, milk, eggs} => {butter}

The above image shows the rules given by association rule mining along with the confidence probability for each rule. This is extremely useful for understanding the probabilities of purchase as well which further helps the owners and managers take informed better decisions.

Although data mining is extremely effective in helping various domains, it does raise certain concerns in terms of privacy. It can be a cause of concern for customers when their individual customer data is used to do targeted marketing to prove a certain rule or hypothesis.

2.11. ENERGY DOMAIN

The overall global markets, the conditions of the market, environmental factors, environmental preservation goals have generated a necessity of bettering the current situation of the energy industry.

The energy domain is now facing a challenge to introduce effective, low cost green ways to generate energy. Hence, there has risen a position for now technologies to help management the energy sources in an effective and efficient manner. The traditional system followed by the energy industry relies mainly on fossil fuel sources such as oil, gas, coal, etc. These resources

are not at all sustainable as the combustion of these fossil fuels directly leads to carbon dioxide emissions (CO_2). Also, these resources are exhaustive and one day they will end. Additionally there are not sustainable environmentally. Another avenue of energy is nuclear energy which generates a lot of nuclear waste which cannot be easily disposed of. This type of energy doesn't have any permanent way of disposal which makes it a liability that is continuously affecting the environment.

This domain is constantly looking for new sustainable energy sources and there is always a constant need and motivation to save energy so as to not deplete existing energy reserves. For instance, the prices of resources such as coal and petrol keep on rising as these resources are being used world-wide. If new renewable sources of energy are not developed there could arise a shortage of these critical resources. Some sources say that if new sources of energy are not developed in the future (next 500 years or so), the world could be looking at serious energy wars over renewable energy resources. This puts the existing sources of energy in a different light. It is now becoming more and more important for the energy domain to economize energy, and to a have sustainable options for the existing energy sources so as to retain the savings over a long time. The need for a dependable and secure energy source is imminent. The best way to combat the issue of energy saving is to produce clean, effective, sustainable and efficient sources of energy; basically all forms of energy that can be saved and reused.

The energy industry is actively looking for ways to predict and project the existing energy consumption as well the possible energy generation of new energy sources. These problems have started getting more and more attention from the research community and the industrial community too. The task of predicting the power that will be generated from new renewable energy sources (e.g. solar power, wind) is highly difficult due to the unsteady nature of weather conditions. These new sources of energy are highly dependent on factors such as the heterogeneous data, unpredictable weather consumption, time periods, etc. which makes the task of prediction extremely strenuous and grueling.

The energy domain consisting of oil, gas, electricity, etc. is a potentially rich are for the implementation of data mining techniques as this domain also stores a data in ample abundance. This domain contains and maintains a plethora of unstructured information in warehouses all over the world. Again, like in most cases, this information is not being put to any good use as it simply sits there, untouched. Hence, many technologies, statistical and

mathematical models are being developed with the help of data mining which can help this industry with predictions and analysis. The implementation of a set of models that are based on current techniques in this industry will result in the growth of a strong suite of tools, applications and techniques that will provide deep meaningful insights into the possibilities of energy saving and predictions of effective energy consumption mechanisms. The data mining tools can help identify possible avenues of research and promise that will introduce the world to new ingenious ways of energy retention along with solid reliable and dependable predictions for newly developed energy resources.

There are several applications and uses of data mining in the energy industry. We shall now discuss some of the applications in brief.

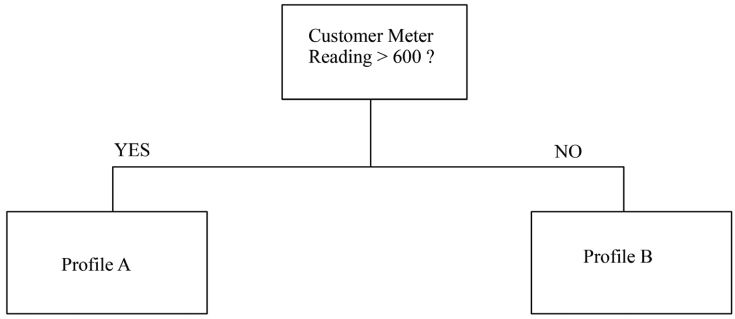
2.11.1 Applications in Electricity

The electricity industry is a type of energy industry that generates electricity and offers gas and electricity on a commercial level to individual customers and organizations. This industry generates electricity and gas with fossil fuel resources that are exhaustive and that generate a lot of carbon dioxide emissions. With the immense pressure on the companies in this industry to be able to predict their energy consumption and energy requirements in the foreseeable future so as to consider avenues where new renewable sources of energy can be applied. As we have seen, this field already stores vast amount of data in the form of customer bills, meter readings, customer electric usage, etc. This data can be then used for the purpose of analysis and research for data mining.

Data mining offers the possibility of analyzing this data and predicting the customer usage patterns based on several factors such as weather, time of the day, electrical appliances owned by the customer, area of the home, etc. What data mining can do is that it can analyze all this information and provide detailed results that can be visually represented and presented to the manager so that they are able to take informed decisions regarding their future plans.

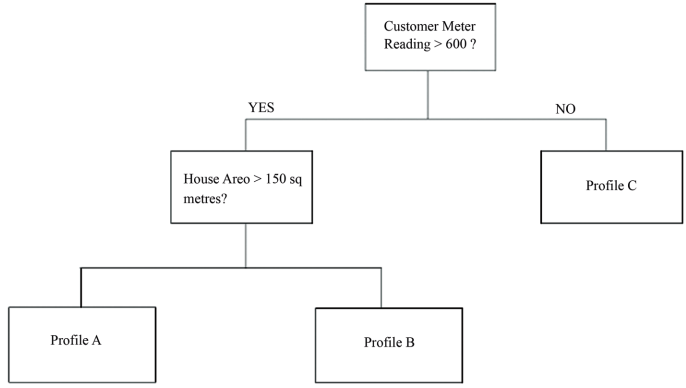
For example, simple classification techniques can be used to classify customers into different groups or profiles based on their energy consumption.

Consider the example below where a customer is classified into different profiles based on their meter readings as follows:



As we can see, a simple decision tree is used to represent the classification of customers into two separate profiles based on their electrical meter reading.

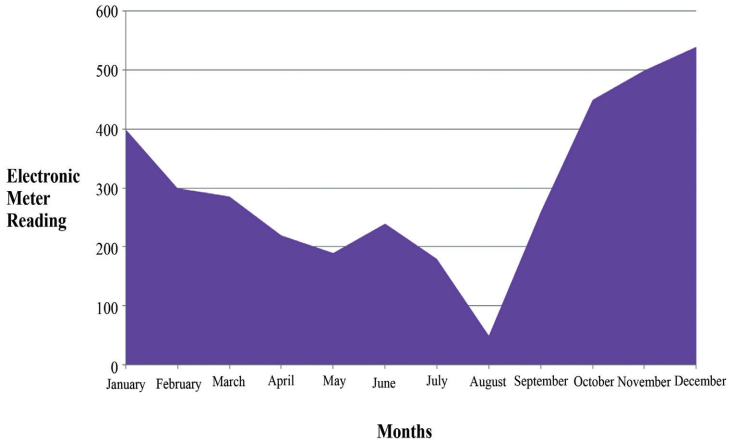
The above classification can further be refined based on other parameters as shown below:



As shown above, the customer is initially classified into profile C first, followed by another classification based on the area of the house that he lives in which is a possible factor that affects the energy consumption and many customers may fall into this category.

Another possible study of existing customer data can be done by analysis of their energy consumption on a monthly basis. In this case, the average user consumption for a group of people belonging to a particular domain is studied and analyzed. The analysis of the customer data helps the electric company makes predictions on the potential energy consumptions based on their existing consumption of energy.

The data can be represented visually as shown below:

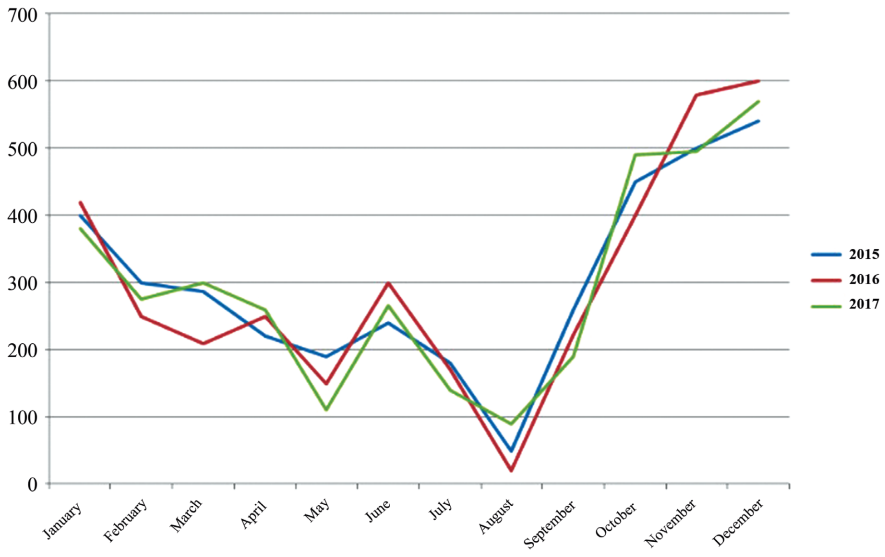


As we can see, the average consumption of a group of customers belonging to a particular profile is analyzed and represented in a visual manner. This data is then provided to the managers of the company so that they can estimate the client's future energy consumption.

Many electric companies such EDF France, use this data to bill their customers. In France, customers are billed a particular amount on a monthly basis that has been calculated by the company based on the customers previous energy consumption. The customer is debited a certain amount every month that has been pre-calculated for them on the basis of their previous year's energy consumption. At the end of the year, if the customer has paid for more than he has consumed, he is reimbursed the amount and if he has paid less than what he consumed, he is sent a bill with the difference in amount. This practice has been used for a long time by many electrical companies.

Sometimes, the pre-calculations done by the electric companies is not very accurate and they end up having to reimburse clients or re-issue bills. Data mining can be a lifesaver in such cases. Data mining can analyze the customer's data for over a period of years and draw conclusions that can be of use to these companies. This information can then be used by the responsible parties to predict energy consumptions more accurately with a margin of error that is not too large. This can help a lot in calculating the budget and help in planning for the upcoming years.

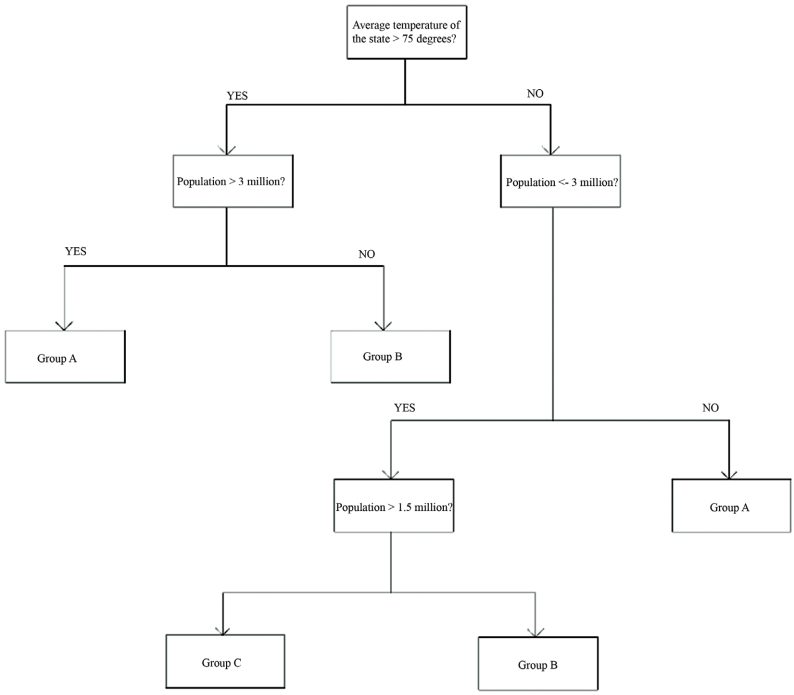
A small sample set showing customer energy consumption over three years is shown below:



As shown by the graph above, the average energy consumption of customers over a few years has been extracted and represented in a visual manner. This data can help in an immense way to predict energy consumption of a global level and thereby save time and efforts at a later date. Predictions for customers can be done in a better way, predictions for customers of a particular region or profile can be done, predictions for individual customers can be done and new plans and services can be developed and proposed to the customers based on their energy consumptions.

Another way of classification of customers can be done on a regional basis. This is because in a large country with many states, the climate may vary from region to region, and based on the regions, the electricity consumption can vary a deal. This information can help electric companies have a consolidated view of states that have similar energy consumptions and then enable them to take decisions based on this information. The managers can decide on the introduction of new billing policies, potential new business needs, the need to develop a new energy generation plant, etc. based on such data to eventually improve their business conditions and reduce the monetary losses and manage their business in an effective way.

A simple classification of regions based on their weather conditions is shown below:



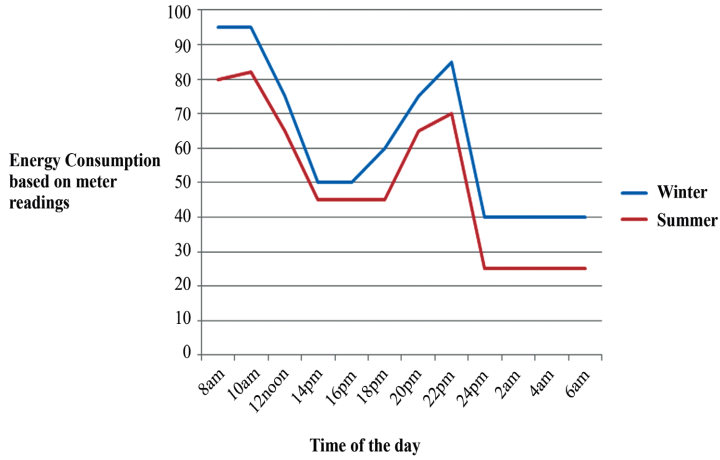
The classification shown the classification of states of a country based on their climatic conditions and their respective populations.

This classification helps the electric company develop global policies and rules that can be applied on a more global level. These rules can be applied based on the initial classification at a high-level. Then on further classification at a sub-level, additional rules can be applied to the data and combined with the global rule.

The electricity consumption of customers can be done on the basis on their daily consumption at different times of the day. The billing policies for the customers can be changed as per time of the day as well as season. Classification of calculation of the electricity bill can be done on the basis of time periods and seasons.

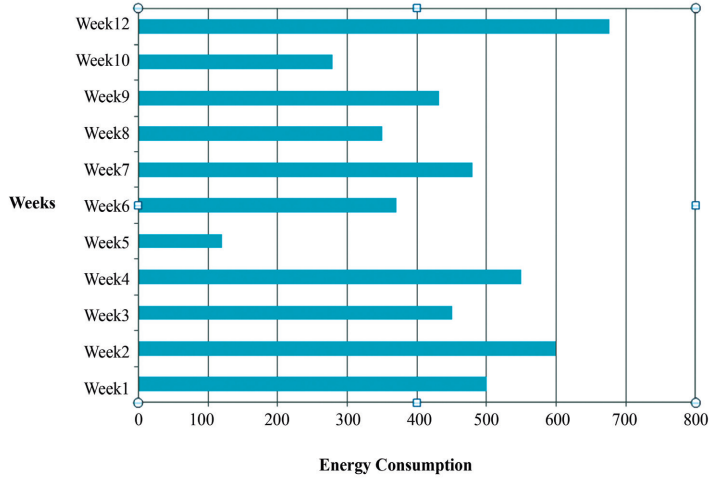
This helps the companies develop and introduce different plans of billing that help save energy and costs for both, the customers and the electric company as well.

A graph showing average electricity consumptions for customers during different times of the day is shown below:



Another way of analysis of customer data can be done on a weekly basis as well. This helps achieve a more high-level view of the customer's consumption and help the experts in many ways. Data mining can help uncover different patterns (if present) between the current week and its associated power consumption. This information could be then analyzed by the electric company's representatives to make informed decisions instead of relying on manual expertise.

An example of weekly analysis is shown below:



Another method of data mining that can be applied on the customer's data is association rule mining. Here, based on the customer's electrical energy consumptions and different factors such as time of the day, area

of house, etc. different rules can be developed and applied to improve the efficiency of energy management.

Consider the following simple rules that can be developed:

Rule 1:
IF ELECTRICAL ENERGY CONSUMPTION > 500
AND
IF HOUSE AREA > 170 SQUARE METERS
THEN BILLING POLICY = A

Rule 2:
IF ELECTRICAL ENERGY CONSUMPTION < 500
AND
IF HOUSE AREA <= 170 SQUARE METERS
THEN BILLING POLICY = B

Rule 3:
IF ELECTRICAL ENERGY CONSUMPTION > 700
AND
IF HOUSE AREA > 250 SQUARE METERS
AND
IF CUSTOMER REGIONAL GROUP = A
THEN BILLING POLICY = C

Rule 4:
IF ELECTRICAL ENERGY CONSUMPTION > 700
AND
IF HOUSE AREA > 250 SQUARE METERS
AND
IF CUSTOMER REGIONAL GROUP = B
THEN BILLING POLICY = D

Rule 5:
IF ELECTRICAL ENERGY CONSUMPTION > 700
AND
IF HOUSE AREA > 250 SQUARE METERS
AND
IF CUSTOMER REGIONAL GROUP = B
AND
IF CURRENT TIME = DAYTIME
THEN BILLING RATE COEFFICIENT = 2.5

Rule 6:
IF ELECTRICAL ENERGY CONSUMPTION > 700
AND
IF HOUSE AREA > 250 SQUARE METERS
AND
IF CUSTOMER REGIONAL GROUP = B
AND
IF CURRENT TIME = NIGHTTIME
THEN BILLING RATE COEFFICIENT =1.5

Rule 7:
IF ELECTRICAL ENERGY CONSUMPTION > 700
AND
IF HOUSE AREA > 250 SQUARE METERS
AND
IF CUSTOMER REGIONAL GROUP = B
AND
IF CURRENT TIME = NIGHTTIME
AND
IF SEASON = SUMMER
THEN BILLING RATE COEFFICIENT =1.2

Rule 8:
IF ELECTRICAL ENERGY CONSUMPTION > 700
AND
IF HOUSE AREA > 250 SQUARE METERS
AND
IF CUSTOMER REGIONAL GROUP = B
AND
IF CURRENT TIME = NIGHTTIME
AND
IF SEASON = WINTER
THEN BILLING RATE COEFFICIENT = 2.8

Rule 9:
IF ELECTRICAL ENERGY CONSUMPTION > 1000
AND
IF HOUSE AREA > 500 SQUARE METERS
AND
IF CUSTOMER REGIONAL GROUP = C
AND
IF CURRENT TIME = NIGHTTIME
AND
IF SEASON = WINTER
AND
IF TIME IS BETWEEN 12 AM TO 6 AM
AND IF CUSTOMER PLAN = SAVINGS
THEN BILLING RATE COEFFICIENT = 1

Rule 10:
IF ELECTRICAL ENERGY CONSUMPTION > 1000
AND

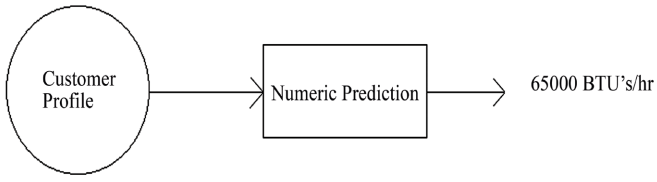
IF HOUSE AREA > 500 SQUARE METERS
AND
IF CUSTOMER REGIONAL GROUP = B
AND
IF CURRENT TIME = NIGHTTIME
AND
IF SEASON = SUMMER
AND
IF TIME IS BETWEEN 12 AM TO 6 AM
AND IF CUSTOMER PLAN = REGULAR
THEN BILLING RATE COEFFICIENT = 2.05

As shown above, several rules similar to those shown above (more detailed and specific) can be applied on the customer data and profiles.

Data mining can be also used to predict what the consumption of electricity will be in the upcoming based on several parameters such as current annual consumption, past consumption, population growth expectancy and availability of resources.

Based on analysis of existing data such as annual electric consumption, historic data related electric energy consumption such as usage over the last 50 years including anomalies, trends, etc., and the growth rate of the current population, data mining can successfully predict the possible electrical energy consumption in the years to come.

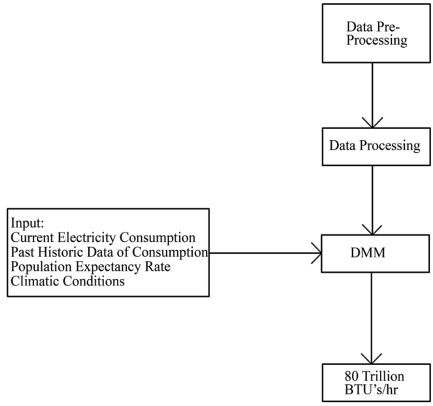
This can be done with the help of statistical machine learning algorithms. For example, based on a customer profile, data mining classification algorithms can predict how much electrical energy he will consume in a month:



The same logic, as shown above, is applied to the existing data and a numerical prediction is made based on the input fed to the data mining

algorithms. Different algorithms such as neural networks, statistical algorithms, Naive Bayes algorithm, regression algorithms, etc., can be used.

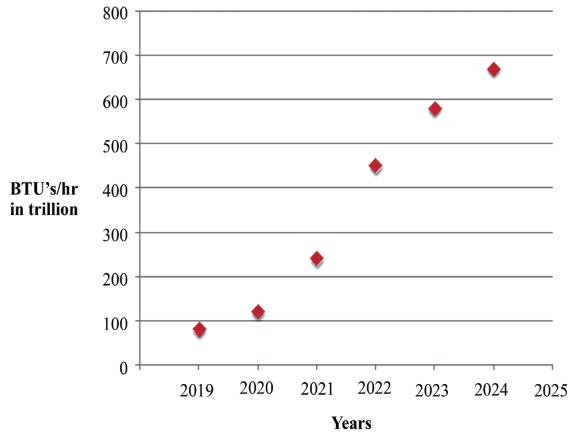
The following flowchart shows the steps followed in order to predict energy annual consumption:



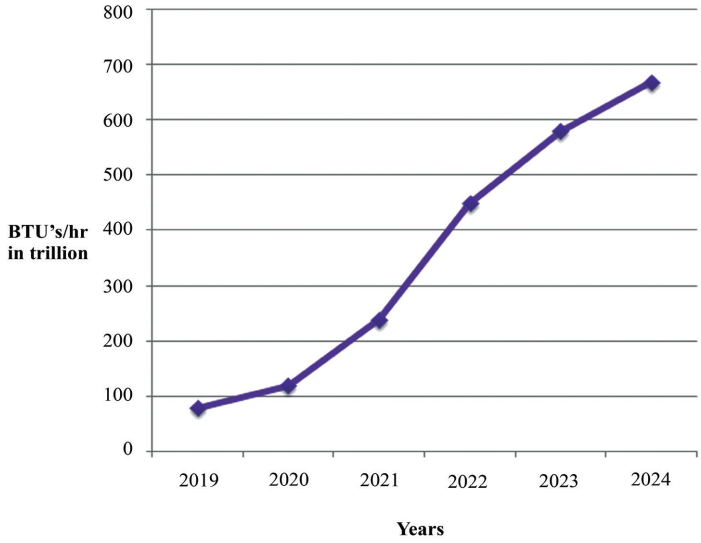
DMM: Data Mining Methods

The above diagram shows the steps followed in the use of prediction approaches and predicts the consumption of a country for the year 2019 based on the data of past years along with additional input parameters.

Similarly, the same algorithm can be applied to predict consumption of several years as shown below:



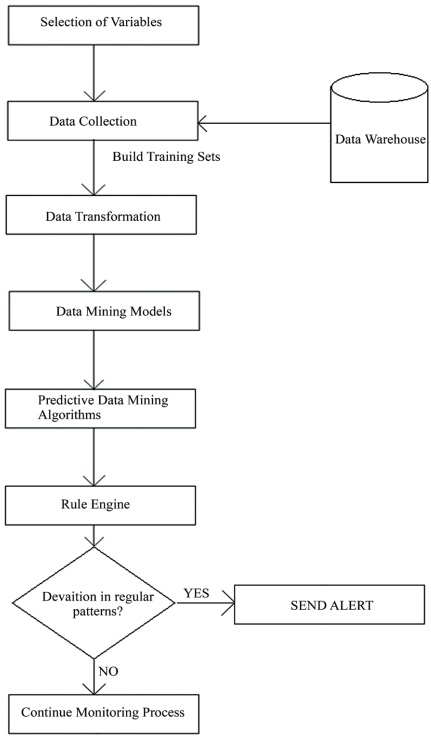
The same data can be represented as follows.



Data mining can additionally be used to predict anomalous behavior such as power outages, disasters, technical difficulties and so on. This is done with the help of outlier detection. For instance, if a particular industry generally makes use of let us say 1 million BTU's/hr on a monthly basis, and suddenly one month they use 20 times the amount. This could be due to a leakage of electricity, or due to some faulty equipment and the concerned parties need to be alerted in such a scenario.

Data mining can help existing monitoring systems by performing analysis of the usage data for anomalous behavior so as to reduce potential loss of resources and ensure safety.

The following flowchart shows the process followed in the event of a deviation when the monitoring is done with the help of data mining.



Rules associated to anomalous or odd behavior can be set up as well. There could exist several rules such as if energy is not generated anymore, if too much energy is being generated, etc.

Researchers have successfully developed models that predict and help in preventing a possible power outage. Generally, it is known that really strong winds during a thunderstorm may cause trees neighboring an electric grid to fall and crash into an electric grid which leads to power outages. Not much is being done to reduce these type of accidents except for scheduling trimming of trees. The cost of arranging and managing a tree cutting operation is costly and the effort that is required is enormous and sometimes, in some regions this is done on a rotational basis which may take months and sometimes even years before all the trees are trimmed.

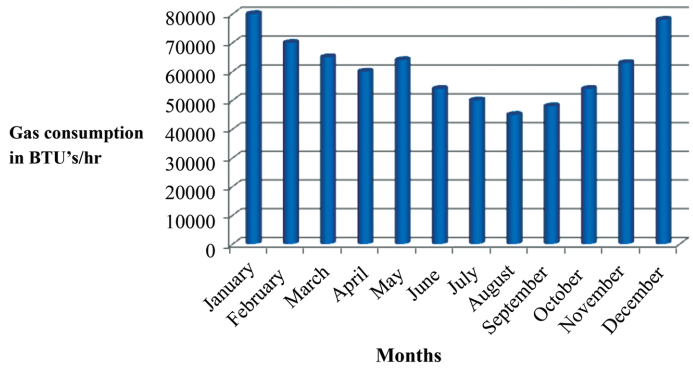
Texas A&M University researchers have developed a model that is capable of predicting a potential vulnerability to the utility assets and present a map of where and when a possible outage may occur. The predictive feature allows the trees in the most critical areas with the highest risk to be trimmed first (Communications, 2018).

2.11.2. Application in the Gas Consumption

Data mining can also be applied to the data related to gas consumption as well. Similar to the consumption of electricity, the consumption of gas is subject to a lot of parameters such as the type of property, the age of the property, how big the property, type of usage of gas, type of heating/cooling, how old the boiler is if any, how hot water is obtained (heating using gas or electricity), region of habitation, etc. All this information is stored by the gas company for informational purposes but is actually never used a lot. Recent times have seen some amazing growth in the field of data mining. Hence, the energy industry is taking an active interest in this field as well. The gas industry could profit greatly with estimations and predictions regarding the consumption of gas.

Although there exist several ways to predict the usage profiles of customers, most of the times it is only valid for a part of the population and not most of the group. Due to generic formulas based on fixed parameters, the gas industry is sometimes unable to get a tighter window of estimation for the consumption of gas which can be inconvenient at times. This is where data mining can play a key role in discovering hidden gems of knowledge regarding the customers.

The following graph shows the consumption of LPG on a monthly basis in BTU's/hr for a small studio apartment:



This data can be mined and rules can be formed using association rule mining that help better understand the customer and his needs. This may help predict the customer's future needs and anticipate his next demands such as a better gas providing and billing plan, a plan that charges based on usage only without extra fees, flexible plans, etc.

From the above data, the following simple rule can be extracted:

IF MONTH = JULY OR AUGUST OR SEPTEMBER

THEN GAS CONSUMPTION < 50000 BTU's/hr

Another rule that can be extracted is the following:

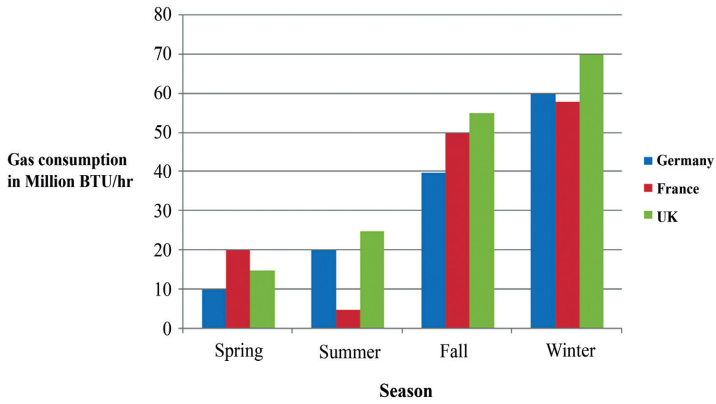
IF MONTH = DECEMBER OR JANUARY

THEN GAS CONSUMPTION > 70000 BTU's/hr

These rules can help develop better customer usage profiles and help the management to look for ways to reduce consumption such as different sources of gas, different consumption rates, levels, etc.

Data mining can be used to consolidate data and present it in a visual manner. The gas companies hold data of gas consumptions across countries and sometimes, this global level data can be mined to form rules.

Consider the following graph that depicts the information consolidated by data mining that shows energy consumptions of different countries at different times of the year:



From the above data, association rules can be extracted such as the following:

IF COUNTRY = FRANCE

AND

SEASON = SUMMER

THEN GAS CONSUMPTION < 10 MILLION BTU/hr

Based on this rule, the gas company can provide better rules for its customers that are customized for the users in France.

Another simple rule that can be derived from the sample data is the following:

From the above data, association rules can be extracted such as the following:

IF COUNTRY = UK

AND

SEASON = WINTER

THEN GAS CONSUMPTION >= 70 MILLION BTU/hr

This rule can help the gas company develop special offers and plans for the winter season for the customers residing in the UK.

Another rule that can be extracted from the above data is the following:

IF SEASON =SPRING

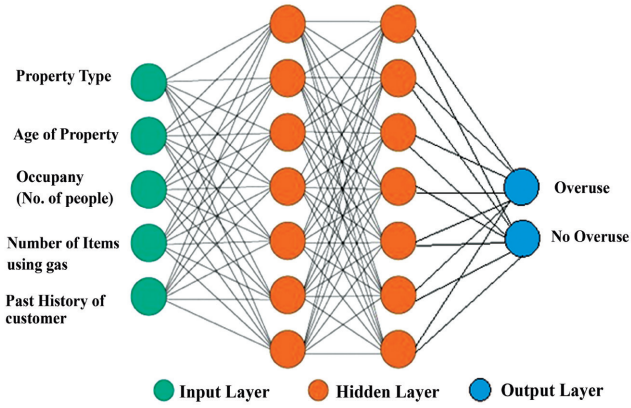
THEN GAS CONSUMPTION <= 20 MILLION BTU/hr

This rule is a global rule that is applicable to all the 3 countries. The gas company can use this rule to further thinks of ways of maintaining low gas consumption for the Spring season for the years to come in the future.

Several data mining methods can be used to predict the usage of a customer. Neural networks can be used to predict if a customer will over-consume a certain quota of gas allocated to him or not.

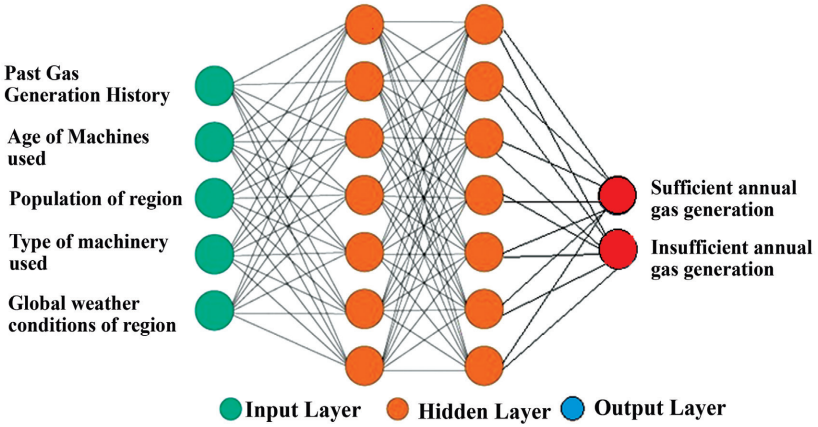
This can be calculated based on several parameters such as past history of consumption, type of property of the customer, type of gas usage, occupancy, age of the property, number of items using gas for heating, etc.

This is shown below:



A neural network can take in the input parameters and based on its internal reasoning and assigning weights to each and every parameter, predict whether a customer will overuse his quota of gas on not. Predictive modeling can also be done on the larger scale where data mining can be used to predict whether the gas power plant in a particular region will be able to generate enough gas for the annual year on not based on specific input parameters such as past consumption history, population, likely weather conditions, type of machinery used, age of machinery, etc.

This is shown below:



As shown above, a neural network with two hidden layers is built which can be used to predict whether a gas power plant can generate sufficient gas for the next annual year based on its current and past generation history and four other parameters that have a significant effect on the decision and the output of the neural network.

Data mining can additionally be used to study the customers' gas consumption data and find meaning patterns and correlations among several input parameters such as property type, occupancy, etc.

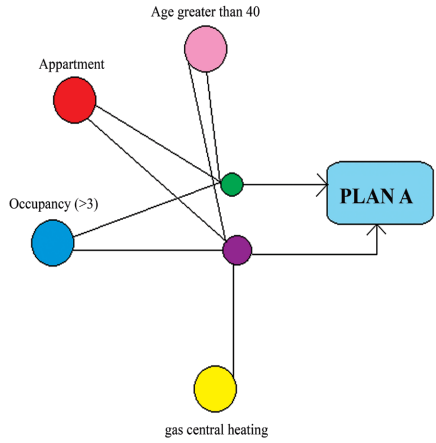
For example, consider the following rules mined from gas consumption data of customers:

Rule Number	Rule
1	{apartment, 40 years old, 3 people} -> {Plan A}
2	{bungalow, 6 years old, 7 people} -> {Plan B}
3	{Studio, 1 year old, 1 person} -> {Plan C}
4	{apartment, 45 years old, 4 people, gas central heating} -> {Plan A}
5	{Bungalow, Illinois, 1 year old, 3 people, electric central heating} -> {Plan C}
6	{Bungalow, Alabama, 0 year old, 4 people, gas central heating} -> {Plan C}
7	{Bungalow, 5 year old, 7 people, gas central heating} -> {Plan B}
8	{Studio, Tennessee, 1 year old, 1 person, electric heating, Electric Boiler} -> {Plan C}
9	{Studio, Philadelphia, 3 year old, 2 people, no heating} -> {Plan C}
10	{apartment, 1 year old, 1 people, electric heating} -> {Plan C}

The above table shows different plans chosen by the customers and gives us specific rules as to what parameters entail the selection of a particular plan by the customer.

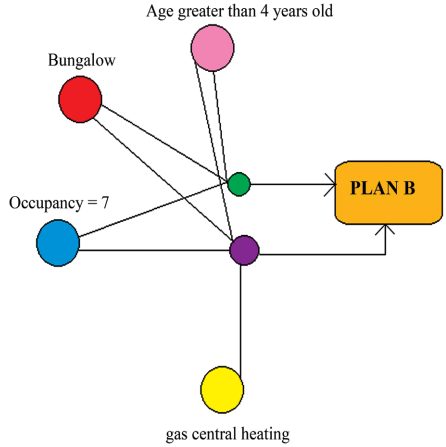
The rules mined from the customer gas consumption data give us interesting tidbits of information that can be used to develop strategies to retain customers and maybe attract new ones by improving upon existing gas consumption plans and services.

The rules can be represented in a visual manner as well. This is shown below:



{apartment, 40 years old, 3 people} => {Plan A}
{apartment, 45 years old, 4 people, gas central heating} => {Plan A}

As shown in the above image, rules 1 and 3 are represented visually which may be more useful in transmitting the information to the experts as the rules can be instantly read and are easy to comprehend. The same can be done for plan B as shown below:



{bungalow, 6 years old, 7 people} => {Plan B}

{bungalow, 5 years old, 7 people, gas central heating} => {Plan B}

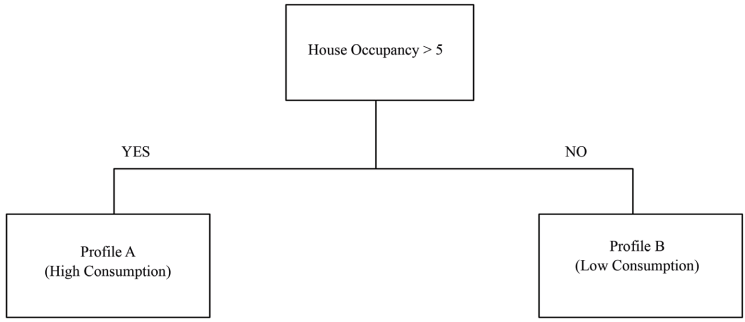
Another possibility is the use of classification algorithms that can be used to classify people into different profiles based on their consumption parameters such as property type, occupancy, frequency of usage, etc.

There exist different types of consumption ranges for different people.

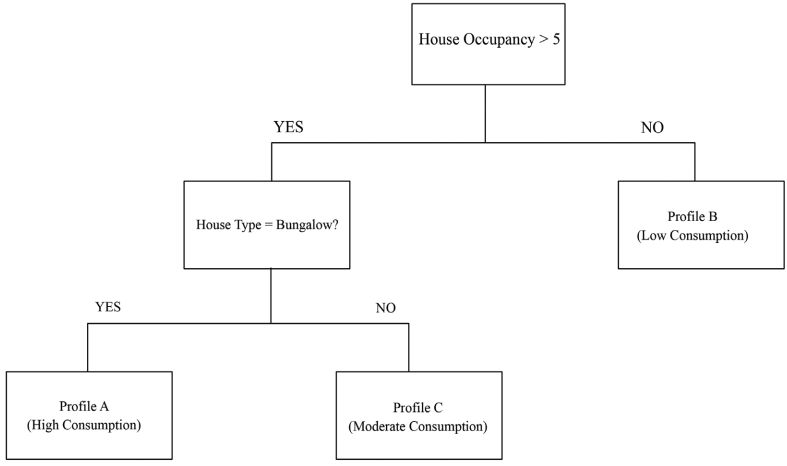
For instance, if there are five people living in a bungalow that makes use of gas central heating and has a boiler that runs on gas, then the chances are that this particular house consumes a lot of energy as compared to single occupancy studio with electrical heating.

Clearly, the usage in both cases varies a lot and it may be of interest to the gas companies to know what factors contribute in high consumption rates.

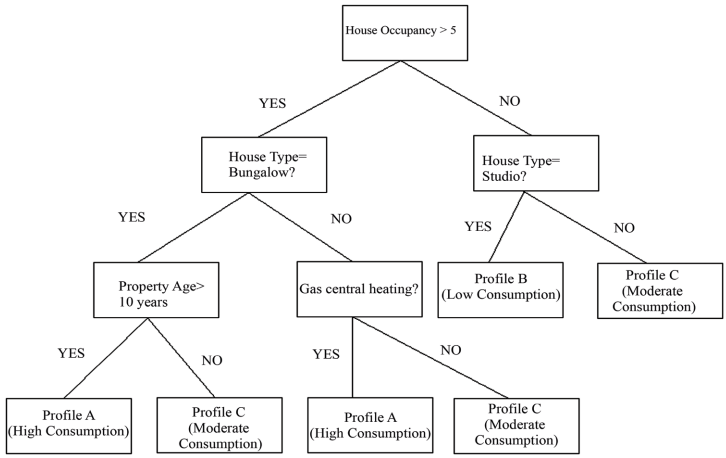
A simple example of classification of customers is shown below:



Below is a more detailed classification of customers that makes use of another parameter for classification namely the type of property:



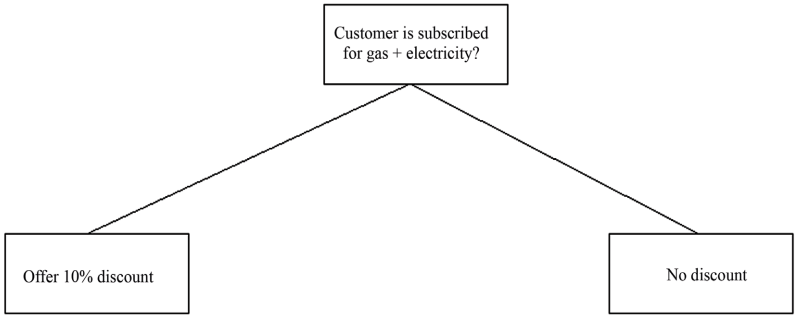
Further, additional parameters can be used to classify the customers into different profiles as shown below:



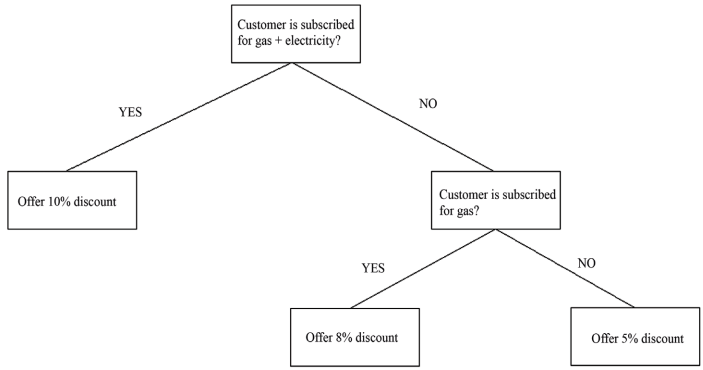
In addition to the analysis of electric and gas consumption separately, sometimes, they can be done together as well. This can happen when the same energy company provides both electricity and gas services.

In such cases, customer data can be mined and simple policies can be developed based on the customer plans. For example, if a customer is a member that uses only the gas services, his profile data may be mined and he may be provided incentives to switch his electric company and vice versa with gas companies. Additionally, loyalty programs can be introduced for those customers that use both electric and gas services.

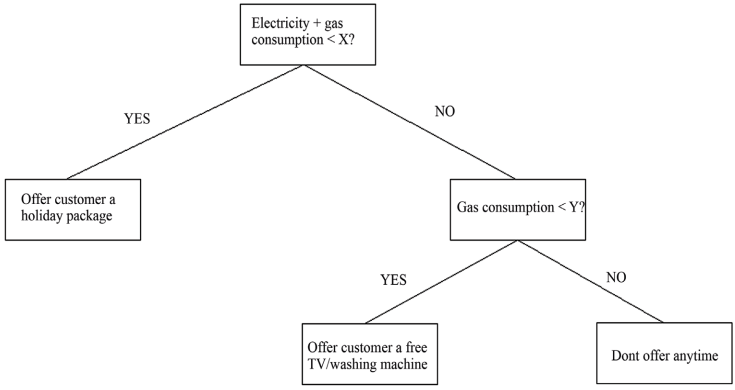
A simple decision tree-based classification is shown below:



The above decision tree can further be modified to make decisions based on additional parameters as shown below:



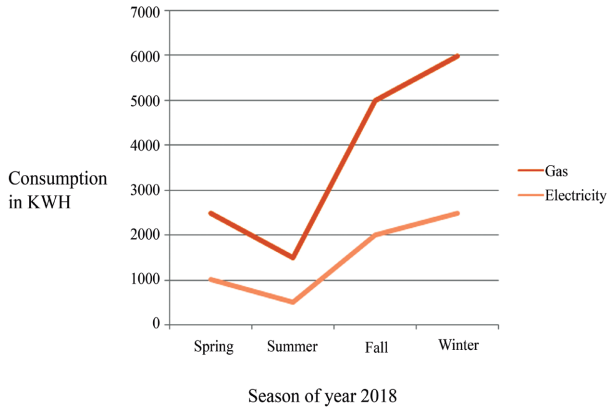
Another possible way is providing incentives to customers to reduce their energy consumption. This can done by setting a particular threshold and if the customers manage to stick to the threshold, the energy company can provide them different bonuses based on their spending habits.



As we can see, rewards and gifts are offered when the customer saves energy thereby motivating the customer to spend his energy resources in a more responsible manner.

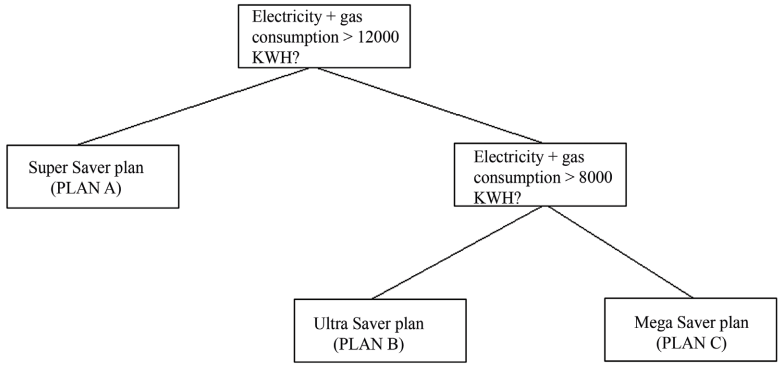
Furthermore, based on the joint electrical and gas consumption of customers, different types of service plans can be proposed to them so as to optimize the energy savings and billing costs for the customer and also to enable customer retention.

Consider the following sample data set that consists of electrical and gas consumptions in Kilowatt hours.

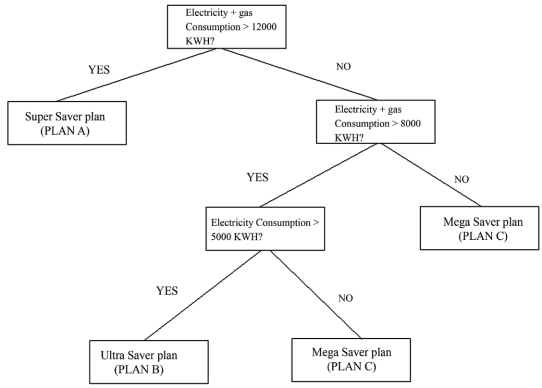


The above data shows the energy consumption data for a customer for the year 2018 across different seasons. This data can further be used to make informed decision and propose the client different and possibly customized service plans based on his spending habits.

One such policy is derived with the help of decision trees as shown below:



Another possible classification based on refinement of the decision tree is as shown below:

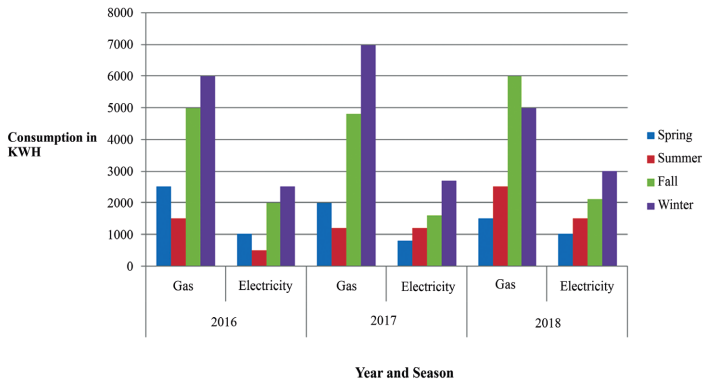


The customer consumption data of both electricity and gas can be also used to predict the customer's future utilization of these resources.

Consider the following sample data set consisting of a customer's usage of gas and electricity for a time period of 3 years:

Year	2016		2017		2018	
Season	**Gas**	**Electricity**	**Gas**	**Electricity**	**Gas**	**Electricity**
Spring	2500	1000	2000	800	1500	1000
Summer	1500	500	1200	1200	2500	1500
Fall	5000	2000	4800	1600	6000	2100
Winter	6000	2500	7000	2700	5000	3000

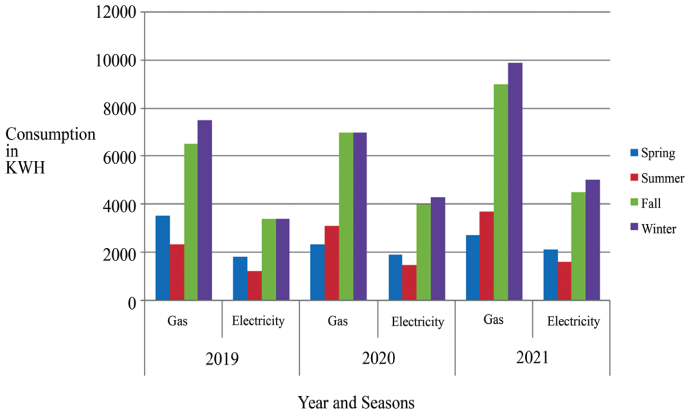
This data can be represented in a visual format as shown below:



Consider the above sample customer data, Based on this data the customer's potential data usage for the upcoming years can be predicted. The data-mining algorithm takes in the above data and also analyzes other

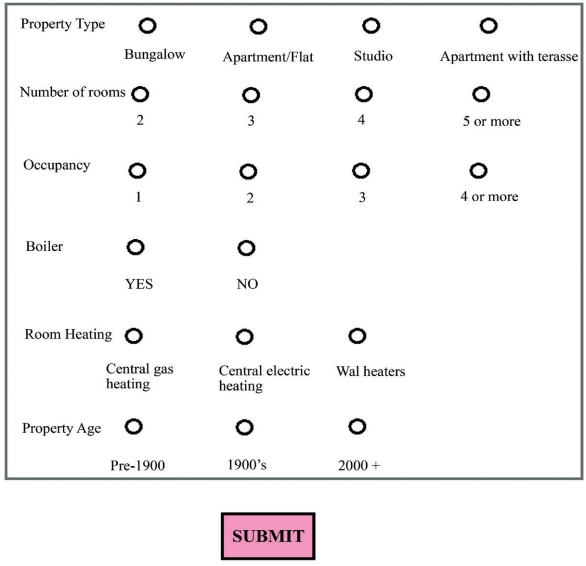
customer data such as his profile, type of property, etc. and applies predictive analysis to predict the values for the next 3 years in this case.

This is shown below in a visual manner:



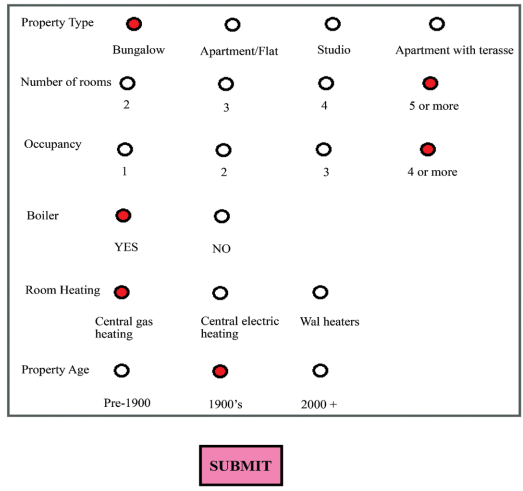
Additionally, with the use of data mining several online prediction tools have been developed to predict the total energy consumption based on a several range of input parameters.

A simple User Interface is shown below:



The above data can be used to give predicts to customers regarding their usage of electricity and gas.

Consider the following data that is input by an existing or potential customer:



The customer enters his profile details such property type, its age, the type of heating, occupancy etc. and on submission of the form, he gets an estimation of the expected consumption of gas and electricity for the annual year.

The sample form shown above makes predictions based on the parameters that are presented as input to the use. This software is able to do this task as it has studied large sets of customer data and developed rules that help predict outcomes.

In the background of this user interface, there exist several rules. Some of them are shown below:

Rule 1:

IF PROPERTY TYPE = BUNGALOW

AND

IF PROPERTY AGE = 1900

AND

IF NUMBER OF ROOMS > = 5

AND

IF OCCUPANCY >= 4

AND

IF BOILER = YES

AND IF ROOM HEATING = CENTRAL GAS HEATING

THEN TOTAL CONSUMPTION (Gas, Electricity) = {25345, 14587} KWH

Here, the outputs are calculated with the help of a formula that takes into account the charges of gas and electricity per hour, the energy consumption of central gas heating, the consumption of boiler, etc.

Rule 2:

IF PROPERTY TYPE = APARTMENT

AND

IF PROPERTY AGE = 1900

AND

IF NUMBER OF ROOMS = 3

AND

IF OCCUPANCY = 2

AND

IF BOILER = YES

AND IF ROOM HEATING = CENTRAL GAS HEATING

THEN TOTAL CONSUMPTION = {18431, 9045} KWH

Rule 3:

IF PROPERTY TYPE = STUDIO

AND

IF PROPERTY AGE = 1900

AND

IF NUMBER OF ROOMS = 2

AND

IF OCCUPANCY = 1

AND

IF BOILER = YES
AND IF ROOM HEATING = CENTRAL GAS HEATING
THEN TOTAL CONSUMPTION = {12045, 4570} KWH

Rule 4:
IF PROPERTY TYPE = APARTMENT WITH TERASSE
AND
IF PROPERTY AGE = 1900
AND
IF NUMBER OF ROOMS = 3
AND
IF OCCUPANCY = 3
AND
IF BOILER = YES
AND IF ROOM HEATING = CENTRAL GAS HEATING
THEN TOTAL CONSUMPTION = {20090, 10080} KWH

Rule 5:
IF PROPERTY TYPE = BUNGALOW
AND
IF PROPERTY AGE = Pre-1900
AND
IF NUMBER OF ROOMS = 5 OR MORE
AND
IF OCCUPANCY = 4 OR MORE
AND
IF BOILER = YES
AND IF ROOM HEATING = CENTRAL GAS HEATING
THEN TOTAL CONSUMPTION = {25790, 16190} KWH

Rule 6:

IF PROPERTY TYPE = APARTMENT
AND
IF PROPERTY AGE = Pre-1900
AND
IF NUMBER OF ROOMS = 3
AND
IF OCCUPANCY = 2
AND
IF BOILER = YES
AND IF ROOM HEATING = CENTRAL GAS HEATING
THEN TOTAL CONSUMPTION = {21420, 10340} KWH

Rule 7:
IF PROPERTY TYPE = STUDIO
AND
IF PROPERTY AGE = Pre-1900
AND
IF NUMBER OF ROOMS = 2
AND
IF OCCUPANCY = 2
AND
IF BOILER = NO
AND IF ROOM HEATING = CENTRAL GAS HEATING
THEN TOTAL CONSUMPTION = {14350, 8515} KWH

Rule 8:
IF PROPERTY TYPE = APARTMENT WITH TERASSE
AND
IF PROPERTY AGE = Pre-1900
AND
IF NUMBER OF ROOMS = 4

AND
IF OCCUPANCY = 4
AND
IF BOILER = NO
AND IF ROOM HEATING = CENTRAL ELECTRIC HEATING
THEN TOTAL CONSUMPTION = {19760, 14580} KWH

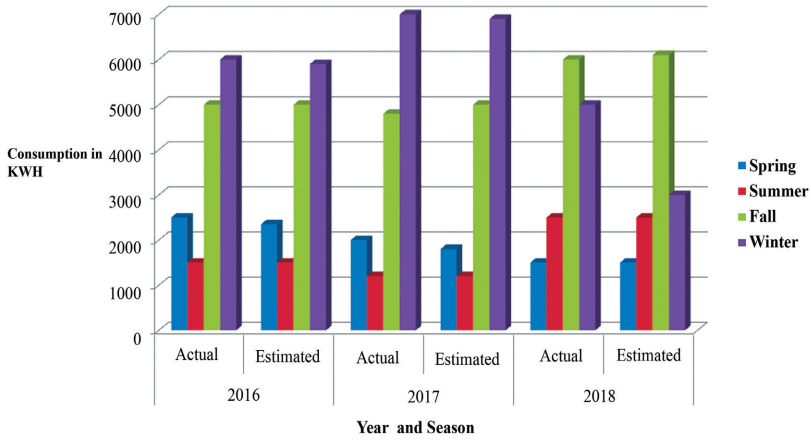
Rule 9:
IF PROPERTY TYPE = APARTMENT WITH TERASSE
AND
IF PROPERTY AGE = 1900
AND
IF NUMBER OF ROOMS = 3
AND
IF OCCUPANCY = 2
AND
IF BOILER = YES
AND IF ROOM HEATING = CENTRAL ELECTRIC HEATING
THEN TOTAL CONSUMPTION = {22490, 12870} KWH

Rule 10:
IF PROPERTY TYPE = APARTMENT
AND
IF PROPERTY AGE = 1900
AND
IF NUMBER OF ROOMS = 2
AND
IF OCCUPANCY = 2
AND
IF BOILER = NO
AND IF ROOM HEATING = WALL HEATING

THEN TOTAL CONSUMPTION = {17760, 15086} KWH

Another possible application of data mining in the domain of electricity and gas related is the possibility of self learning. Data mining can compare the predictions made by its model to the actual values that were generated and then learn from the discrepancies to improve upon its algorithms.

A sample data set that compares the actual and predicted values is shown below:



2.11.3. Solar Power Energy Prediction

The sun is a formidable source of natural energy. The rays of the sun, also known as sunlight can be used for a lot of purposes such as heating, electricity generation, lighting, generation energy for cooling systems, in large-scale industries and so on. The sun generates energy in the form of radiation and is one of most natural and clean energy resources available to mankind. This solar energy in the form of radiation can be a source of generating new renewable energy. Sunlight is a useful resource that is freely available and is a source of renewable energy that helps reduce costs of operating large fossil fuel generation plants, reduces carbon emissions and other deadly emissions too.

The use of solar energy as a renewable alternate source of energy is a growing field of research in the energy domain. The research has led to development of solar systems that have immensely helped the energy industry in their quest to generate clean environment friendly forms of energy. Solar power plants have proved to have a lot of advantages such as reduction of the environmental impact of the use and combustion of fossil

fuels, reduction of emissions of harmful gases, greenhouse gases, etc. Solar systems are proven to have extremely low emissions of pollutants such as carbon monoxide, sulfur dioxide, volatile organic compounds, nitrous oxides and the greenhouse gas as compared to traditional fossil fuel generation plants which make solar systems a better, clean and more safe energy option. Energy industries are investing a lot into the research of development of efficient solar systems due to its highly promising advantages.

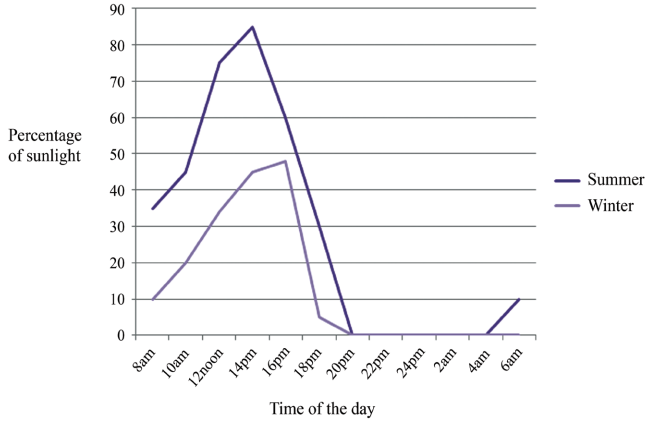
But, sunlight or solar energy is not always a reliable factor. It is not like a machine that one can switch on and off. The availability of sunlight and the radiation waves is dependent entirely on the weather conditions and geo-location. The solar systems are functional only if there is solar energy and may vary on a daily or hourly basis. For instance, if the sun sets at 6 p.m. in a particular country, the solar energy can no longer be used. But, this might not be true for another country where the sun sets at 8 p.m. in summers.

The data related to the availability is highly volatile and is subject to constantly changing weather conditions. Additionally, the cost of setup of a solar system on a small scale is a little high and customers may not be too keen to invest their money in an unreliable energy system. Energy companies constantly find it is difficult to sell solar energy as a possible option to their customer due to its unreliable nature and cost. This is why for the development of solar powered systems and to guarantee its success there is an undeniable need for prediction of solar radiations. Recently, the energy domain is actively looking at data mining as a solution for prediction of solar energy and development of models that are reliable and sustainable. However, if we want reap the benefits of solar power to its maximum potential; there should be accurate means of prediction of solar radiation power. In order to get the predictions, many applications were introduced but the energy industry still prefers the use of data mining for the predictions.

Data mining makes use of its predictive models such as neural networks, statistical algorithms, regression algorithms and so on to predict the power of sunlight. This prediction and the forecast are done by learning the existing historic data that has been captured by the energy industry regarding the solar radiation levels across the world. Data mining gives the possibility of predicting frequently occurring patterns in the available sunlight and its power on the basis of parameters such as wind resistance, solar radiation power etc. A regular pattern that is quite frequent and obvious is the inefficient use of solar energy in the early morning time. This is a problem that can be fixed.

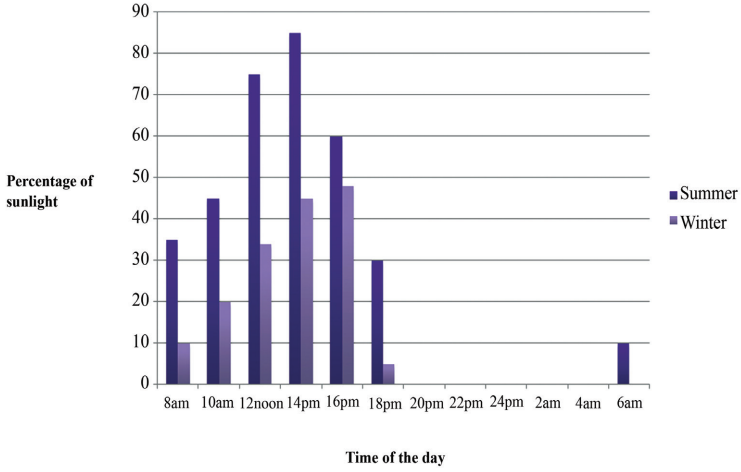
Data mining can be used to effectively predict the power of the solar radiation based on the time of the day in different seasons.

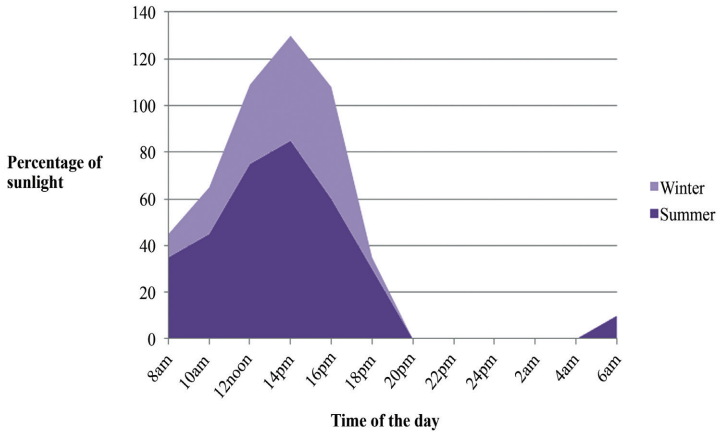
A sample data set that predicts the solar power in percentage is shown below:



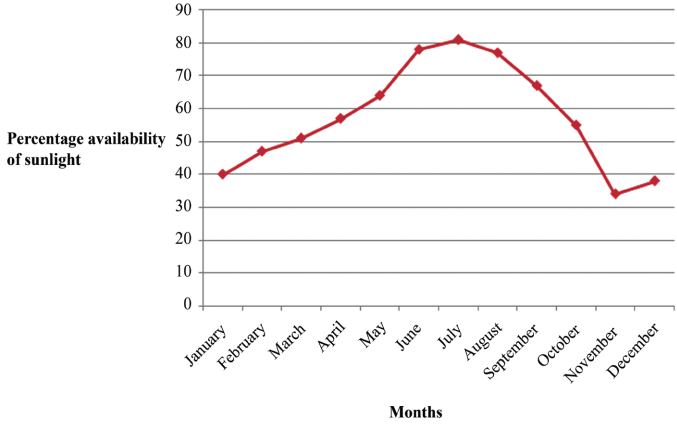
As shown above, the strength of the solar power in different seasons is predicted for a particular day in a particular region. Here, we can see that in the wee hours of morning there is some light that can be utilized.

This information can also be represented in following ways:



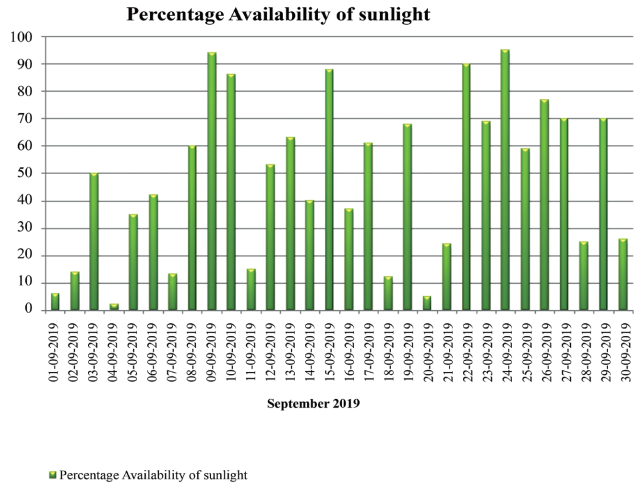


Predictions can be done on a monthly basis as well. Here, the output is the total percentage of availability of sunlight throughout the month as an average.



The same prediction can be done on a more fine-grained level such as for each day of the month.

The following graph shows the availability of sunlight for each day in the month of September for the year 2019.

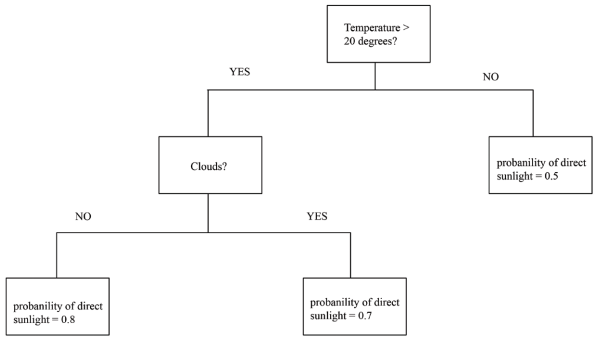


Predictions of this type can be done on a large-scale for the entire annual scale, which will help the energy industry utilize the available solar energy and further evaluate total costs. With the help of predictive data like this, the energy industries are capable of predicting valid plans to its customers that make use of solar energy for heating purposes.

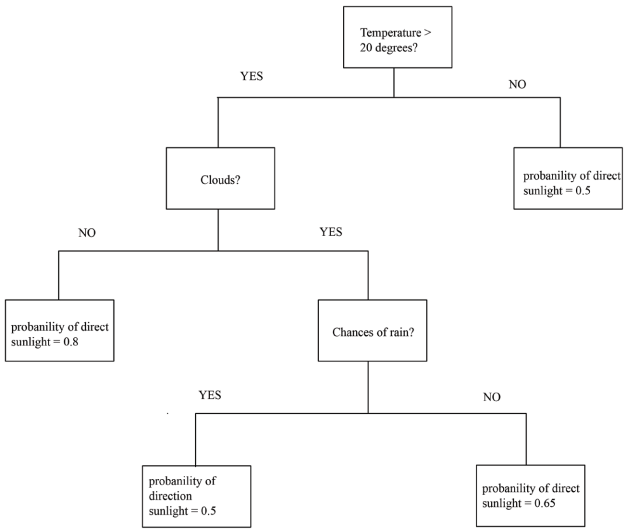
The energy industry will benefit immensely from predictions of this type as now they have concrete predictions that can be put forth to their customers. The executives of energy companies can work on the development of an action plan to market solar systems. They can provide solar system options in combination with traditional options so as to ensure that the customer has an alternate option on days when the power of sunlight is less.

Another way of classification of existing geographical and solar energy data can be done on the basis of the climatic conditions of the region under consideration.

A simple sample decision tree that makes decisions based on factors such as temperature and wind resistance to determine the power of sunlight is shown below:

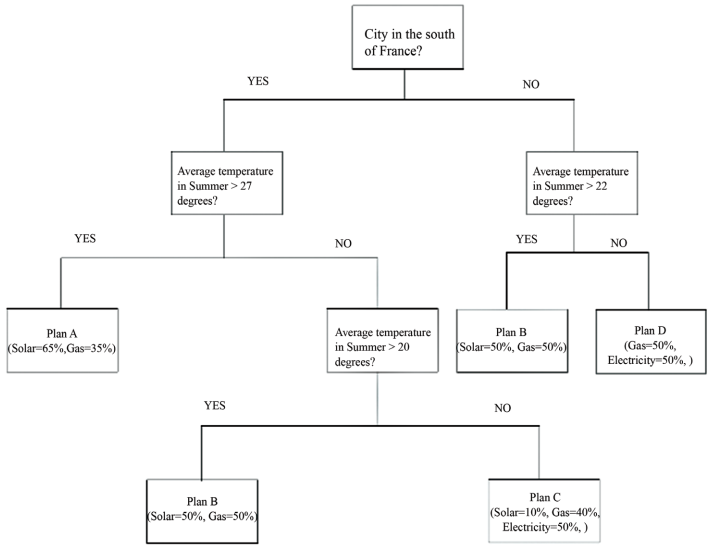


This decision tree can further be enhanced as follows:



In addition to predicting the probability, the type of climate of the region can play a part in the decision factors as well. This can help develop different kinds of service plan to different customers belonging to different regions based on their respective geographical locations.

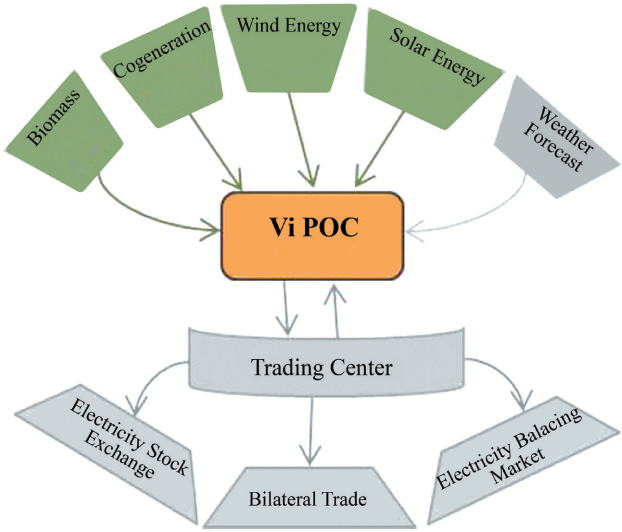
This is shown below:



The above graph shows a sample decision tree that suggests services plans for the use of solar energy in country of France. The average temperatures in summer are used as a decision parameter for selection of the most appropriate plan for the region under consideration.

There exist several implementations that use data mining successfully for the management of renewable energy sources. One such implementation is the Vi-POC (Michelangelo et al., 2014) that makes use of data mining algorithms to aid energy producers in their decision-making and management capabilities. Vi-POC is fully called as the Virtual power operating center. It is a framework for collecting, storage, analysis and querying data that is acquired from energy plants that are developing renewable energy such as wind energy, geothermal energy, photovoltaic, etc. (Michelangelo et al., 2014).

This approach makes use of predictive algorithms and adaptive data mining algorithms to predict outcomes. The architecture of this system is as shown below:



(Michelangelo et al., 2014).

The above image demonstrates a high-level view of this system. The system gathers data from a wide range of renewable energy power production plants such as biomass, solar power, wind, geothermal etc. In addition to this data, weather forecasting data is also fed into the system. The purpose of feeding weather forecast data is to enrich the data obtained from the different renewable energy power plants.

Vi-POC analyses the data that it collected and processes the data by applying prediction algorithms on the data. The results are sent into different outputs such as trading centers that make use of this information for helping their energy buying and selling processes. They can also use this information for development of strategies for the future based on existing data.

Vi-POC strives to get accurate results as the results generated can impact a lot of financial decisions. This is done by monitoring of production sites on a regular basis and collecting data regularly. This system strives to create an environment that is automatized and can help in real-time monitoring operations as well.

2.12. EDUCATION

The rise in the number of educational institutes all over the world is staggering. The educational sector is looking at an exponential rate of growth and hence they are becoming more and more complex over time. They have the same

level of complexity as any enterprise in the other industries. They are now responsible to not only provide good level education, but also manage a number of activities such as marketing, advertising, student management, recruitment, planning, coordination, etc. Institutes at that teach at bachelor level and above may also be responsible for organizing campus placements for its students. Also, they need to manage the regulatory, financial and statutory aspects of running an educational institute. They need to ensure smooth running of operations and be well prepared for all possible scenarios of risk. They need to ensure safety of their students and their staff and also make sure the institute is well guarded from possible external threats. They also need to handle the cash flow, financial aspects, planning for the next years and so on. They need to prepare for the admissions process and make sure that everything is being handled in an efficient and fair manner. Regulatory checks of security, cleanliness and accounts needs to be done as well because they are responsible for the physical and mental health of their students and their staff. They are responsible for managing financial aid for students and deciding the basis on which they provide aid and hence are always in need of donations from organizations. This indirectly means that each institute needs to develop and manage a network of connections and alumni's that would be willing to help the institute.

On top of all the administrative and managerial responsibilities, the institutes face a lot of competition from other equivalent institutes that could offer more incentives to students for enrolling in their institute. Additionally, as the number of educational institutes keeps on growing at a rapidly fast rate, these institutes are subject to a lot of demands from both the students and the corporate domain as well. They need to ensure that their institute stands out in the crowd and make sure that it is run efficiently. They need to employ modern management practices in order to stay afloat in the current economic conditions. In addition to better managerial practices and method, these institutes are in constant need of state of the art technologies to keep up with the growth in the technology and computer industry.

Hence, recently we have seen the rise of automated software solutions that help the administrative process. There exist many software solutions for educational institutes these days. These softwares serve to help these institutes mange their internal and external operations effectively. One such example is the use of websites for promotion and admissions. All colleges have started a process of developing websites that are like paper brochures highlighting the facilities provided by the institute, their capacity, etc. Also, the admission process is now online. Potential applicants are now invited

to send in their applications via a website containing an array of forms that ask for personal information such as their marks, activities, their desired course and so on. They are then asked to upload documents online such as their mark sheets, birth certificate, passing certificate, etc. The dossier of the applicant is then transmitted to the institute where each applicant's case is studied by someone and then a decision is made and sent via email. Hence, the entire process of enrollment is now online.

Furthermore, the postacceptance process (procedures to be followed after a student has accepted to enroll in the institute) is also done online. Here, student is asked to upload additional documents, pay fees, assigned a number and asked to choose his elective classes, etc. The whole process is now done via the online world using internet.

In addition to the student enrollment process, the finance and the bookkeeping aspects are also handled online. The management of student fees, the staff enrollment process, the management of scholarships, donations, alumni, etc. is done online by educational institutes.

Although, the development of softwares for automation purposes has improved the life of educational institutes by automating their operations, it is not enough for institutes to stay afloat. Despite having state-of-the-art technology and resources, these institutes still continue to face a lot of challenges in terms of marketing, student retention, prediction of student growth, making informed managerial decisions, etc. They are unable to predict whether a student will enroll in their university, whether a student will perform well, whether a student will drop out, etc. These institutions are eager to predict the paths of the students, the alumni, predicting what courses will be taken up by students, the number of students that will enroll, etc. One of the biggest hurdles faced by this industry is the explosion of educational data and information that is being stored and that keeps growing without being put to too much use. The institutes are in need of making better decisions and improve the quality of their service in order to make profits and ensure retention. Hence, the use of data mining seems to be an ideal solution as it can find meaning in the sea of information. The process of data mining can be applied to higher education institutes and be used to help them retain students, improve their performance, their management processes, lower the risk of attrition and manage the funds. Several data mining techniques have been successfully employed to this effect.

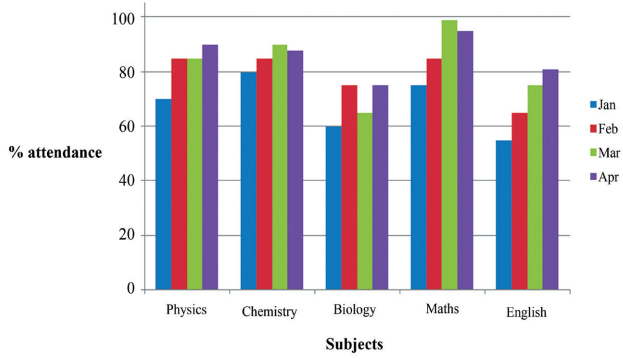
We shall now take a look at a few data mining techniques that can be applied on student information stored by educational institutes. The

employment of data mining in the field of education is called as Educational Data Mining (EDM) (Scheuer & McLaren, 2012). It is a branch of research that is dedicated to the development of methods and solutions that explore the distinctive type of data that is seen in an education setting and the employment of those methods and solutions to understand the students in a more detailed and better manner and understand the settings that they work, grow and learn in. The main areas of focus of this discipline are the mining of the performance of students, the enrollment data and improvement of the current processes of educational institutes. EDM is a learning science and an area of research of data mining.

One of the primary uses of data mining is visualization. Data mining helps consolidate the student data and represent it in a visual manner that can be further used by the board to make decisions and highlight important information.

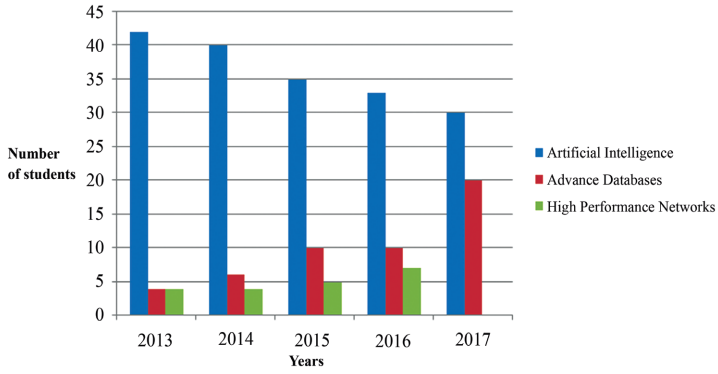
The educational institutes can use the student data perform an analysis of the activities of students for their courses and use this information to get a general overview of the learning experience of the student. For instance, many colleges make use of id card/badges that are needed to be swiped when they enter or exit a building. This information can be used to see where the badge is swiped the most. Another example is recording of attendance in courses. This information can be used to determine how many students are present for each of the courses and further analyze why they are not attending the course.

Consider the following sample data that could be extracted from student data:



The above data shows the percentage of attendance for different subjects. This information can be used by the management and educators to assess

their course materials, evaluate student performance, etc. Another sample set showing the preference of elective courses for a group of 50 students is shown below:

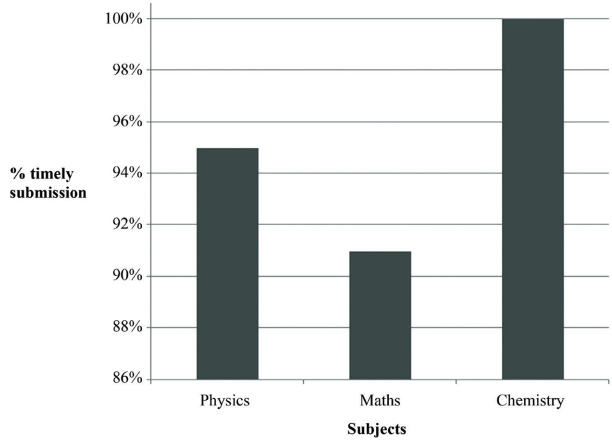


The above diagram shows that most students prefer to enroll in the artificial intelligence elective course and very few of them show an interest in the High Performance Networks course.

This information can be used by the management to invest more funds in the first course and maybe decide to discontinue the latter course.

Another event that can be monitored is the timely submission of assignments.

This is shown below:



The above graph clearly shows that the subject of chemistry has a timely submission rate of 100%. There could be many reasons for this particularly

interesting behavior. The reason for this can be analyzed and then used by teachers of the other subjects so as to increase their submission rates.

This is where data mining can help immensely. In the above example it can be found out by use of data mining that the submission deadline for the subject of Chemistry is usually 12 noon as compared to 8 am and 9 am for Physics and Maths, respectively. This relationship between submission time and the % of timely submission can be discovered with the help of data mining. It helps us understand that students prefer late submissions as compared to early submissions maybe because they spend the previous night working on the assignment and then fell asleep in the wee hours of the morning and simply forgot to send in their assignment as they were asleep!

Such associations and correlations are possible with the help of data mining.

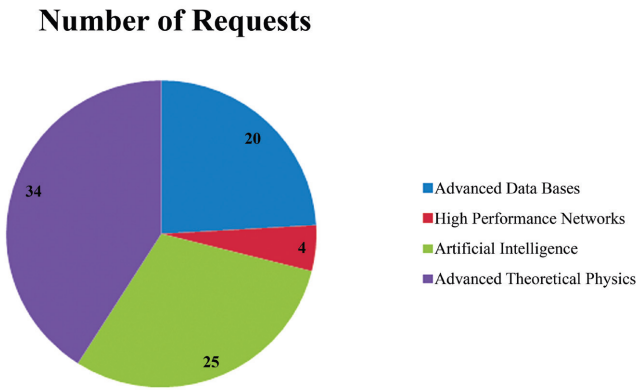
But, for the purpose of providing a visual representation of the consolidated data, generally, a combination of statistics and visualization are used. Softwares that analyze statistical data such as SPSS can be used for this task. Such tools can easily analyze and process student data such as entry and exit times of students in the campus, the most checked out books, books that are in high demand, e-books and papers that more most researched, downloads of e-learnings materials, time spent of e-learning websites, etc. Such tools can be successfully used to provide summaries of students' usage data on a daily, weekly or monthly basis. It can help look for patterns and trends in the student behavior, form rules on the sequence of events followed by students during exams, the order in which they prefer finishing their tasks, which subjects are more researched by students and so on.

The use of statistical analysis is helpful in obtaining reports on the time spent by students on their assignments, his submissions rates and times and predictions regarding his future submissions as well.

For example, the library books request system can use detailed analysis too. Consider the following table that shows the total number of books that were requested per subject as follows:

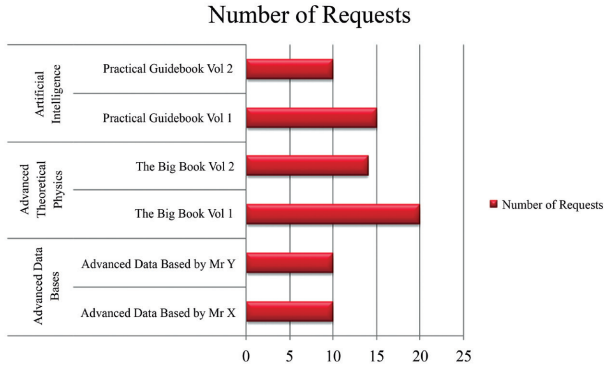
Subjects	**Number of Requests**
Advanced Data Bases	20
High Performance Networks	4
Artificial Intelligence	25
Advanced Theoretical Physics	34

This information can be visualized as a pie chart as shown below:



Further, this information can be refined at a deeper level to obtain details on the book titles that were requested for each subject along with the number of requests for each book.

This is shown below:



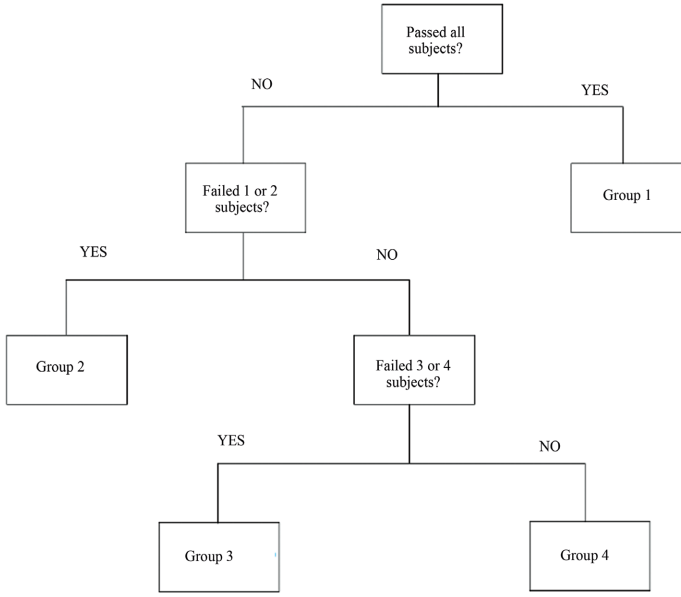
The above diagram shows the information at the refined level showing the total number of requests for each of the individual books for each of the subjects.

This information can be used by the management to make executive decisions such as buying more books for the library, adding the e-book version of a particular book that is most requested, providing papers related to that book, starting a special rotation policy for those particular books, etc.

Data mining can also be used for classification purposes as well. Based on specific factors such as performance and score, they can be classified into different groups.

Classification analysis is often done to group like-minded students together. This analysis can also be done to group students into different teams for events, to analyze student performance to form student societies, etc.

One such example is shown below:



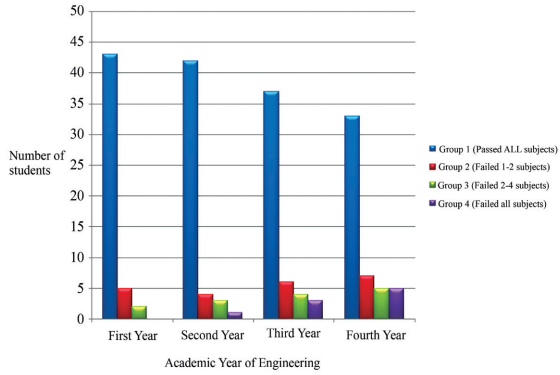
The above example shows a classification method that is used to group students based on their performance in their courses.

The application of this information on student data for the 4 years of college for one particular discipline (Engineering) is shown below:

Grouping Students (group of 50) based on their results:

Group No. **College Year**	**Group 1 (Passed ALL subjects)**	**Group 2 (Failed 1–2 subjects)**	**Group 3 (Failed 2–4 subjects)**	**Group 4 (Failed all subjects)**
First Year	43	5	2	0
Second Year	42	4	3	1
Third Year	37	6	4	3
Fourth Year	33	7	5	5

This information can be represented in a more visual manner as shown below:



As discussed before, the prediction of student's performance individually or as a group is one of the most popular and requested applications of EDM.

When we try to predict a student's performance, we estimate a value that is unknown and this value describes the student. In the education industry this unknown value is generally the student's performance, their score or rank. This value is usually a numeric or continuous value.

Different predictive data mining methods can be used to predict the performance of a student. On such method that is often used is regression analysis where a relationship between different factors is found. Classification, as explained previously can be used to group individuals on the basis of quantitative features and characteristics present in the training set.

Moreover, for the purposes of analysis and prediction neural networks, rule-based systems, Bayesian networks, regression analysis and correlation analysis can be used as well. This use of different analysis methods is employed to help predict the student's performance (predict his success in a course for example). These data mining methods can further be used to predict a student's final grade based on variables that can be easily extracted from the logged data of the student.

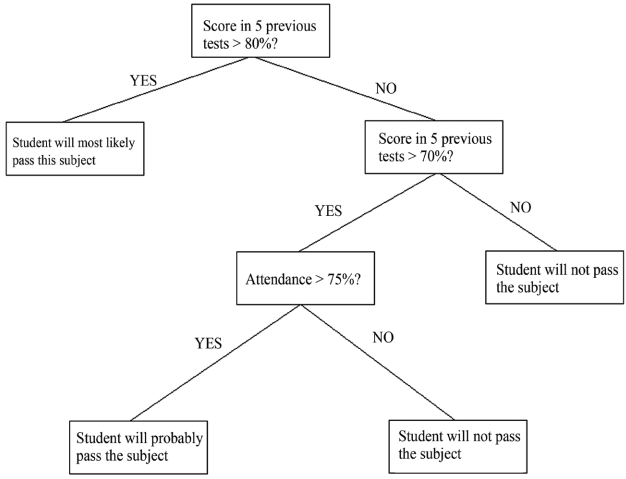
A wide range of rule-based systems have successfully been applied in the prediction of a student's performance in an e-learning setup using fuzzy association rule mining.

Some of the regression techniques that have been applied for different purposes are shown below:

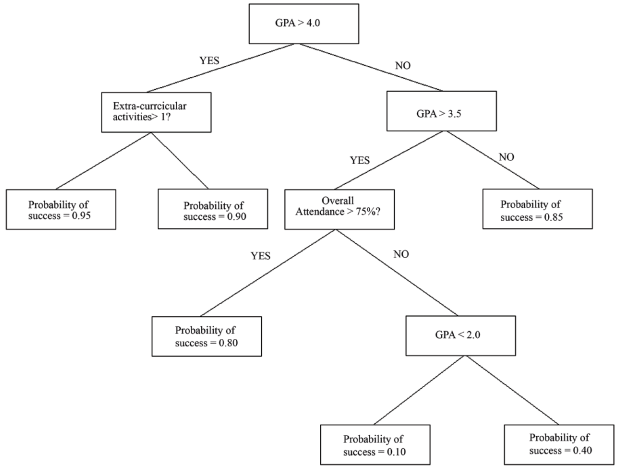
Type of Regression Analysis	**Application**
Linear Regression	Prediction of Student's marks

Stepwise Linear Regression	Time that will be spent on a learning page/ website
Multiple Linear Regression	Prediction of Exam results
Multiple Linear Regression	Identification of variables that are directly related to a student's success in college

Let us take a look at a simplistic decision tree-based classification for students based on their score:

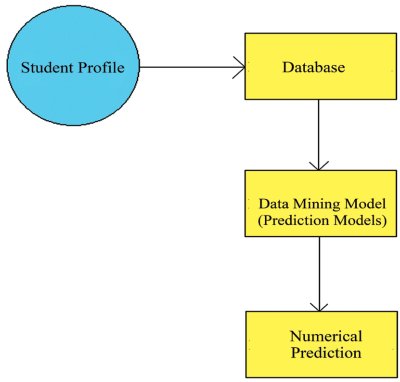


Another example of a decision tree that predicts the success of a student in college is shown below:



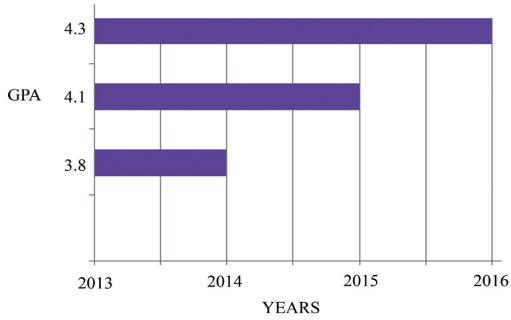
Additionally the prediction of marks that a student will obtain can also be done with the help of numeric prediction values.

The process is shown below:



Based on the past data, training sets are formed and these sets are used to construct rules and formulas that are applied on the testing data set.

For example consider the following table with the GPA of a student for 3 years:



A rule could be developed based on the training set that factors in the student's profile could be something as follows:

IF STUDENT PASSED ALL TESTS OF ALL SUBJECTS

AND

IF GPA > 3.7

AND

STUDENT ATTENDANCE> 80%

AND

IF NUMBER OF HOURS SPENT IN LIBRARY > 120
THEN PREDICTED GPA = AVERAGE GPA OF ALL YEARS

If all these conditions hold true for this student then the predicted GPA value for this student would be the average of the last three years.

Estimated GPA (2016) = 3.8 + 4.1 + 4.3 / 3 = 4.06

Many such rules can be developed based on the input values of the student and his profile and his GPA.

Some of the rules that can be generated could be as shown below:

RULE 2:
IF STUDENT PASSED ALL TESTS OF ALL SUBJECTS
AND
IF GPA > 3.7
AND
STUDENT ATTENDANCE > 70%
AND
IF NUMBER OF HOURS SPENT IN LIBRARY = 100
THEN PREDICTED GPA = (AVERAGE GPA OF ALL YEARS – 0.25)

If all these conditions hold true for this student then the predicted GPA value for this student based on the GPA values provided in the above would be the subtraction of 0.25 from the average of the last three years.

Estimated GPA (2016) = (3.8 + 4.1 + 4.3 / 3) – 0.25 = 3.81

RULE 3:
IF STUDENT PASSED ALL TESTS OF ALL SUBJECTS
AND
IF GPA = 3.7
AND
STUDENT ATTENDANCE> 60%
AND
IF NUMBER OF HOURS SPENT IN LIBRARY > 100

AND

IF ELECTIVE COURSE = “ADVANCE DATABASES”

THEN PREDICTED GPA = (AVERAGE GPA OF ALL YEARS – 0.15)

If all these conditions hold true for this student then the predicted GPA value for this student based on the GPA values provided in the above would be the subtraction of 0.15 from the average of the last three years.

Estimated GPA (2016) = (3.8 + 4.1 + 4.3 / 3) – 0.15 = 3.91

RULE 4:

IF STUDENT PASSED ALL TESTS OF ALL SUBJECTS

AND

IF GPA > 3.6

AND

STUDENT ATTENDANCE> 86%

AND

IF NUMBER OF HOURS SPENT IN LIBRARY = 100

AND

IF ELECTIVE COURSE = “ADVANCE DATABASES”

THEN PREDICTED GPA = (AVERAGE GPA OF ALL YEARS + 0.17)

If all these conditions hold true for this student then the predicted GPA value for this student based on the GPA values provided in the above would be the addition of 0.17 to the average of the last three years.

Estimated GPA (2016) = (3.8 + 4.1 + 4.3 / 3) + 0.17= 4.23

RULE 5:

IF STUDENT PASSED ALL TESTS OF ALL SUBJECTS

AND

IF GPA > 3.6

AND

STUDENT ATTENDANCE> 90%

AND

IF NUMBER OF HOURS SPENT IN LIBRARY > 130

AND

IF ELECTIVE COURSE = “ARTIFICIAL INTELLIGENCE”

AND

IF ENGLISH LANGUAGE SCORE > 75/100

THEN PREDICTED GPA = (AVERAGE GPA OF ALL YEARS + 0.22)

If all these conditions hold true for this student then the predicted GPA value for this student based on the GPA values provided in the above would be the addition of 0.22 to the average of the last three years.

Estimated GPA (2016) = (3.8 + 4.1 + 4.3 / 3) + 0.22 = 4.28

RULE 6:

IF STUDENT PASSED ALL TESTS OF ALL SUBJECTS

AND

IF GPA (latest year) > 4.0

AND

STUDENT ATTENDANCE> 90%

AND

IF NUMBER OF HOURS SPENT IN LIBRARY > 150

AND

IF ELECTIVE COURSE = “ARTIFICIAL INTELLIGENCE”

AND

IF ENGLISH LANGUAGE SCORE > 80/100

THEN PREDICTED GPA = (AVERAGE GPA OF ALL YEARS + 0.40)

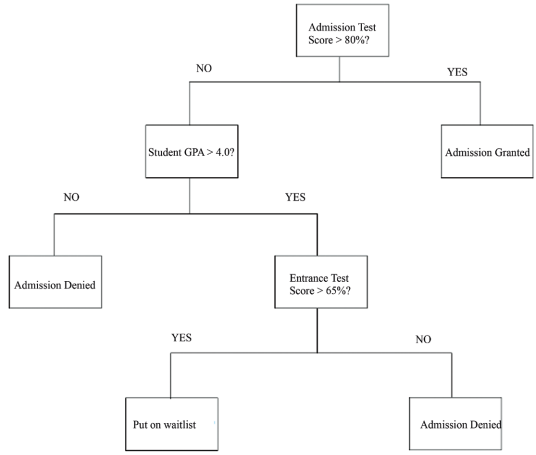
If all these conditions hold true for this student then the predicted GPA value for this student based on the GPA values provided in the above would be the addition of 0.4 to the average of the last three years.

Estimated GPA (2016) = (3.8 + 4.1 + 4.3 / 3) + 0.40 = 4.46

Another managerial application of data mining is the enrollment process. Colleges are now capable of enrolling thousands of students over a wide variety of courses and disciplines. Hence, the enrollment process is now online. But, even though the potential applicants apply online, the applications are evaluated online. This task is a tedious task that requires a lot of manual labor from the staff at the admissions office.

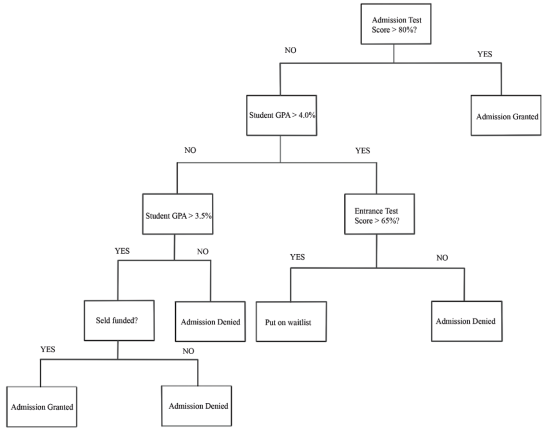
Data mining can be of use in the task of pre-screening. There exist a lot of factors that play a part in the enrollment process and using these parameters, applications can be screened and only those applications that pass the screening process are sent into be evaluated manually.

Consider the following decision tree that can be used to help the decision process:



The above tree does a small pre-screening on two factors: Entrance test score and GPA.

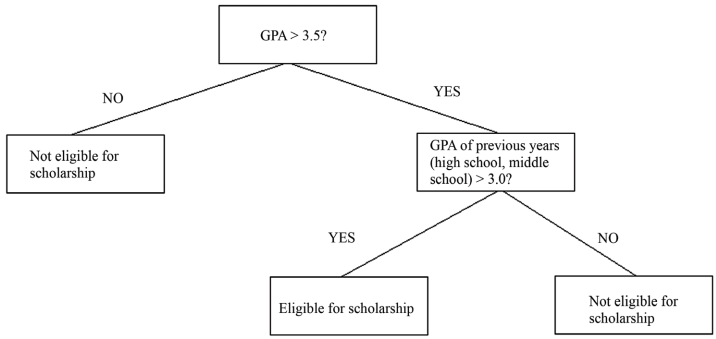
This decision tree can further be extended to consider additional variables as shown below:



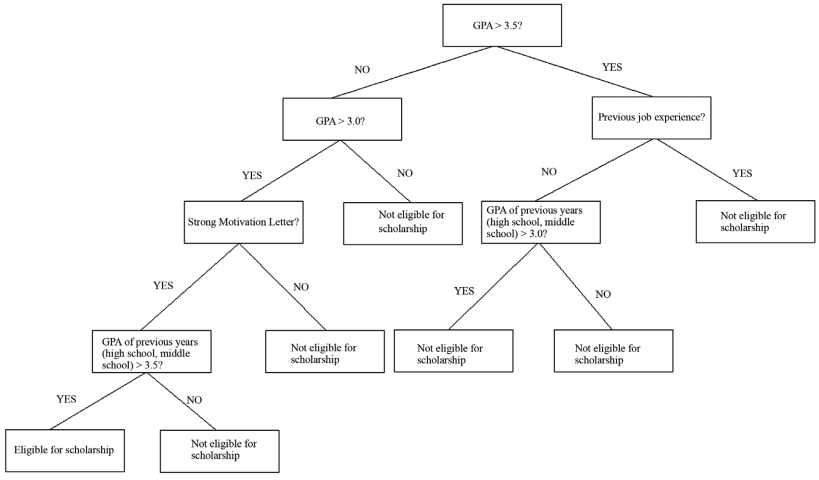
As seen from the decision tree above another parameter relating to the funding information has been included in the decision process.

Colleges and educational institutes usually have some form of scholarship grants and programs. In addition to the enormous amounts of applications that they receive for enrollment, they also receive several scholarship applications well. It is their job to analyze each application based on factors such as grades, test scores, their motivational letters, their profile, their existing work experience level and so on. This process can be automated by application of data mining techniques that can pre-screen the scholarship applications and only accept the ones that fit the criteria provided.

A sample is shown below:



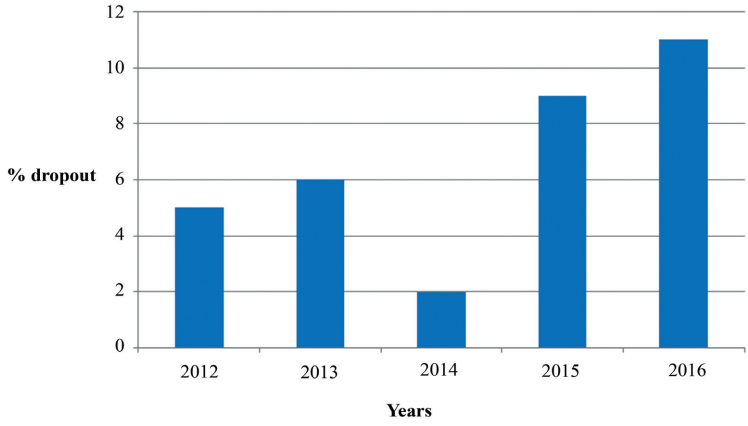
A more detailed screening process is shown below:



Another application of data mining is the prediction of dropout rates. Based on the historic data regarding dropouts over the few years, predictions regarding the future rates can be made. Additionally, individual student

data can be monitored for warning signs and alerts can be sent out to the respective parties.

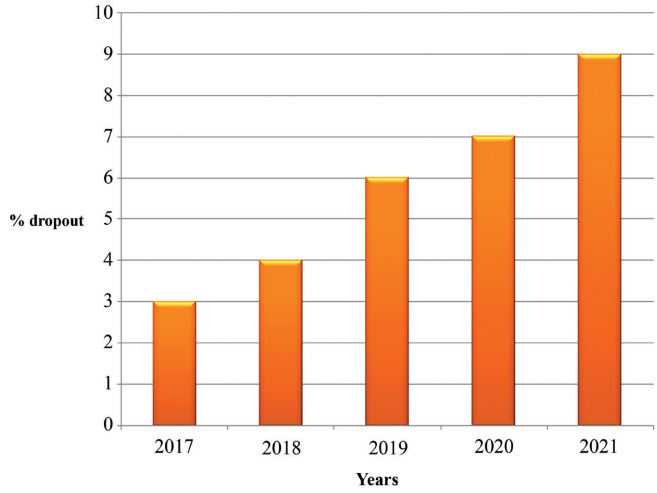
Consider the following table showing the dropout rates for the last 5 years of a college:



Based on this data, predictions regarding the next few years can be made. Prediction algorithms discover correlations between different parameters such as the population enrolled, staff members, overall student performance, students' motivation levels, etc.

These correlations and associations help predict the future rates.

A sample prediction is shown below:



In addition to predicting the overall dropout rate, data mining can be put to use in the prediction of dropout of individual students as well. This can be done with the help of a monitoring system.

The dropout data from the previous years is processed and a training set is formed. Based on this training set, rules are formed and associations are mined. These constructed rules are then applied on each incoming student data and the risk of dropout is calculated.

There are several ways that the risk of dropping out can be calculated. One of the most common ways is predictive modeling using classification trees to classify students as low risk and high risk. Another way is the use of regression models and neural networks to predict if a student will drop out.

Consider the following rules that can be mined from association rule mining:

Rule 1:
IF student fails 4 or more subjects for more than 2 semesters
THEN DROPOUT = YES

Rule 2:
IF the student fails 3 or more subjects for 3 semesters
THEN DROPUT = YES

Rule 3:
IF the student fails 3 or more subjects for 4 semesters
AND
IF elective subject = ‘High Performance Networks’
THEN DROPUT = YES

Rule 4:
IF the student fails 3 or more subjects for 5 semesters
AND
IF elective subject = ‘High Performance Networks’
THEN DROPUT = YES

Rule 5:
IF the student fails 2 or more subjects for 6 semesters
AND
IF elective subject = 'High Performance Networks'
AND
IF ATTENDANCE < 50%
THEN DROPUT = YES

Rule 6:
IF the student fails 2 subjects for 4 semesters
AND
IF elective subject = 'Artificial Intelligence'
AND
IF ATTENDANCE > 65%
THEN DROPUT = NO
Consider the following test set against which the rules are applied:

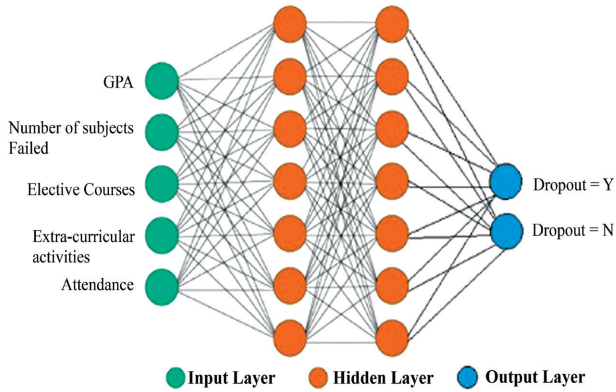
Student ID	Number of Subjects Failed	Number of semesters	Elective Subject	Attendance
1	5	3	Database systems	20%
2	4	5	Artificial Intelligence	35%
3	2	4	Artificial Intelligence	70%
4	4	6	High Performance Networks	20%
5	6	3	Database systems	22%

If the above rules are applied to this data we get the following predictions:

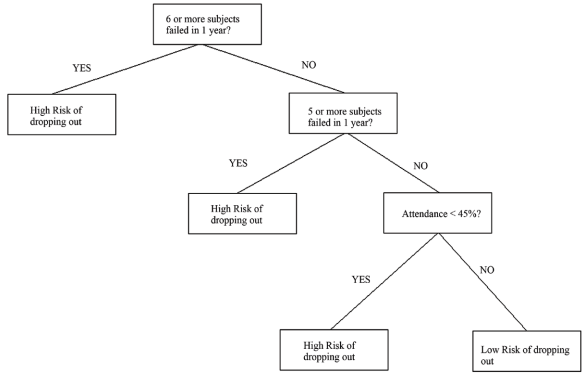
Student ID	Number of Subjects Failed	Number of semesters	Elective Subject	Attendance	Dropout Prediction (Y/N)

1	5	3	Database systems	20%	Y
2	4	5	Artificial Intelligence	35%	Y
3	2	4	Artificial Intelligence	70%	N
4	4	6	High Performance Networks	20%	Y
5	6	3	Database systems	22%	Y

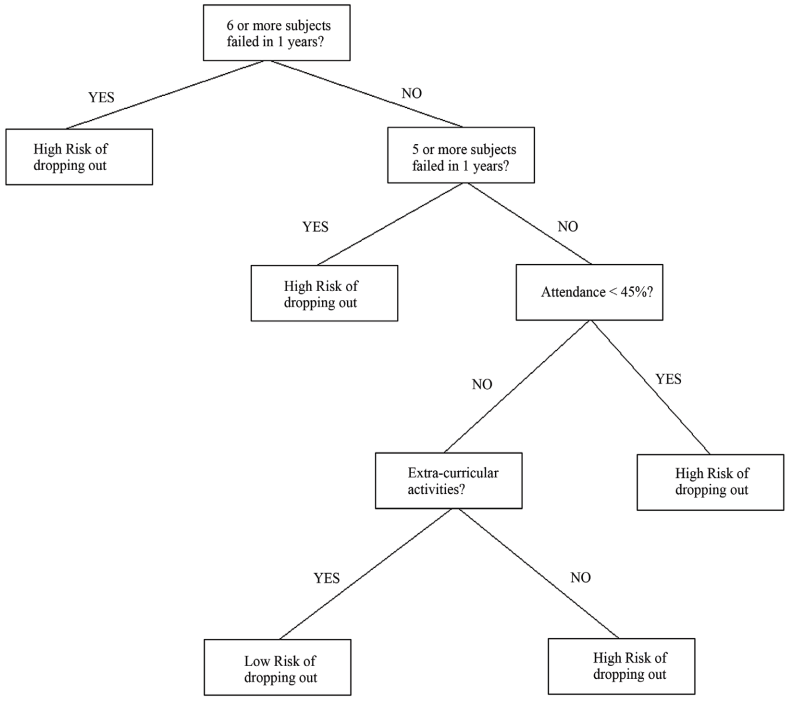
The use of a neural network to predict dropout is shown below:



A decision tree used to classify students as 'high risk' or 'low risk' is shown below:

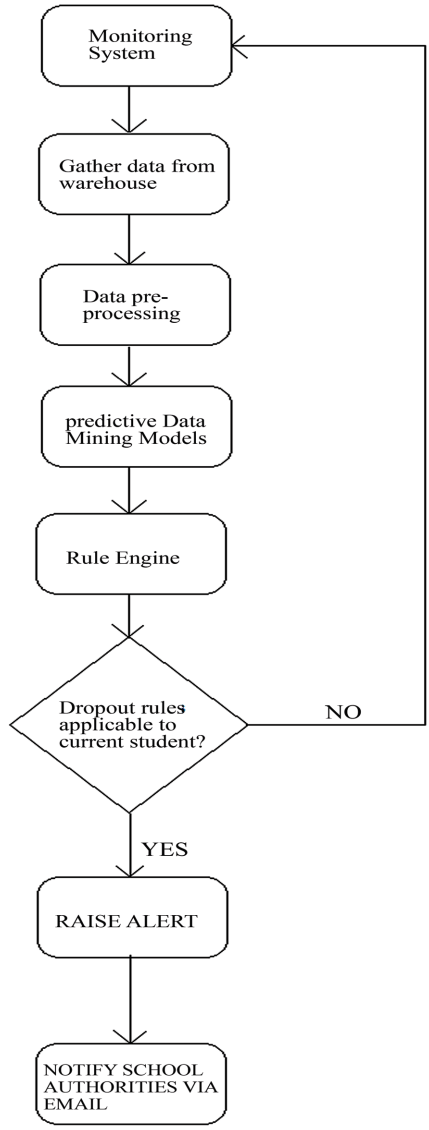


This decision tree can further be defined as follows:



A monitoring system that constantly monitors student data can also be put into place to detect early signs of problems that lead to a student dropping out (bad grades continuously, low attendance, etc.).

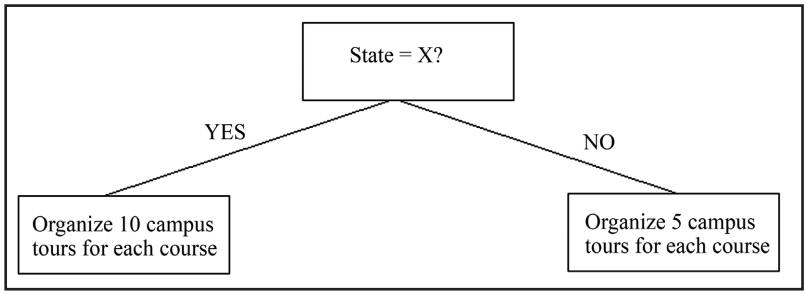
This can be done with the help of predictive models that form training sets and rules that help monitor the current status of each student and generate alerts based on the inputs it received.



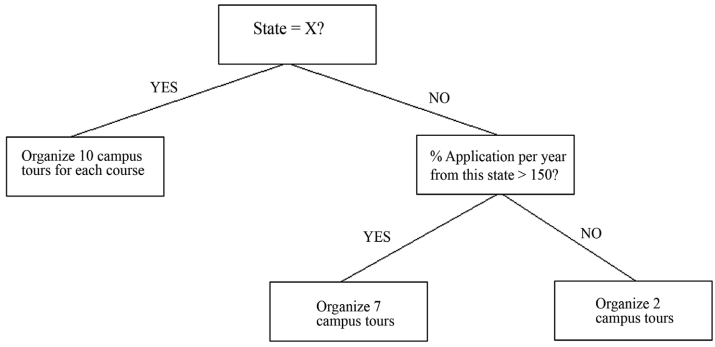
A well-known and well-used application of data mining is also present in the marketing domain of educational institutes. Colleges and schools can run targeted campaigns in order to promote their school and their courses.

For instance, if data mining discovers that students from a city are more favorable to their university, they may decide to invest more into the campaign in that city by increasing campus tours, offering more visitations.

A sample decision tree is shown below:



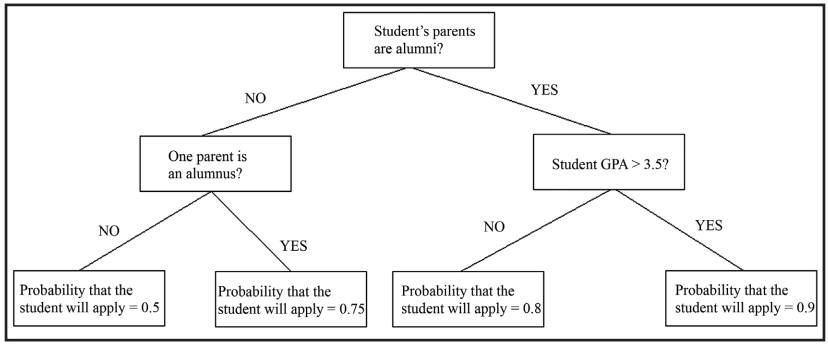
A more refined decision tree is as shown below:



Data mining can also be used to predict what profiles of students would most likely apply to their courses. This can be done with the help of prediction algorithms that take in several factors such as age, educational background, regional information, family history data, etc.

For instance, if a student's parents were both students of the same school or university, it would play a major part in student's decision to attend the same university.

An example using a decision tree is shown below:



3
CHAPTER

APPLICATIONS OF DATA MINING IN ENGINEERING

CONTENTS

3.1. INTRODUCTION

Data mining use in the field of engineering is a quite well known. The use of data mining in engineering is similar to its use in the field of science. Due to the swift rise in the development of computer and information technology With the rapid development of computer and information technology over the last decades, the amount of data in the field of engineering and science has been increasing at a fast pace and will continue to do so for the years to come as well.

This data is either being stored in large storage spaces and devices owing to the constant influx of information. Too add to this, this data is freely and easily available as it is made available to the world via the Internet. These volumes of data have changed the fields of science and engineering by transforming them from being 'poor' to 'rich' in terms of data. This growth spurt of data has led to the calling of new methods that help conduct the research in these fields.

Both fields engineering and science, usually collect a massive amount of data on a regular basis. Both these practices often collect large quantities of data and usually stored in large data warehouses and that require a lot pre-processing because the data is highly complex.

Also, lots of human communication is also present in these fields. There is some sort of human communication every second on the internet. There is an outpour of human interactions on the online front in several forms such news, articles, online forums and discussions, blogs, papers, messages, etc. on the web that also include social networking sites. Usually with the complex nature of data, these fields often make use of different visualization techniques such as graphs, charts, networks. In addition, a lot of processes followed in the field of engineering require responses in real-time and this presents a new area of research for data mining too.

By deduction, the application of data mining in these fields has started to become extremely popular.

The field has many applications in several fields such as telecommunications, software engineering, social sciences, energy domain, etc. Data mining can be used to make better managerial decisions, improve engineering design processes, improve manufacturing processes, predict the trends in the engineering data, detect fault in engineering networks, monitor software and hardware systems and help provide results in order to improve

upon the processes in general. It can be used to isolate network faults, detect anomalies in the network grids, and possibly reduce the reaction time for real-time systems.

Data mining in science and engineering is a rich and very active domain for due to the unique challenges faced in this domain. The employment of data mining in these fields requires the development of futuristic, scalable data mining systems capable of handling real-time engineering and science data.

We shall now discuss some applications of data mining in the fields of engineering and software engineering.

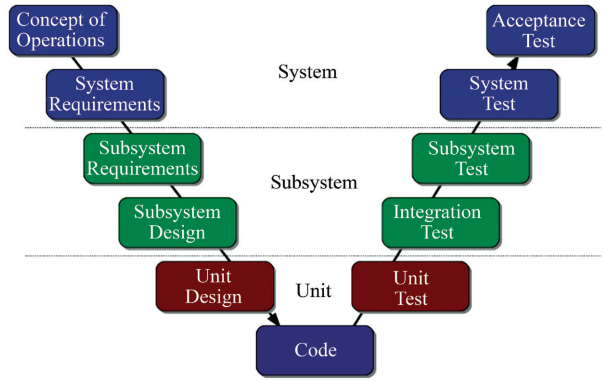
3.2. SOFTWARE SYSTEMS

Software plays a crucial role in the current times as most industries have turned to software for implementing and managing their systems. Software is crucial to many industries such as the banking industry, the governments, the engineering and the medicine and healthcare industry. It is vital to the growth of businesses, societies and governments. The software systems developed for sensitive industries such as the government and banking industry are extremely critical, complex and often need to be highly operational, reactive and follow strict standards of development. Additionally, most softwares that are being developed in the recent times are complex and hence more difficult to conceptualize. The complexity of softwares compounded with external dependencies and different programming approaches and standards, tends to slow down the development process and thereby the maintenance activities as well. Sometimes, the complex nature of the software systems along with limited understanding of business needs may lead to faults, rise in the number of defects and finally delay the delivery process. All these factors may lead to increase in the cost of software development and loss for the client.

Due to the intricate nature of software systems and the associated complexity each and every company that develops and sells softwares usually employ some sort of process to help manage software processes.

One of the most common software development (SDLC) process is the V model that follows software activities in a sequential manner.

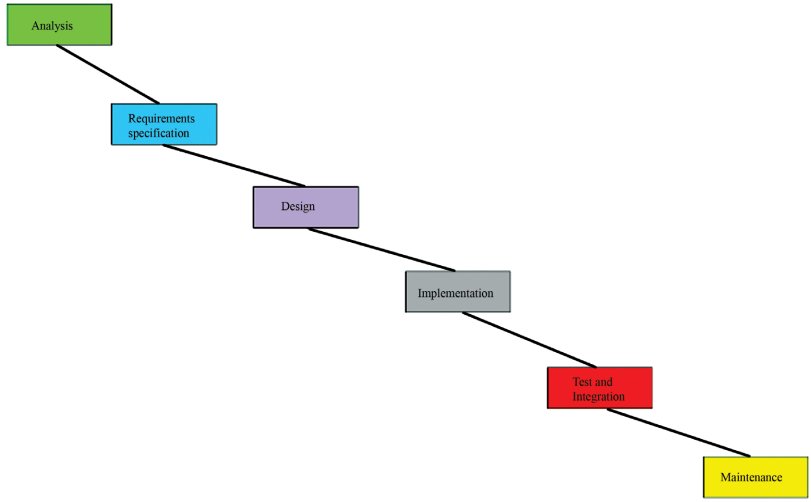
The model is shown below:



("Automation Learn – SDLC View," 2018).

The model follows a linear cycle and goes through each of the step until the end of the process.

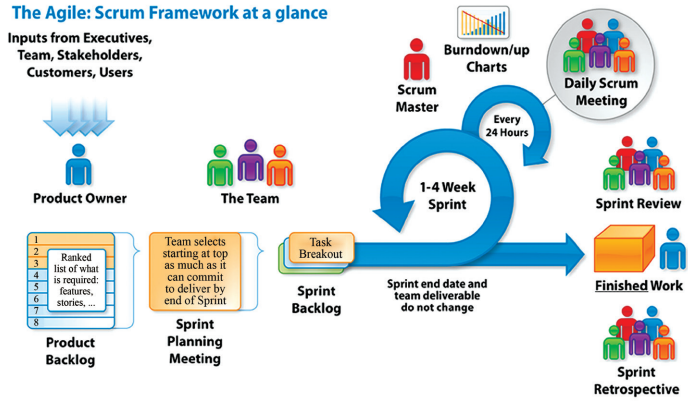
Another well-used model is the waterfall model as shown below:



This process is not very flexible as it does not permit to go back to the previous stage in case, which makes it impossible to modify the implementation without breaking the cycle.

Another process that is usually implemented is the Agile process that is a type of incremental model.

This is shown below:



("Automation Learn – SDLC View," 2018).

There exist manymore processes for software development management. But, these models and processes are usually based on previous experiences and not on empirical data. These processes do not factor in all the variables and hence are sometimes 'flying blind' and not completely able to ensure on time delivery. This does not happen due to lack of planning and low quality, but due to the difficulty in being able to predict all the variables and scenarios. The problem with software is that although it involves variables such as code count, defect count, number of developers involved, number of analysts, total number of hours put in, budget, experience level of those involved, these variables constantly vary for each company and for each software development team. Hence, this makes it a challenge to base the development process on just these factors.

In order to assess the quality of software, there have been several software metrics that have been in place since a long time. They have been in play to validate the software and its processes that produce the software. But, there exist many pitfalls and hurdles in the use of software metrics. However, there are pitfalls associated with the use of metrics as usually the people responsible for obtaining the metrics are tempted to only use the metrics that are easily obtained. This does not bode well as the parties responsible do not have all the information and hence will make uniformed decisions related to the project. Hence sometimes metrics do not truly serve their purpose. Although they seem interesting, there are high chances that this information is not informative and is potentially not relevant to the current project, or may even be invalid or maybe something that cannot be worked upon or is not an actionable item. The truly useful and meaningful metrics

may be difficult to obtain and understand. On the other hand, software teams and software engineering activities are generating a lot of data that can be harnessed. If this data is properly processed and captured it can be harnessed using data mining techniques to provide useful insights into the process of software development and management. Alternatively, software engineering activities generate a vast amount of data that, if harnessed properly through data mining techniques, can help provide insight into many parts of software development processes. The application of data mining methodology on this vast amount of data may help understand different parts of the software processes and organizations and the software development teams may benefit vastly from such methods that help uncover hidden information within the large amounts of software data. The software data serves as an interesting path of research and potential data mining implementations and contributions.

Data mining has been put into use for many purposes in the study of software systems such as improving performance of the system, detecting bugs, detect plagiarism of software, analysis of faults within the system, detect intrusions, check the system for vulnerabilities, etc. Data mining for software and engineering systems can be applied on different types of data such as static data or dynamic data depending on the type of software system. There exist several methods that have been developed in this particular domain that use different data mining techniques such as machine learning, pattern recognition, visualization, neural networks, statistics, etc.

We shall now take a look at a few applications of data mining in the software engineering.

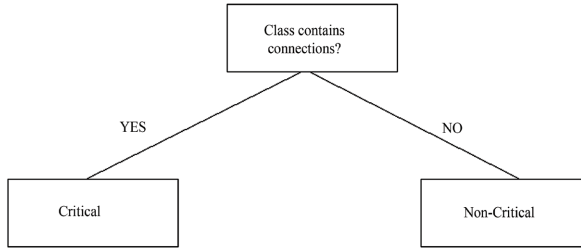
A very useful and common application of data mining is the use of data mining to classify parts of code into different segments as per their level of criticality.

We consider a simple software system with 5 classes. Data mining classification can be used to classify these 5 classes into two groups: critical to the system and non-critical.

Class Name	**Group**
Class 1	Critical
Class 2	Critical
Class 3	Non-Critical
Class 4	Non-Critical
Class 5	Non-Critical

This classification of code can be done on the basis of several factors such as type of code, number of lines of code, structures used, etc.

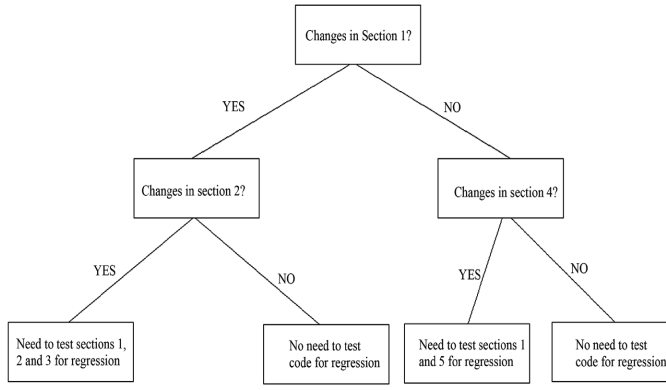
A sample decision tree is shown the problem:



Another possible application of data mining is the prediction of coupling between different parts of the code for maintenance. This would predict the impact of changes performed on one part of the code on the other parts so as to estimate the possible areas that need to be tested for regression.

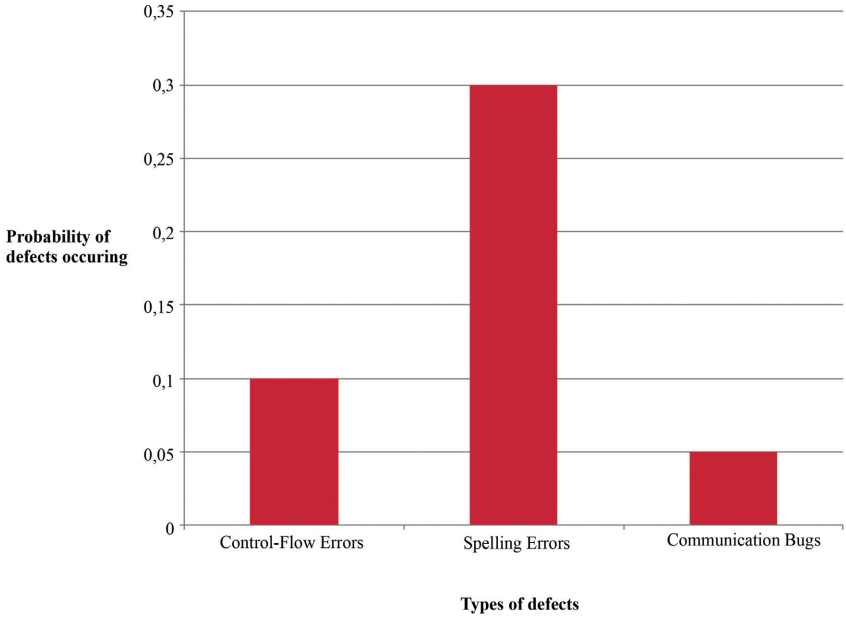
This can be done with the help of classification algorithms such as decisions trees, CART, etc.

Consider a software system that is divided into 5 sections: Sections 1–5. We see below a decision tree that predicts the need for regression thus highlight coupling between different sections of the software.



Another potential use of data mining in software systems maintenance is its use to predict type of defects that can arise based on the past defects. Based on historic data, training sets are formed and used to develop rules that help predict the type of bugs that may occur in the future.

This is shown below:



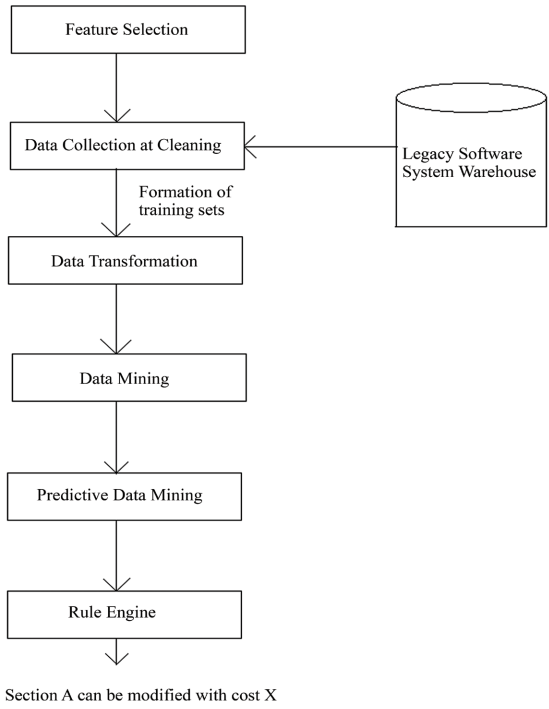
In addition to detecting and predicting the types of bugs or defects, data mining can be used to detect potential areas of improvement for legacy code.

For instance, for a programming language such C++, C, making changes in the code in complex legacy systems is costly and has a huge risk associated to it. Data mining can help predict the cost of changes to such systems along with their added value with the new changes that would be made.

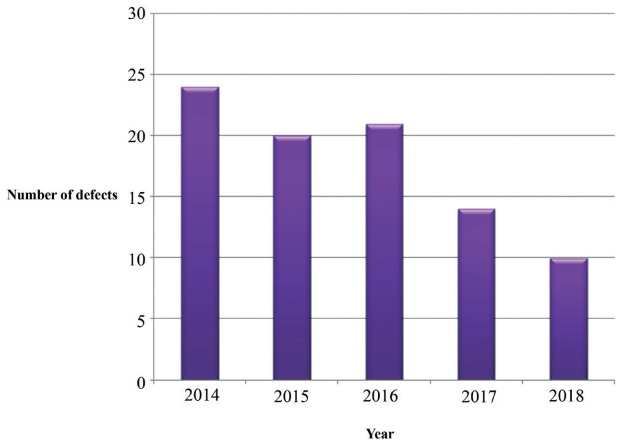
It can use predictive modeling techniques to return a particular section of code, which on modification will not risk the chances of breaking the existing software. This can be done with the help of predictive regression analysis, which takes in sets of training data that consist of rules that define the areas of risk within the code.

Consider a legacy code that can be classified into different sections or areas and these sections can be analyzed for possibilities of improvement.

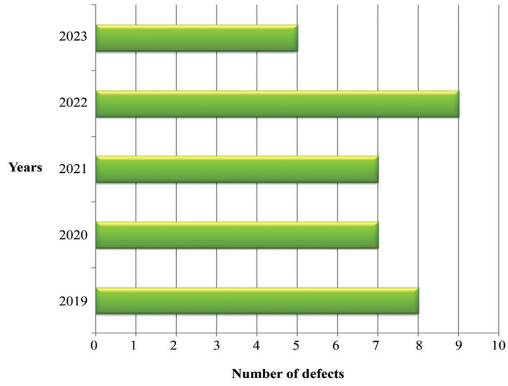
The process would be as shown below:



Predictive models can be used as shown above to predict which section of the legacy code can be modified along with the cost of its modification.



Used to predict the number of defects that applications will have in the next few years.



Now that we have seen some possible applications of data mining in software engineering, we will take a lot at some successful existing implementations in this field.

3.3. APPLICATIONS IN SOFTWARE MANAGEMENT

One of the applications of data mining in software was proposed by Canfora and Cerulo (2005), which proposed the analysis of the impact of a change request on different parts of the source code. The research goal was to identify the work products of areas that would possibly be impacted by a new change/feature request or a defect fix. The authors of this paper studied open source projects to extract change request data and related data from defect tracking systems and source code versioning systems to uncover which possible source files would be affected or impacted by this particular request. This model discovers links between changes done in the past to the files impacted and based on the historical data and retrieval techniques, derives sets of files that will be affected by a particular request for change.

Research proposed by Hassan (2006) presents ways in which software entities and the historic data of software systems can be used to help managers better manage their teams. Hassan says that managers need to prevent the inflation of faults and at the same time ensure their quick discovery and repair which making sure the software continues to evolve in a graceful manner and can handle new client requirements. He provides a summary of the issues faced by software development managers such as prediction of defects, allocation of resources, etc. and also provides solutions to fix these issues.

Yet another research proposal in the field of software is done by Mockus et al. (2003) aims to predict the amount of effort remaining along with its distribution that is needed to finish a project. A predictive model is proposed based on the notion that each and every software change and modification can bring about the need for repairs at a later time in the lifecycle of the software (Mockus et al., 2003). Then they use the model they proposed to predict and plan the development resource allocation task for the existing projects. The model is a highly efficient model that investigates and predicts the efforts remaining in a successful manner. The results returned by this model also confirm that there is a relationship between new features/modifications and defect fixes.

A unique research performed by Atkins et al. (1999) attempts to provide quantitative values for the effects of a software tool on the effort of a developer. The focus of this research is to find relations between the tools and the software effort of developers. This is because software tools are supposed to improve software quality, but sometimes they are expensive and difficult to implement, install and maintain, mainly in large enterprises. The method proposed in this paper evaluates the tools and finds relations between usage statistics of the tool with estimates of the effort of developers. The approach proposed in this paper is not expensive and non-intrusive in its implementation and is purely observational. On the plus side, it includes controls for variables that are confounding. The results generated by this model are helpful to managers to effectively quantify the effect of a software tool on the effort of a developer. This model can help provide a view of the cost versus the benefit of the usage of the tool and help managers take decisions.

3.4. APPLICATIONS IN SOFTWARE DEVELOPMENT TASKS

The application of data mining in software development is useful in a lot of ways. The development of software is a complex and creative process and each program differs from the other. Due to the limited amount of data that is available at the beginning of software projects or shortage of exploitable data, it is difficult to provide perceptions that can help with the process of development of the project. During the initial phase of the project during the development phase, there is not enough relevant data that can help in the process. But as the development goes on, there is enough data generated, the effort can be measured and empirical data can be obtained that helps the

development tasks. One application proposed by Mens and Demeyer (2001) attempts to determine ways to apply metrics to software objects that are still under evolution. This paper focuses on evolution of software as a key point and tries to make a difference between predictive and retrospective analysis. The retrospective analysis is the most common form of analysis. This paper proposes a naming method/taxonomy to classify different segments of code with respect to their evolution.

The first classification is named 'e*volution-critical.*' This section needs to be evolved in order to improve quality of software and its structure or at least refactored in order to avoid the effects of aging software.

The second classification is called as '*evolution-prone*.' This section consists of parts of code that are unstable and are likely to be evolved more often as these parts correspond to software requirements that are constantly changing and are volatile.

The last classification is '*evolution sensitive.*' This section/class consists of code or parts of code that are highly coupled and can cause regressions like a ripple effect when it is changed.

An approach to software testing was proposed by Liblit et al. (2005) that makes use of an algorithm that performs dynamic analysis in order to isolate bugs in the software program through a process of sampling of predicates during the life of execution of the software program. This paper proposes ways to simplify predicates those are tested again and again and thus are redundant. This approach deals with these predicates that can point to more than one defect and thereby help identify multiple bugs at the same time. This work is a new fresh approach towards analysis of the quality of software as compared to the traditional static analysis methods that are extremely popular in the field of software engineering.

An approach towards testing the quality of code was proposed by Livshits and Zimmermann (2005) where they present a technique for discovering common error patterns in software, which combines mining of revision histories with dynamic analysis of the code, including correlation of method calls and bug fixes with revision check-ins. When this technique was applied to large systems with substantial historical data, the authors have been successfully able to uncover errors and discover new patterns that are specific to the applications (Livshits & Zimmermann, 2005). It was found that the errors detected using this approach was not known before and hence helped improve the quality of the software system. Another approach that helped detecting coupling in software was proposed by Zimmermann

et al. (2005) where they developed a tool that detects coupling in software systems and predicts possible changes in code.

The principal aim of this tool was to perform an inference and suggest possible or most likely changes based on the modifications done by a developer/programmer. Also, the goal was to prevent errors and issues that arise due to changes that are incomplete.

Their goal is to infer and suggest likely changes based on changes made by a programmer, but also to prevent errors due to incomplete changes. This tool makes use of association rules that help uncover links and correlations between changes made by a programmer and its impact and these rules sometimes help reveal coupling that is generally not detected by regular analysis of the program.

The effectiveness of this tool increases exponentially with an increase in historical data of the software. The predictive power of this tool is best when the tool has a lot of historic data of the software to make better predictions. It cannot be said that all the predictions made by this tool are valid, but in the best case scenario, the potential changes in the code are reported and the user is then able to re-evaluate the impact instead of omitting it altogether.

A managerial application in the field of software was proposed by Mockus et al. (1999) who proposed a model that analyzes legacy code and looks at changes in code that correspond to good decision in terms of business. This is a pure application of data mining, which is very prevalent. This paper proposes to analyze each proposed changed and evaluate it in terms of the profits it can return.

A quantification of a possible change to code is done.

They sate “[e]ach change to legacy software is expensive and risky but it also has potential for generating revenues [sic] because of desired new functionality or cost savings in future maintenance” (Taylor et al., 2010).

The authors studied a large software system and performed an analysis of the code on the basis of the changes (and factors such as cost of the change and quality) and propose inferences using measures of change that were extracted from change management and version control systems.

3.5. APPLICATIONS IN SOFTWARE DEVELOPMENT RESEARCH

The research of different ways to improve software processes is a budding

area of research within the software community. A lot of research on the open source data is being done these days. Researchers are constantly on the lookout for different patterns, correlations and anomalies in the software projects and the processes. The research community within the field of software engineering tries to gain an insight into the workings of the software development process world over and find links, patterns and relations between them. They aim to find and unearth characteristics that are common to all software systems and thereby deduce and form rules, standards and norms that can be applied at a global level to all the software systems.

One such analysis that is done by researchers is the analysis of open source projects that are available freely to determine trends and gain insight into the processes. Although this task is a difficult, it helps gain insight into the open source community and can be used to form training sets to check for software plagiarism as the freely available code is accessible to everyone.

Another pair, Gall and Lanza (2006) look at ways to analyze, filter and visualize the evolution of software processes. Identification of architectural decay and trends of logical coupling between unrelated files are also shown. They also try to identify decay of architecture and look for trends that show coupling between files that are not related.

An avenue of research in software engineering is the study of evolution of software. This is a common field of interest of data miners in the software industry. A research conducted by Ball et al. (1997) serves to uncover different and better ways that help understand the development history of a program through the applications of clustering and partitioning methods.

An area of research that is very active and well-known in the software engineering communities is the extraction of software contributors and finding correlations among contributors.

One such method of extraction was proposed by Alonso et al. (2006) where they tried to characterize the different roles of the participants of the project based on their rights of contribution. Another such approach was discussed by Zhang et al. (2007) that tried to concentrate on the performance of each developer individually. This work focused on calculating the effort and performance of each programmer or developer at an individual level so as to understand the contribution of the developer to the project.

Yet another area of interest in the research community of software engineering is the need for development of software tools that help better the process of collection, gathering and analysis of software components

and metrics. There exist many groups of researchers that have successfully developed tools that simplify the collection process of metrics. Some of these tools are re-usable and some of these tools tend to be more software specific.

An example of such a tool that is in use is a tool named 'GlueTheos' which was developed, created and written by a group Robbles et al. (2004). It is a tool that is used for collection purposes; gathering data from sources such as OSS. Although this tool is limited in its analysis features and visualization capabilities, its architecture is scalable and extensible.

Another architecture that collected and recovered software metrics in a non-invasive manner was proposed by Scotto et al. (2006). This approach tries to implement different collections techniques and tools for the metrics such as web-based and distributed tools that safely collect the software metrics. These tools collect and gather the data regarding the software metrics and manage to aggregate the information obtained with the lowest possible level of interaction required from the end users of the system.

3.6. PRACTICAL APPLICATION OF DATA MINING IN SOFTWARE ENGINEERING

In the previous section, we took a look at some applications of data mining in the field of engineering. In this section, we shall take a look at a practical example of data mining implementation.

In this section, we apply two data mining predictive classification algorithms (Zero R and One R) on a data set and evaluate the results. The algorithms will be applied to a dataset from a bank that contains its customer information. We apply data mining principles to this data so as to predict if a customer is likely to open a bank account with the bank or not based on the characteristics of the customer such as age, profession, etc. Hadoop MapReduce technique is used to apply data mining on the banking data set.

Before diving into the application, we shall first take a look the two classification algorithms.

3.6.1. ZeroR

ZeroR or also known as Zero Rule is an algorithm of the classification technique, it is the simplest and most naïve algorithm as it does not use any of the explanatory classes to predict the target class ("ZeroR," 2018). Simply, it predicts the majority value of the target class for all objects.

It ignores the values of all the predictor classes and very simply relies on the target or majority class. This algorithm is not frequently used for mining purposes. Due to its simplistic nature, it is generally used by other classification methods that use it is to determine a baseline value for their performance.

3.6.2. Algorithm

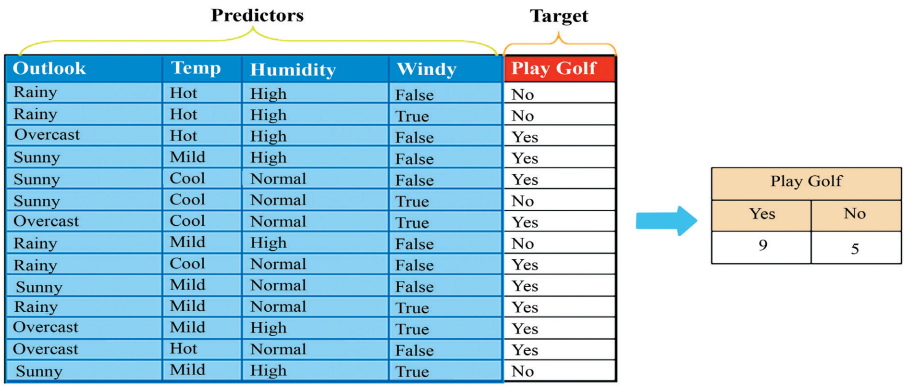


Figure 3.1: Frequency Table

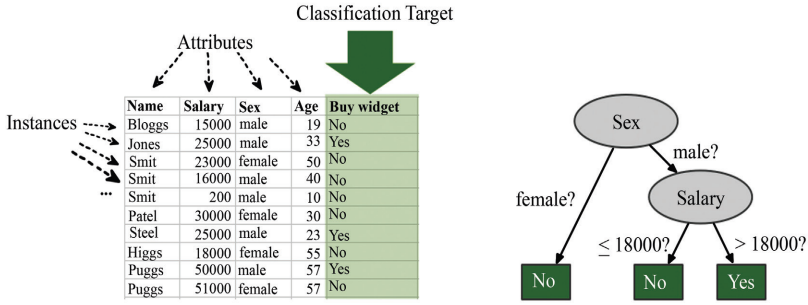
A frequency table of the target class is constructed from the available data. A frequency table is simply a count of each value of the class. The following figure shows a simple example to clarify the idea.

The above figure shows that table with the explanatory or predictor classes are describing the climate status. The target class is the decision of playing golf. For the ZeroR algorithm, the frequency table on the right is constructed with the number of yes decisions (9) and the number of no decisions (5). And the rule of the classifier is to predict yes for all records without taking into consideration the predictor classes. Though it is a very simple algorithm with no predictive power, it is useful to provide a benchmark for other classifiers so as to know what is the least acceptable accuracy for any classifier (Sayad, 2018).

3.6.3. OneR

OneR or also known as One Rule, as the name suggests it uses only one of the explanatory classes to predict the target class. The classifier is built by choosing the class which best describes the target class; it generates

rules such as the one shown in the figure below for each of the explanatory classes then chooses the class that has the least error when applying its rule to predict the target class.



("Data Requirements for Process Mining — Flux Capacitor", 2018)

3.6.4. Algorithm

Similarly to the ZeroR we construct a frequency table for each of the predictors against the target class. Using the same example in Figure 4, the frequency tables for the predictor classes would be as shown in the Figure below:

Frequency Tables

		Play Golf	
		Yes	No
Outlook	Sunny	3	2
	Overcast	4	0
	Rainy	2	3

		Play Golf	
		Yes	No
Temp.	Hot	2	2
	Mild	4	2
	Cool	3	1

		Play Golf	
		Yes	No
Himidity	High	3	4
	Normal	6	1

		Play Golf	
		Yes	No
Windy	False	6	2
	True	3	3

To choose the best predictor class, for each value of a predictor we sum up the frequencies of the dominant target value; for Humidity predictor for instance we add: 4 (frequency of "no" when the value is "high") and 6 (frequency of "yes" when the value is "normal"). We repeat this for all predictors and the best predictor is the one with largest sum. In this example the best predictor is Outlook with a sum of 10. And the rule for this predictor

is: If sunny or overcast predict yes, if rainy predict no. In the case study discussed later in the report we will see how to choose the explanatory class that best describes the target class from a practical programming point of view (Sayad, 2018).

Now that we have seen the ZreroR and OneR algorithms, we shall take a look the concept of MapReduce and take a look at the Hadoop framework that is used to process data.

3.7. MAPREDUCE

3.7.1. Definition and Conceptual View

Google MapReduce (Bappalige, 2014) is an emerging parallel programming technique designed to process big data, present in clusters in parallel using commodity hardware. It is a programming model and has a framework that attempts to introduce a level of abstraction in the computation of large-scale data, by hiding the system levels details from the developers. MapReduce libraries are available in many programming languages: java, Ruby, Python, C++ (White, 2012).

MapReduce programming is similar to functional programming languages like LISP and ML, it has 2 stages: Map and Reduce.

Map: Data is read, filtered and processed in small discrete 'chunks' in parallel. The input data is split and converted into another sorted dataset with the individual elements as tuples or key/value pairs.

Reduce: Output from the Map stage is taken as the input, to perform a summary/aggregation operation. The input data is combined into a smaller set of tuples.

3.7.2 Hadoop

Hadoop ("Apache Hadoop," 2018) is an open source software framework of Apache that enables distributed processing of big data using simple programming models. It is scalable up to thousands of machines, each of which offers local storage and computational capabilities. The Hadoop libraries are capable of detecting and handling failures at application level rather than relying on hardware.

The Hadoop framework consists of the following modules:

- ***Hadoop Common***: This module contains utilities and libraries

required by the other Hadoop modules.

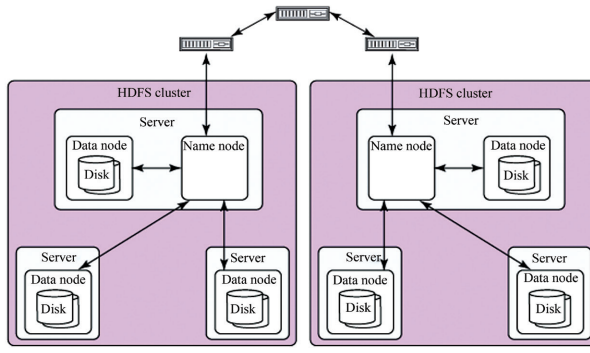
- ***Hadoop Distributed File System (HDFS):*** It is distributed file system that stores data on the commodity machines and provides high-throughput access to the application data (Hanson, 2018).
- ***Hadoop YARN:*** A framework for resource management in clusters and job scheduling.
- ***Hadoop MapReduce:*** A model for parallel processing of big data sets.

The two primary and important components at the core of Apache Hadoop are the Hadoop Distributed File System (HDFS) and the MapReduce parallel processing framework (Hanson, 2018). We will see them in detail below.

Hadoop Distributed File System (HDFS): HDFS is a portable, scalable, and distributed file-system written in Java for the Hadoop framework (Bappalige, 2014). In a Hadoop instance each node has a single name node, and a cluster of data nodes form the HDFS cluster. Name nodes manage the file system namespace and regulate the client access to files and data nodes store data as blocks within files (Hanson, 2018). Using a block protocol that is specific to HDFS, every data node serves blocks of data over the network. The TCP/IP layer is used for communication by the file system and RPC (Remote Procedure Call) is used for communication between clients.

3.7.3 Hadoop Architecture

The following figure illustrates the Hadoop architecture in detail. HDFS implements a master/slave design with the name nodes as master and data nodes as slaves.



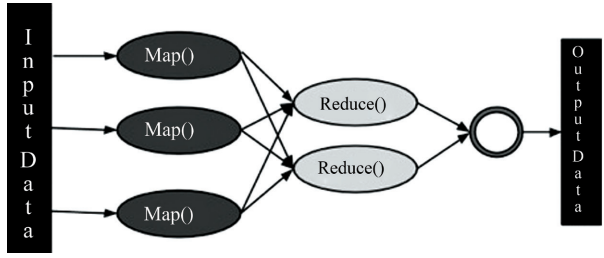
HDFS Architecture (Hanson, 2018).

3.7.4. Hadoop MapReduce

MapReduce is a software framework that enables developers to easily write applications that process large amounts of data up to terabytes in a parallel manner on large clusters containing thousands of nodes of commodity hardware in a reliable and fault-tolerant manner ("Hadoop MapReduce," 2018).

In a MapReduce job, the input data set is generally split into separate chunks that are parallel processed by the map function. The map output is then sorted by the framework, later used as input to the reduce function. Generally, the input and the output are stored in a file-system. The MapReduce framework manages scheduling, monitoring and re-execution of the failed tasks. Most of the computing task is done on nodes with data on the local disk itself to reduce the network traffic. On successful completion of the scheduled tasks, the data is collected and reduced by the cluster to form a proper result which is sent back to the Hadoop server.

The MapReduce works as follows:



("Hadoop MapReduce," 2018).

3.7.5 The Environment Setup

It is arguably one of the hardest steps in the implementation of MapReduce. With our choice of Hadoop, the most appealing environment is to have a platform with a GUI and necessarily running JVM. The logical next step is installing an IDE such as Eclipse and plugging in the MapReduce. This is a very complex task, and will be discussed briefly.

The first step in the setup is the installation of a Hadoop plugin for Eclipse IDE.

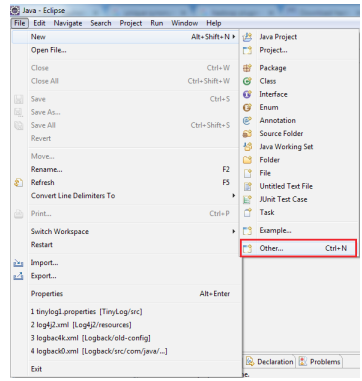
We go to the following site to download the plugin:

http://www.java2s.com/Code/Jar/h/Downloadhadoop0202eclipsepluginjar.htm

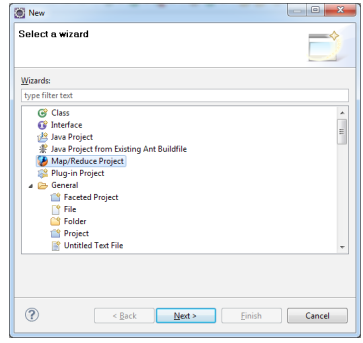
Once we download the plugin, we then put in our eclipse installation folder, i.e. the eclipse 'plugins' folder.

Once we have copy-pasted the plugin in this folder, we need to restart eclipse.

To check if the plugin has been added correctly in eclipse, we go to the File -> Other as shown below:



The following wizard dialog box opens up:



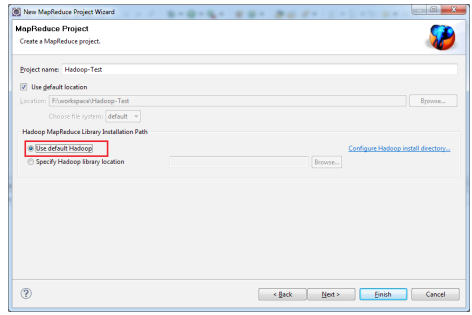
Once, we can see that we have a Map/reduce Project option, it is proof that the plugin was installed correctly.

We now create a simple MapReduce project.

The steps followed for creation of a Hadoop MapReduce project in eclipse is shown below:

First we click on the 'Map/Reduce Project' link in the dialog box shown above.

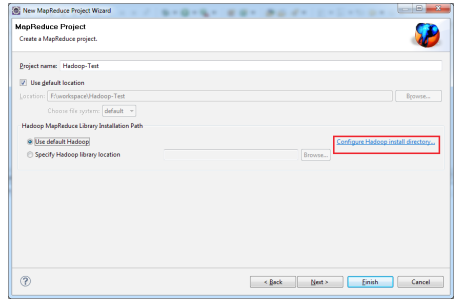
The following window opens up:



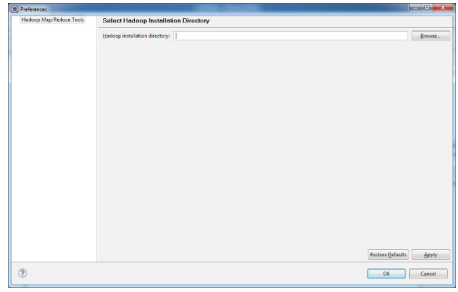
We give a name to project and select the Installation path of Hadoop. (Note that here we assume that Hadoop is already installed on the machine).

For the creation, we choose the 'Use default settings' option as shown above. But before terminating the process, the Hadoop installation directory needs to be configured.

In order to do this, we click on the 'Configure Hadoop install directory' link as shown below:

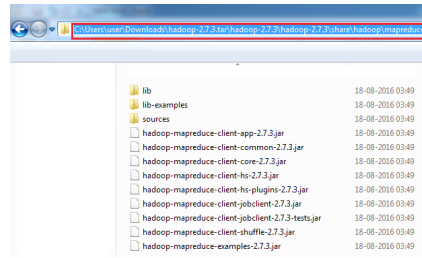


The following window opens up:

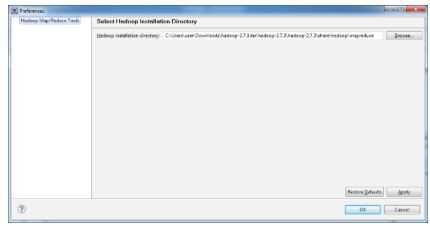


Here we provide the location of the 'MapReduce' folder in the Hadoop installation directory.

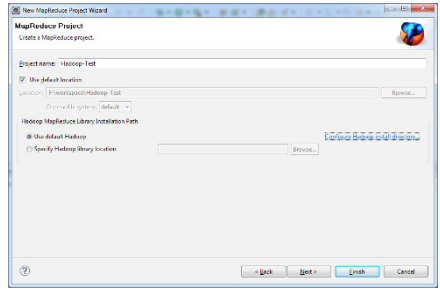
This is shown below:



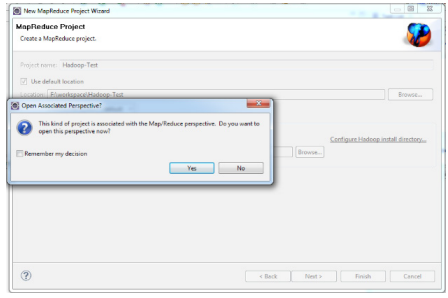
The location of the directory is updated as shown below:



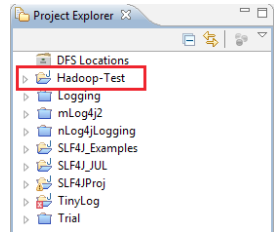
Once the installation directory is selected, click on finish.



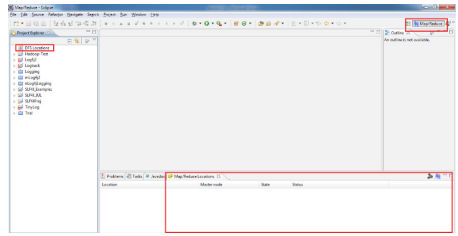
The following window opens up:



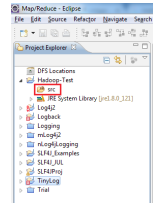
Click on ‘Yes.’ A new project is created as shown below:



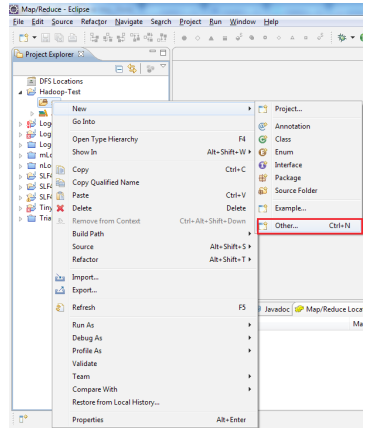
The project is opened in Hadoop perspective which adds some features as shown below:



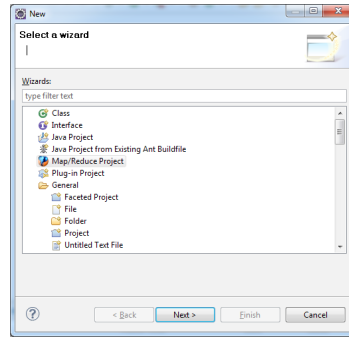
To create new classes in your project, first drop down to the 'src' level as shown below:



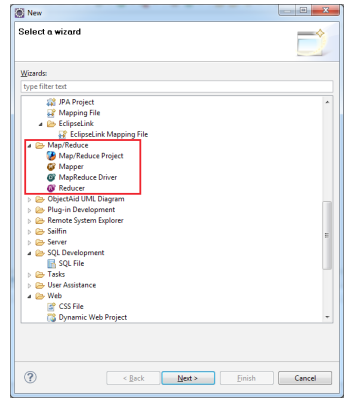
Right-click on the folder and go to New -> Other as shown below:



The following dialog box opens up:



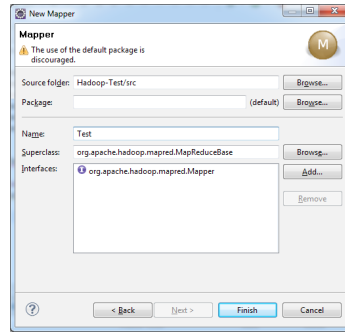
Scroll Down to the 'Map/Reduce' option as shown below:



As seen from the above image, we can now proceed to start developing using Hadoop MapReduce in Eclipse.

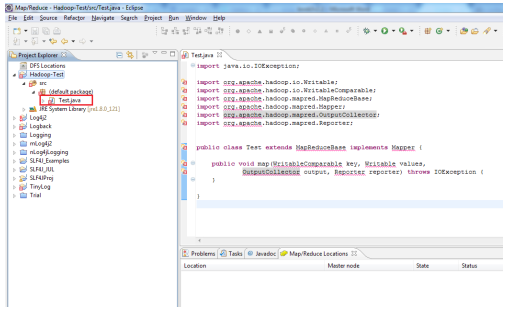
We shall now see an example of creation of a mapper. We click on the 'Mapper' option that we saw previously.

The following dialog box appears:

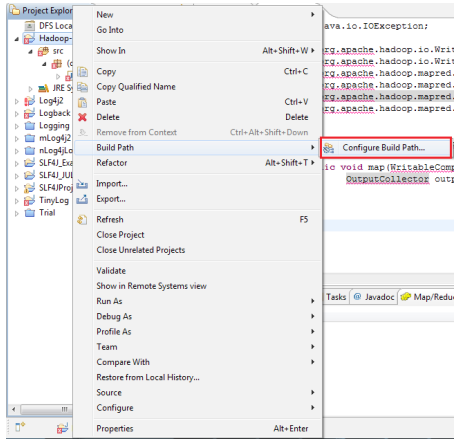


We enter the name of the Mapper and click on Finish.

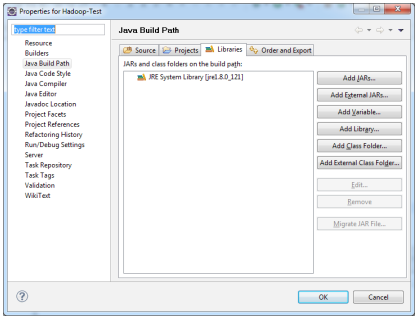
The following java class is created in the project structure as shown below:



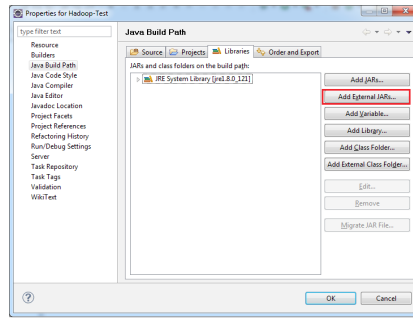
But as we can see, it is full of errors as it is unable to find the MapReduce jars. The jars need to be added to the build path. To do this, we right click on the project and go the sub-menu 'Build Path' and click on 'Configure Build Path.'



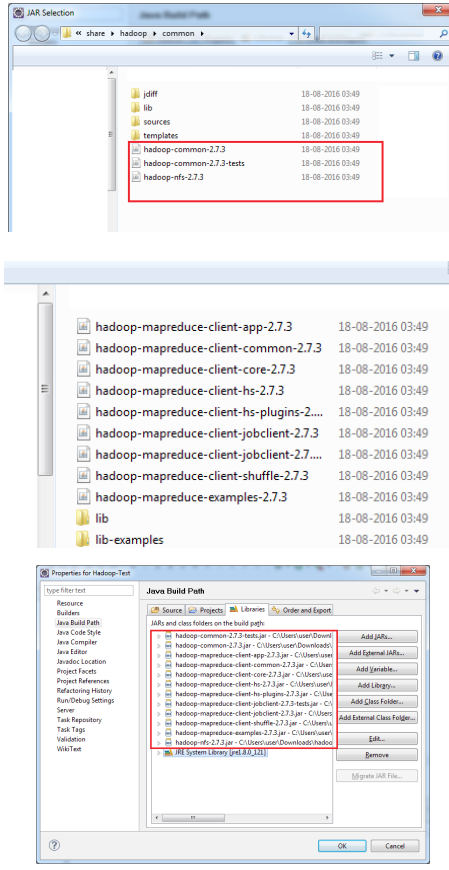
The following window opens up:



We click on ‘Add External Jars.’

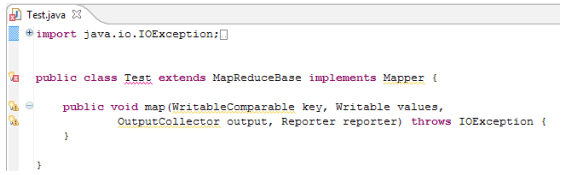


Here, we provide the path of the ‘Hadoop common’ and ‘MapReduce’ jars.

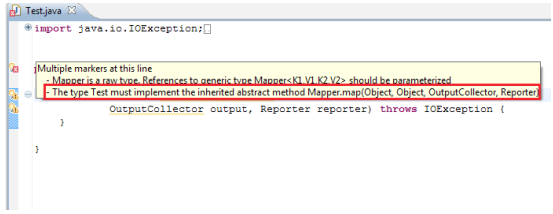


WE click on the button 'OK.'

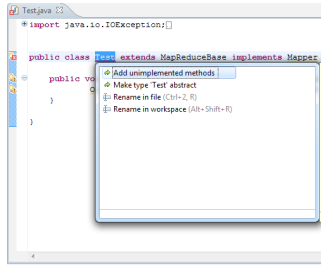
As we can now see in the diagram below, all of the errors except one have disappeared.

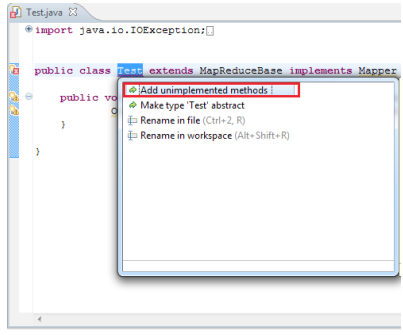


The error is the following:

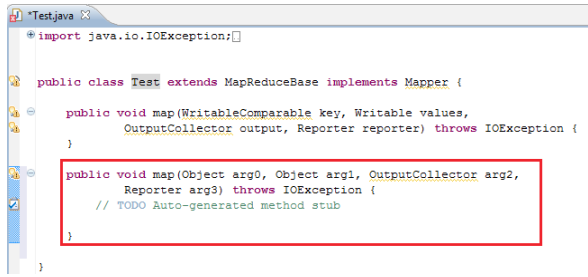


This error can be fixed by clicking on the error symbol and choosing the first available option:

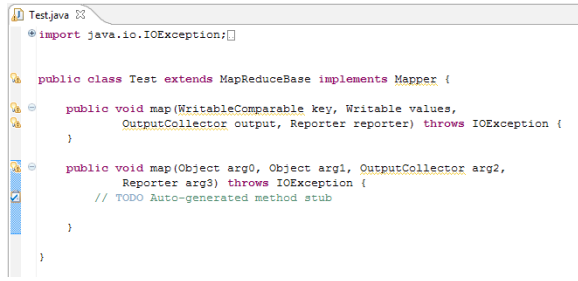




A new method is generated as shown below:



On saving the java file, the error goes away. This is shown below:



Now that we have seen how to create a sample Map/reduce project in Eclipse, we shall now take a look at the approach for the application of data mining with MapReduce to the banking dataset.

3.7.6. Practical Approach for Data Mining Implementation with MapReduce

With the environment set, we introduce in this section the following steps to implement the MapReduce program. The implementation is on a single node cluster because of its simple configuration requirements. However, with MapReduce, the program for one or few nodes is easily scalable to hundreds or thousands of nodes in a transparent fashion.

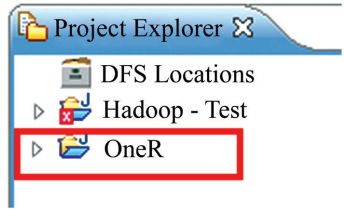
3.7.7 Case Study

The first step is to find a data set and a need for prediction to implement data mining. The dataset used in this example (Moro et al., 2011) is of a Portuguese bank available to public for research purposes. The file is composed of records (45,211), each record holds data about the bank's clients: age, job, education. The last class, which is the target class, is of categorical type with values yes or no, which indicates whether or the client accepted the product (bank term deposit) offered by the bank. The objective is to analyze the data to be able to predict the target class or the customer's

decision using the client's data (explanatory classes).

Now we introduce how we applied the ZeroR and oneR algorithms to the data.

We create a Map/reduce project as shown below:



This project is made up of a Mapper class, a reducer class, a driver class and a test class as shown below:

The mapper class is as follows:

```
package BankOR;

import org.apache.hadoop.MapReduce.Mapper;
import org.apache.hadoop.io.Text;
import org.apache.hadoop.io.LongWritable;
import org.apache.hadoop.io.IntWritable;
import org.apache.commons.logging.Log;
import org.apache.commons.logging.LogFactory;

import java.io.IOException;
import java.util.StringTokenizer;

public class BankORMapper  extends Mapper <LongWritable,Text,Text,Text>
{
  String tempString=null;
        String targetValue = null;
 private static Log log = LogFactory.getLog(BankORMapper.class);

   public void map(LongWritable key, Text value, Context context) throws
IOException, InterruptedException {
    // create iterator over record assuming ';'-separated fields
```

```
StringTokenizer iterator = new StringTokenizer(value.toString(),”,”);

// TODO check number of tokens in iterator
    int tokensCount = iterator.countTokens() ;


// TODO Loop iterator until reaching the target class and save its value in
targetValue String
    int i;
    for ( i = 0 ; i < tokensCount ; i ++ )
    {
        targetValue = iterator.nextToken().toString();
    }

    // Iterate again for each class except the target class
    iterator = new StringTokenizer(value.toString(),”,”);

    for ( i = 0 ; i < tokensCount - 1 ; i ++ )
    {
        tempString = iterator.nextToken().toString();
        switch (i)
        {
            // First column: age
            case 0 :
                String ageCategory = null;
                Integer ageInteger = new Integer(tempString);
            int ageInt = ageInteger.intValue();
                if ( ageInt < 30 )
                    ageCategory = “young”;
                else if (ageInt > 50 )
                    ageCategory = “old”;
                else ageCategory = “average”;
context.write( new Text (“age”), new Text (ageCategory+” “ + targetValue));
```

```
                    break;

                // Second column: job
                case 1 :
context.write( new Text (“job”), new Text (tempString+” “ + targetValue));
                    break;
// Third column: marital status (categorical: “married”,”divorced”,”single”; note: “divorced” means divorced or widowed)
                case 2 :
context.write( new Text (“marital”), new Text (tempString+” “ + targetValue));
                    break;


                // forth column: education
                case 3 :
                    context.write( new Text (“education”), new Text (tempString+” “ +
                                                    targetValue));
                    break;

                // fifth column: has credit yes or no
                case 4 :
context.write( new Text (“credit”), new Text (tempString+” “ + targetValue));
                    break;

                // seventh column: housing yes or no
                case 6 :
context.write( new Text (“housing”), new Text (tempString+” “ + targetValue));
                    break;

                // Eigth column: loan status
                case 7 :
context.write( new Text (“loan”), new Text (tempString+” “ + targetValue));
                    break;
```

```
                    // Ninth column: contact
                    case 8 :
context.write( new Text (“contact”), new Text (tempString+” “ + targetValue));
                        break;

                    // Eleventh column: last contact month
                    case 10 :
context.write( new Text (“month”), new Text (tempString+” “ + targetValue));
                        break;

                    // sixteenth column: poutcome outcome of the previous
marketing campaign
                    case 15 :
context.write( new Text (“poutcome”), new Text (tempString+” “ + targetValue));
                        break;
                }
            }
        }
    }
```

The driver class is as follows:

The Driver class:

```
package BankOR;

import org.apache.hadoop.MapReduce.Counters;
import org.apache.hadoop.MapReduce.Counter;
import org.apache.hadoop.MapReduce.CounterGroup;
import org.apache.hadoop.conf.Configured;
import org.apache.hadoop.conf.Configuration;
```

```
import org.apache.hadoop.io.Text;
import org.apache.hadoop.io.IntWritable;
import org.apache.hadoop.MapReduce.Job;
import org.apache.hadoop.util.Tool;
import org.apache.hadoop.util.ToolRunner;
import org.apache.hadoop.fs.Path;
import org.apache.hadoop.MapReduce.lib.input.FileInputFormat;
import org.apache.hadoop.MapReduce.lib.output.FileOutputFormat;
import org.apache.hadoop.MapReduce.lib.input.TextInputFormat;
public class BankORDriver extends Configured implements Tool {
  public int run(String[] args) throws Exception {
    // check the CLI
    if (args.length != 2) {
System.err.println("usage: hadoop jar -classpath $CLASSPATH:BankOR.jar
BankOR.BankORDriver <inputfile> <outputdir>");
        System.exit(1);
    }
    // setup the Job
    Job job = new Job(getConf());
    job.setJarByClass(BankORDriver.class);
    job.setMapperClass(BankORMapper.class);
    job.setReducerClass(BankORReducer.class);
    job.setInputFormatClass(TextInputFormat.class);
    job.setOutputKeyClass(Text.class);
    job.setOutputValueClass(IntWritable.class);
        job.setMapOutputValueClass(Text.class);
    // setup input and output paths
    FileInputFormat.addInputPath(job, new Path(args[0]));
    FileOutputFormat.setOutputPath(job, new Path(args[1]));

    // launch job syncronously
    return job.waitForCompletion(true) ? 0 : 1;
  }
```

```
  public static void main(String[] args) throws Exception {
    Configuration conf = new Configuration();
    conf.set("MapReduce.output.key.field.separator", ",");
    System.exit(ToolRunner.run(conf, new BankORDriver(), args));
  }
}
```

The test class is as follows:

```
package BankOR;

import org.junit.*;

public class BankORTest {
  private static MapDriver<LongWritable, Text, Text, Text> mapDriver;
  // TODO declare the reduceDriver object
  private static ReduceDriver<Text, Text, Text, FloatWritable> reduceDriver;

  @Before
  private static void setUp() {
    BankOR.BankORMapper mapper = new BankOR.BankORMapper();
    mapDriver = MapDriver.newMapDriver(mapper);
    // TODO instantiate a reducer object and reducer driver
    BankOR.BankORReducer reducer = new BankOR.BankORReducer();
    reduceDriver = ReduceDriver.newReduceDriver(reducer);
  }

  @Test
  private static void testMapper(LongWritable key, Text value, String output) throws IOException {
    String[] outputStringArray = output.split(" ");
    String outputKey = outputStringArray[0];
    String outputValue = outputStringArray[1]+ " "+ outputStringArray[2];
    mapDriver
      .withInput(key, value)
      .withOutput(new Text(outputKey), new Text(outputValue))
      .runTest();
  }
  private static void testReducer(Text key, List<Text> values, String output) throws IOException {
  // TODO implement the testReducer method
    String outputKey = "";
    String[] outputStringArray = output.split("\\s+");
    int i;
    for ( i = 0 ; i < outputStringArray.length - 1 ; i++)
    {
      outputKey= outputKey + outputStringArray[i] + " ";
    }
    String outputValue = outputStringArray[outputStringArray.length - 1];
    reduceDriver
      .withInput(key, values)
      .withOutput(new Text(outputKey), new FloatWritable(Float.parseFloat(outputValue)))
      .runTest();
  }

  public static void main(String[] args) {
    if (args.length != 2) {
      System.err.printf("usage: %s <map | reduce> <inputfile>\n", "ReceiptsTest");
      System.exit(1);
    }
    String value=null, output=null;
    BufferedReader reader;
    try {
      reader = new BufferedReader(new FileReader(args[1]));
      value = reader.readLine();
      output = reader.readLine();
    }
    catch(IOException e) {System.out.println("error reading from input file " + e.toString());}
    setUp();
    if (args[0].equals("map")) {
      try { testMapper(new LongWritable(0), new Text(value), output);}
      catch (Exception e) {System.err.println("error running test: " + value.toString() + " " + output);}
      finally {System.out.println("success");}
    }

      else if (args[0].equals("reduce")) {
      // TODO create a list object for reduce input
      List<Text> reduceInput = new ArrayList<Text>();
      // TODO tokenize the first line from the input file
      StringTokenizer iterator = new StringTokenizer(value, " ");
      // TODO pull out the key from the tokenized line
      Text key = new Text(iterator.nextToken());
      // TODO loop through tokens to add Text to reduce input list
      while(iterator.hasMoreTokens())
        reduceInput.add(new Text (iterator.nextToken()) );

      try { testReducer(key, reduceInput, output);}
      catch (Exception e) {System.err.println("error running test: " + value.toString() + " " + output);}
      finally {System.out.println("success");}
    }

    return;
  }
}
```

The Reducer class is a complex class that performs of lot of operations. The Reducer class is as shown below:

package BankOR;

import org.apache.hadoop.MapReduce.Reducer;

```
import org.apache.hadoop.io.Text;
import org.apache.hadoop.io.FloatWritable;
import org.apache.hadoop.io.IntWritable;
import java.io.IOException;


public class BankORReducer  extends Reducer <Text,Text,Text,FloatWritable> {


  public void reduce(Text key, Iterable<Text> values, Context context) throws IOException,
InterruptedException {
        float recordCount = 0;
        String attributeName = key.toString();
        String[] valueStringArray;
        Text output = null;
        /*
        * the percentage of contributon of the age class in describing the target class
        */
        float ageContribPercent;
        float jobContribPercent;
        float maritalContribPercent;
        float educationContribPercent;
        switch (attributeName)// switching classes or columns
        {
                case "age":// first column is age
                        float youngYesCount = 0;
                        float youngNoCount = 0;
                        float youngDomCount = 0;// the count of the dominant value
(yes or no)
                        String youngDomTargetValue = null;

                        float averageYesCount = 0;
                        float averageNoCount = 0;
                        float averageDomCount = 0;// the count of the dominant value
(yes or no)
                        String averageDomTargetValue = null;

                        float oldYesCount = 0;
```

```
float oldNoCount = 0;
float oldDomCount = 0;// the count of the dominant value (yes
or no)
String oldDomTargetValue = null;
for ( Text value : values)
{
        recordCount ++ ;
valueStringArray = value.toString().split("\\s+"); // splitting value to class value and target
class value
        switch (valueStringArray[0])
        {
            case "young":
                if ( valueStringArray [1].equals(
"yes"))
                    youngYesCount ++;
                if ( valueStringArray [1].equals(
"no"))
                    youngNoCount ++;
youngDomTargetValue = youngYesCount >= youngNoCount ? "yes" : "no";
youngDomCount = youngYesCount >= youngNoCount ? youngYesCount : youngNoCount;
                break;

            case "average":
                if ( valueStringArray [1].equals(
"yes"))
                    averageYesCount ++;
                if ( valueStringArray [1].equals(
"no"))
                    averageNoCount ++;
averageDomTargetValue = averageYesCount >= averageNoCount ? "yes" : "no";
averageDomCount = averageYesCount >= averageNoCount ? averageYesCount :
averageNoCount;
                break;

            case "old":
                if ( valueStringArray [1].equals(
"yes"))
                    oldYesCount ++;
```

```
                                        if ( valueStringArray [1].equals(
"no"))
                                            oldNoCount ++;
oldDomTargetValue = oldYesCount >= oldNoCount ? "yes" : "no";
oldDomCount = oldYesCount >= oldNoCount ? oldYesCount : oldNoCount;
                                        break;

                        }
                    }
ageContribPercent = ( youngDomCount + averageDomCount + oldDomCount ) /
recordCount ;
output = new Text ("AGE: young => " + youngDomTargetValue + " average => " +
averageDomTargetValue +" old => " + oldDomTargetValue + "\n");
                    context.write(output, new FloatWritable (ageContribPercent)
);

                    break;

                case "job":
//type of job (categorical: "admin.","unknownJob","unemployed","management","house
maid","entrepreneur","student",
                     //    "blue-collar","self-employed","retired","technician","s
ervices")
                    float adminYesCount = 0;
                    float adminNoCount = 0;
                    float adminDomCount = 0;// the count of the dominant value
(yes or no)
                    String adminDomTargetValue = null;

                    float unknownJobYesCount = 0;
                    float unknownJobNoCount = 0;
float unknownJobDomCount = 0;// the count of the dominant value (yes or no)
                    String unknownJobDomTargetValue = null;

                    float unemployedYesCount = 0;
                    float unemployedNoCount = 0;
float unemployedDomCount = 0;// the count of the dominant value (yes or no)
                    String unemployedDomTargetValue = null;
```

```
float managementYesCount = 0;
float managementNoCount = 0;
float managementDomCount = 0;// the count of the dominant value (yes o r no)
String managementDomTargetValue = null;

float housemaidYesCount = 0;
float housemaidNoCount = 0;
float housemaidDomCount = 0;// the count of the dominant value (yes or no)
String housemaidDomTargetValue = null;

float entrepreneurYesCount = 0;
float entrepreneurNoCount = 0;
float entrepreneurDomCount = 0;// the count of the dominant value (yes or no)
String entrepreneurDomTargetValue = null;

float studentYesCount = 0;
float studentNoCount = 0;
float studentDomCount = 0;// the count of the dominant value (yes or no)
String studentDomTargetValue = null;
// blue-collar
float blueYesCount = 0;
float blueNoCount = 0;
float blueDomCount = 0;// the count of the dominant value (yes or no)
String blueDomTargetValue = null;
// self employed
float selfYesCount = 0;
float selfNoCount = 0;
float selfDomCount = 0;// the count of the dominant value (yes or no)
String selfDomTargetValue = null;

float retiredYesCount = 0;
float retiredNoCount = 0;
```

```
float retiredDomCount = 0;// the count of the dominant value
(yes or no)
String retiredDomTargetValue = null;

float technicianYesCount = 0;
float technicianNoCount = 0;
float technicianDomCount = 0;// the count of the dominant
value (yes or no)
String technicianDomTargetValue = null;

float servicesYesCount = 0;
float servicesNoCount = 0;
float servicesDomCount = 0;// the count of the dominant value
(yes or no)
String servicesDomTargetValue = null;

for ( Text value : values)
{
        recordCount ++ ;
valueStringArray = value.toString().split("\\s+"); // splitting value to class value and target
class value
        switch (valueStringArray[0])
        {
            case "admin.":
                if ( valueStringArray [1].equals(
"yes"))
                    adminYesCount ++;
                if ( valueStringArray [1].equals(
"no"))
                    adminNoCount ++;
adminDomTargetValue = adminYesCount >= adminNoCount ? "yes" : "no";
adminDomCount = adminYesCount >= adminNoCount ? adminYesCount : adminNoCount;
                break;

            case "unknown":
                if ( valueStringArray [1].equals(
"yes"))
                    unknownJobYesCount
```

```
++;
                                        if ( valueStringArray [1].equals(
“no”))
                                            unknownJobNoCount
++;
unknownJobDomTargetValue = unknownJobYesCount >= unknownJobNoCount ? “yes” :
“no”;
unknownJobDomCount = unknownJobYesCount >= unknownJobNoCount ?
unknownJobYesCount : unknownJobNoCount;
                                        break;

                                    case “unemployed”:
                                        if ( valueStringArray [1].equals(
“yes”))
                                            unemployedYesCount
++;
                                        if ( valueStringArray [1].equals(
“no”))
                                            unemployedNoCount
++;
unemployedDomTargetValue = unemployedYesCount >= unemployedNoCount ? “yes” :
“no”;
unemployedDomCount = unemployedYesCount >= unemployedNoCount ?
unemployedYesCount : unemployedNoCount;
                                        break;

                                    case “management”:
                                        if ( valueStringArray [1].equals(
“yes”))
                                            managementYesCount
++;
                                        if ( valueStringArray [1].equals(
“no”))
                                            managementNoCount
++;
managementDomTargetValue = managementYesCount >= managementNoCount ? “yes”
: “no”;
managementDomCount = managementYesCount >= managementNoCount ?
managementYesCount : managementNoCount;
                                        break;
```

```
                                        case “housemaid”:
                                            if ( valueStringArray [1].equals(
“yes”))
                                                housemaidYesCount ++;
                                            if ( valueStringArray [1].equals(
“no”))
                                                housemaidNoCount ++;
housemaidDomTargetValue = housemaidYesCount >= housemaidNoCount ? “yes” : “no”;
housemaidDomCount = housemaidYesCount >= housemaidNoCount ? housemaidYesCount
: housemaidNoCount;
                                            break;

                                        case “entrepreneur”:
                                            if ( valueStringArray [1].equals(
“yes”))
                                                entrepreneurYesCount
++;
                                            if ( valueStringArray [1].equals(
“no”))
                                                entrepreneurNoCount
++;
entrepreneurDomTargetValue = entrepreneurYesCount >= entrepreneurNoCount ? «yes»
: «no»;
entrepreneurDomCount = entrepreneurYesCount >= entrepreneurNoCount ?
entrepreneurYesCount : entrepreneurNoCount;
                                            break;

                                        case “student”:
                                            if ( valueStringArray [1].equals(
“yes”))
                                                studentYesCount ++;
                                            if ( valueStringArray [1].equals(
“no”))
                                                studentNoCount ++;
studentDomTargetValue = studentYesCount >= studentNoCount ? “yes” : “no”;
studentDomCount = studentYesCount >= studentNoCount ? studentYesCount :
studentNoCount;
                                            break;
```

```
                                        case "blue-collar":
                                            if ( valueStringArray [1].equals(
"yes"))
                                                blueYesCount ++;
                                            if ( valueStringArray [1].equals(
"no"))
                                                blueNoCount ++;
                                            blueDomTargetValue =
blueYesCount >= blueNoCount ?
"yes" : "no";
blueDomCount = blueYesCount >= blueNoCount ? blueYesCount : blueNoCount;
                                            break;

                                        case "self-employed":
                                            if ( valueStringArray [1].equals(
"yes"))
                                                selfYesCount ++;
                                            if ( valueStringArray [1].equals(
"no"))
                                                selfNoCount ++;
                                            selfDomTargetValue =
selfYesCount >= selfNoCount ?
"yes"
: "no";
selfDomCount = selfYesCount >= selfNoCount ? selfYesCount : selfNoCount;
                                            break;


                                        case "retired":
                                            if ( valueStringArray [1].equals(
"yes"))
                                                retiredYesCount ++;
                                            if ( valueStringArray [1].equals(
"no"))
                                                retiredNoCount ++;
retiredDomTargetValue = retiredYesCount >= retiredNoCount ? "yes" : "no";
retiredDomCount = retiredYesCount >= retiredNoCount ? retiredYesCount : retiredNoCount;
                                            break;
```

```
                                    case "technician":
                                        if ( valueStringArray [1].equals(
"yes"))
                                            technicianYesCount ++;
                                        if ( valueStringArray [1].equals(
"no"))
                                            technicianNoCount ++;
technicianDomTargetValue = technicianYesCount >= technicianNoCount ? "yes" : "no";
technicianDomCount = technicianYesCount >= technicianNoCount ? technicianYesCount
: technicianNoCount;
                                        break;

                                    case "services":
                                        if ( valueStringArray [1].equals(
"yes"))
                                            servicesYesCount ++;
                                        if ( valueStringArray [1].equals(
"no"))
                                            servicesNoCount ++;
                                        servicesDomTargetValue =
servicesYesCount >=
servicesNoCount ? "yes" : "no";
servicesDomCount = servicesYesCount >= servicesNoCount ? servicesYesCount :
servicesNoCount;
                                        break;
                                }
                            }

jobContribPercent = ( adminDomCount + unknownJobDomCount + unemployedDomCount
+ managementDomCount + housemaidDomCount + entrepreneurDomCount
+ studentDomCount + blueDomCount + selfDomCount + retiredDomCount +
technicianDomCount + servicesDomCount ) / recordCount ;

output = new Text ("JOB: admin => " + adminDomTargetValue + " unknownJob =>
" + unknownJobDomTargetValue +" unemployed => " + unemployedDomTargetValue
+ "\n management => " + managementDomTargetValue + " housemaid => " +
housemaidDomTargetValue + " entrepreneur => " + entrepreneurDomTargetValue + " \n
student => " + studentDomTargetValue + "blue-collar => " + blueDomTargetValue + "
self-employed => " + selfDomTargetValue + " retired => " + retiredDomTargetValue + "
technician => " + technicianDomTargetValue + " services => " + servicesDomTargetValue
+ "\n");
```

```
context.write(output, new FloatWritable (jobContribPercent) );


                    break;


                    case "marital":
//marital status (categorical: "married","divorced","single"; note: "divorced" means
divorced or widowed)

                    float marriedYesCount = 0;
                    float marriedNoCount = 0;
                    float marriedDomCount = 0;// the count of the dominant value
(yes or no)
                    String marriedDomTargetValue = null;

                    float divorcedYesCount = 0;
                    float divorcedNoCount = 0;
                    float divorcedDomCount = 0;// the count of the dominant value
(yes or no)
                    String divorcedDomTargetValue = null;

                    float singleYesCount = 0;
                    float singleNoCount = 0;
                    float singleDomCount = 0;// the count of the dominant value
(yes or no)
                    String singleDomTargetValue = null;

                    for ( Text value : values)
                    {
                              recordCount ++ ;
valueStringArray = value.toString().split("\\s+"); // splitting value to class value and target
class value
                              switch (valueStringArray[0])
                              {
                                        case "married":
                                                  if ( valueStringArray [1].equals(
"yes"))
                                                            marriedYesCount ++;
```

```
                                        if ( valueStringArray [1].equals(
"no"))
                                            marriedNoCount ++;
marriedDomTargetValue = marriedYesCount >= marriedNoCount ? "yes" : "no";
marriedDomCount = marriedYesCount >= marriedNoCount ? marriedYesCount :
marriedNoCount;
                                        break;

                                    case "divorced":
                                        if ( valueStringArray [1].equals(
"yes"))
                                            divorcedYesCount ++;
                                        if ( valueStringArray [1].equals(
"no"))
                                            divorcedNoCount ++;
                                        divorcedDomTargetValue =
divorcedYesCount >=
divorcedNoCount ? "yes" : "no";
divorcedDomCount = divorcedYesCount >= divorcedNoCount ? divorcedYesCount :
divorcedNoCount;
                                        break;

                                    case "single":
                                        if ( valueStringArray [1].equals(
"yes"))
                                            singleYesCount ++;
                                        if ( valueStringArray [1].equals(
"no"))
                                            singleNoCount ++;
                                        singleDomTargetValue =
singleYesCount >=
singleNoCount ? "yes" : "no";
singleDomCount = singleYesCount >= singleNoCount ? singleYesCount : singleNoCount;
                                        break;
                                }
                            }
maritalContribPercent = ( marriedDomCount + divorcedDomCount + singleDomCount )
/ recordCount ;
output = new Text ("MARITAL: married => " + marriedDomTargetValue + " divorced =>
```

```
" + divorcedDomTargetValue +" single => " + singleDomTargetValue + "\n");
                    context.write(output, new FloatWritable
(maritalContribPercent) );
                    break;

                    case "education":
                    // (categorical: "unknownEdu","secondary","primary","terti
ary")
                    float unknownEduYesCount = 0;
                    float unknownEduNoCount = 0;
float unknownEduDomCount = 0;// the count of the dominant value (yes or no)
                    String unknownEduDomTargetValue = null;

                    float primaryYesCount = 0;
                    float primaryNoCount = 0;
                    float primaryDomCount = 0;// the count of the dominant value
(yes or no)
                    String primaryDomTargetValue = null;

                    float secondaryYesCount = 0;
                    float secondaryNoCount = 0;
                    float secondaryDomCount = 0;// the count of the dominant
value (yes or no)
                    String secondaryDomTargetValue = null;

                    float tertiaryYesCount = 0;
                    float tertiaryNoCount = 0;
                    float tertiaryDomCount = 0;// the count of the dominant value
(yes or no)
                    String tertiaryDomTargetValue = null;

                    for ( Text value : values)
                    {
                              recordCount ++ ;
valueStringArray = value.toString().split("\\s+"); // splitting value to class value and target
class value
                              switch (valueStringArray[0])
                              {
```

```
                                        case "unknown":

                                                if ( valueStringArray [1].equals(
"yes"))
                                                        unknownEduYesCount
++;
                                                if ( valueStringArray [1].equals(
"no"))
                                                        unknownEduNoCount
++;
unknownEduDomTargetValue = unknownEduYesCount >= unknownEduNoCount ? "yes"
: "no";
unknownEduDomCount = unknownEduYesCount >= unknownEduNoCount ?
unknownEduYesCount : unknownEduNoCount;
                                                break;

                                        case "primary":

                                                if ( valueStringArray [1].equals(
"yes"))
                                                        primaryYesCount ++;
                                                if ( valueStringArray [1].equals(
"no"))
                                                        primaryNoCount ++;
primaryDomTargetValue = primaryYesCount >= primaryNoCount ? "yes" : "no";
primaryDomCount = primaryYesCount >= primaryNoCount ? primaryYesCount :
primaryNoCount;
                                                break;

                                        case "secondary":

                                                if ( valueStringArray [1].equals(
"yes"))
                                                        secondaryYesCount ++;
                                                if ( valueStringArray [1].equals(
"no"))
                                                        secondaryNoCount ++;
secondaryDomTargetValue = secondaryYesCount >= secondaryNoCount ? "yes" : "no";
secondaryDomCount = secondaryYesCount >= secondaryNoCount ? secondaryYesCount :
```

```
secondaryNoCount;
                                                    break;

                                                case "tertiary":

                                                    if ( valueStringArray [1].equals(
"yes"))
                                                        tertiaryYesCount ++;
                                                    if ( valueStringArray [1].equals(
"no"))
                                                        tertiaryNoCount ++;
                                                    tertiaryDomTargetValue =
tertiaryYesCount >=
tertiaryNoCount ? "yes" : "no";
tertiaryDomCount = tertiaryYesCount >= tertiaryNoCount ? tertiaryYesCount :
tertiaryNoCount;
                                                    break;
                                            }
                                    }

educationContribPercent = ( unknownEduDomCount + primaryDomCount +
secondaryDomCount + tertiaryDomCount ) / recordCount ;
output = new Text ("EDUCATION: unknown => " + unknownEduDomTargetValue + "
primary => " + primaryDomTargetValue + " secondary => " + secondaryDomTargetValue
+" tertiary => " + tertiaryDomTargetValue + "\n");
                                    context.write(output, new FloatWritable
(educationContribPercent) );
                                    break;
                        }

    }
}
```

In the following table we breakdown the algorithms and classes by explaining what is happening in each phase: driver, mapper and reducer.

Point of comparison	**ZeroR**	**OneR**
Driver	Normal configuration for mapper, reducer classes. Input class format is TextInputFormat. Output key is Text, Output value is float writable.	
Mapper	**Output key** It is not of use in the reducer so we just output the text "target" for each record. **Output value** The value of the target class	**Output key** The name of each the predictor class, e.g., "education," "job" … for each record. **Output value** text of two strings separated by space: value of predictor and value of target class.
Unit Test record (input data set for both algorithms)	57;entrepreneur;married;secondary;no;2971;no;no;cellular;17;nov;361;2;188;11;other;no	
Unit Test outcome	Target no	Age average no
Reducer	Count yes and no count and outputs the percentage of the value with the higher percentage.	Count yes and no for each value of each class. Outputs a rule for each predictor expecting the target value of the higher count for each value of the predictor. Outputs the accuracy for each rule if applied the data. The rule with the highest accuracy is the concluded rule of the algorithm.
Unit Test outcome	Yes 60.0	AGE: young => yes average => no old => no 0.5871424

The above table shows the logic used in each algorithm, now we will examine the results and detail the experiments done on data using the OneR algorithm.

3.7.8. Results

3.7.8.1. ZeroR Results

ZeroR learning algorithm was run on the whole data, it gave an accuracy of 53.97%. As mentioned before, this percentage is useful as a benchmark. With this result in hand, we now examine the results of the OneR with the

capability of judging how good it is, compared to the benchmark.

3.7.8.2. OneR Results

The first step in the application of this algorithm is the building of the classifier. Here, we examine what is the least possible portion of data can be used to build the classifier. Generally, the more data used for learning the more accurate the classifier would be as it learns from a wider range of data. In the search for a good threshold to build a classifier with good accuracy we started by learning from 10% of the data.

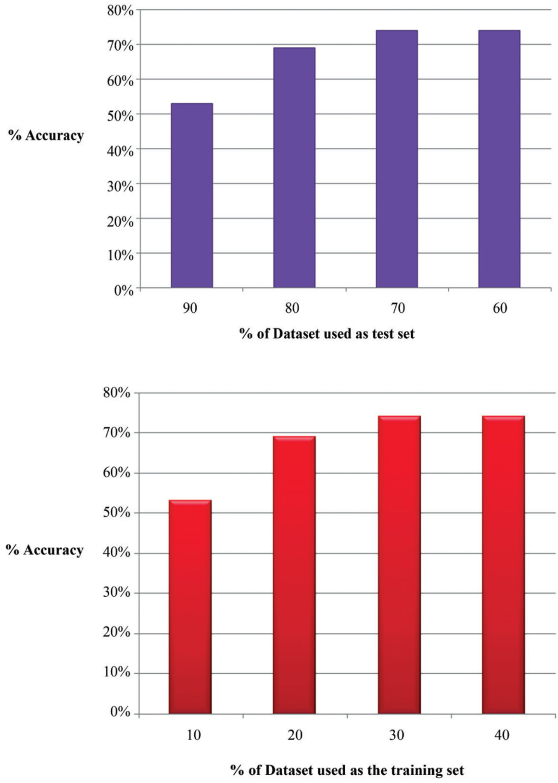
Next, we use the rule obtained and modify the reducer to predict the target class according to the rule then check the prediction against the actual value on the remaining 90% of the data and outputs the accuracy. We repeat the same procedure for 20%, 30% of the data until the accuracy reached is satisfactory and remains almost constant with the increase of data used in learning.

The following table shows the obtained results.

Training set %	**Dataset %**	**Rule**	**Accuracy**
10	90	Job: If Student => yes, otherwise => no	53% (lower than ZeroR, totally rejected)
20	80	Job: management, student => yes, otherwise => no	69 % (Satisfactory)
30	70	Job: management, student, retired => yes, otherwise => no	74% (Good)
40	60	Job: management, student, retired => yes, otherwise => no	74% (Good and the same as the previous experiment)

As can be seen from the table, using only 10% of the data produced a rule that is not acceptable as its accuracy is less than ZeroR. Increasing the proportion of data used in building the classifier increases accuracy. We noticed that there is no change between 30% and 40% so we wanted to ensure that learning from 30% of the data provides a rule that forms a good representation of the data. Thus, we repeated the OneR algorithm for another 30% of random data and the same rule was derived for four trials, where we ensured that each record was used at least once in building the classifier. Consequently, we were able to derive a rule using only 30% of the data which provides a good representation of the whole data as it gives an accuracy which is much higher than the accuracy obtained from ZeroR.

The above data can be represented in a visual manner as shown below:



The results obtained confirm that Data mining proves to be a very powerful process to be applied in business entities to help make efficient decisions. In our case study, the bank will concentrate more on customers satisfying the rule obtained from OneR, so that higher percentages of offers are accepted. In real life, the data that is available is much bigger which makes the learning process even stronger and more precise. Moreover, MapReduce model is a suitable tool to implement data mining, because of its parallel nature and ability to perform efficiently in a distributed environment. Its main strength points are the model's applicability to almost all data analysis problems, handling distributed systems challenges implicitly, sending programming tasks to where data resides in the cloud, rather than the opposite which minimizes network traffic.

4
CHAPTER

APPLICATIONS OF DATA MINING IN MEDICINE

CONTENT

4.1. INTRODUCTION

The healthcare and medicine industry is an industry that is constantly growing and evolving at a high speed. There are new inventions and marvelous discoveries on a daily basis in this field. There exist a great number of research communities (government and private) that are continuously working to find solutions to all types of diseases that plague the world. Additionally, the medicine industry is an industry that has been in existence since centuries and will continue to grow. This industry affects every person and hence naturally it generates and produces a lot of data every second. Data about patients such as their personal information, medical condition, treatment plan, course of action long-term, etc. are stored. Additionally, there is another different bank of information in terms of medicine and drugs that are stocked and produced daily. All this information is highly complex in nature and each hospital, pharmaceutical company, and different organizations possess different types of information at their end.

This industry keeps on gaining more and more data that is simply being stored passively. Data is being generated each second and stored and nothing is being done to use this data. Hence sometimes this industry is deemed to be rich in information yet poor in its knowledge. There are large volumes and quantities of data stored world over within the medicine and healthcare sectors. These vast volumes and stocks of data are usually never "mined." This industry has always lacked the presence of effective means, methods and tools that provide useful insight in this data pool and uncovers hidden relationships between different elements that play a part in this industry.

For example, doctors are constantly faced with difficult decisions all the time regarding diagnosis, treatment plan and choice of medicines. They have no reference expect their own experience and that of other doctors and patients. This situation could benefit from a more reliable way of providing help to patients. The same goes for the medicine section where new drugs are being developed daily that need to be researched first and then tested. Additionally, the pharmaceutical industry may benefit greatly if they were to know what drugs are the most needed, what is preferred, etc.

The healthcare, medicine and pharmaceutical industry store large volumes of data that potentially have hidden precious knowledge within it. Together, all this data does not make sense on its own. The potential for discovery, uncovering of hidden rules, patterns and relationships often tends to go unexploited in this sector. But, it is fertile land for the field of

data mining. Data mining is being used in several domains for commercial as well as scientific purposes. It has found numerous applications in the business and scientific domain so far. The industry of medicine has started appealing more and more to the field of data mining due to infinite applications of data mining in various sectors. Data Mining has been used and implemented successfully in a ton of applications such as, customer relationship management, engineering, marketing, expert predictions, energy data mining, mobile computing, mobiles, and web mining.

The data mining techniques are gaining a lot of popularity as it is capable of handling large quantities of data and extract worthwhile information from within this data. Data mining can mine the information in a precise and effective manner; being able to uncover relationships, correlations and association among different objects in the medical industry. The medical industry proves to be an excellent source of knowledge as it has been in existence since a long period of time and hence has already decades of data that has been safely stored over time. It has boundless applications in the medicinal field due to its limitless applications in terms of techniques and methods that can be used to mine the data in an effective, appropriate and detailed manner. Valuable, useful information and knowledge can be gleaned from the employment of data mining techniques in the medicine and healthcare system. Scientific data mining is slightly different form regular commercial data mining as the nature of the data sets often vary a lot as compared to those of traditional data mining applications that are market driven.

The field of data mining can also be exceptionally beneficial in making insightful predictions that can be further used for decision-making processes. Data mining algorithms that can be applied in the medicine and healthcare industry can play a noteworthy part in the analysis, diagnosis and prediction of several types of diseases. Thanks to the new and continuously advancing researches in the medicine and healthcare domain, more and more data is available for use of data mining. Owing to the current advancements in the medical field along with radical changes, the current world is acquiring more and more information and techniques; data mining can be one of the most optimal approaches that can help approximate the future trends and events and help predict future consequences.

The main hurdle faced by the medicine and healthcare systems is the cultivation or construction of useful information using different heterogeneous sources with a lot of raw data which is why data mining is

well suited to this field. It can help use these bundles of data to generate something useful, meaningful and potentially very important that will help improve the way of life for a lot of people, improve the system and help save costs.

Over the last decade, with the rise of database management systems and data warehouses, the use of data mining in the field of medicine and healthcare has increased significantly. It has been successfully use to determine trends, patterns using its analysis methods and helped in better decision making due to its infinite potential and effectiveness. Successful implementations have been made in this industry in terms of the predictions too as data mining has fortuitously been able to predict different kinds of diseases and symptoms. Several Data Mining methods such as clustering, regression analysis, association and classification have been used in the medicine and healthcare domain. For example, all the data that is stocked can be used to first and foremost build medical profiles of patients. These profiles can be then further used to predict the possibility of a patient falling sick and being diagnosed with diseases such as blood pressure, cholesterol, heart attacks, cancer, etc. Significant knowledge can be gained from these profiles that can be used everywhere in these sectors.

Due to the depth of data in the sectors of medicine, Hospitals and Pharmaceuticals, there exist a variety of applications of data mining in these areas in the section of medicine, medical devices, hospital management and so on. The process of unearthing of information is long complex process involving analysis of the domain data, pre-processing, processing, creation of data sets and training sets, transformation of data and then finally discovery of information.

We shall now take a look at some of the applications of data mining in the medicine and healthcare industry.

Some of the most common applications are:

- Developing Patient Profiles
- Prediction of diseases
- Treatment Effectiveness calculation
- Patient and Healthcare management (Hospital Management)
- Fraud Detection
- Medical Device Analysis
- Pharmaceutical industry applications
- Marketing

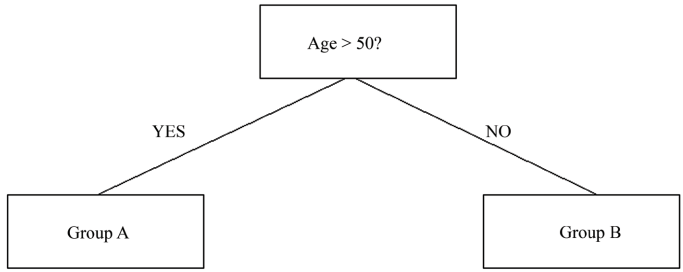
We shall now look at each of these applications in brief:

a. Developing Patient Profiles: Owing to the rise of better database management systems, the medicine and healthcare industry can now store their patient and customer data in a better and organized manner. This data is vast and well-organized across different sectors such as the hospital, pharmacy and the concerned medical professionals. The healthcare industry contains a lot of data about all of their patients and this data can be used to build profiles on patients so that this information can further be used for research purposes. The hospitals can benefit from customer/patient profiles to quickly extract information about them and also maybe look for patients with similar medical histories.

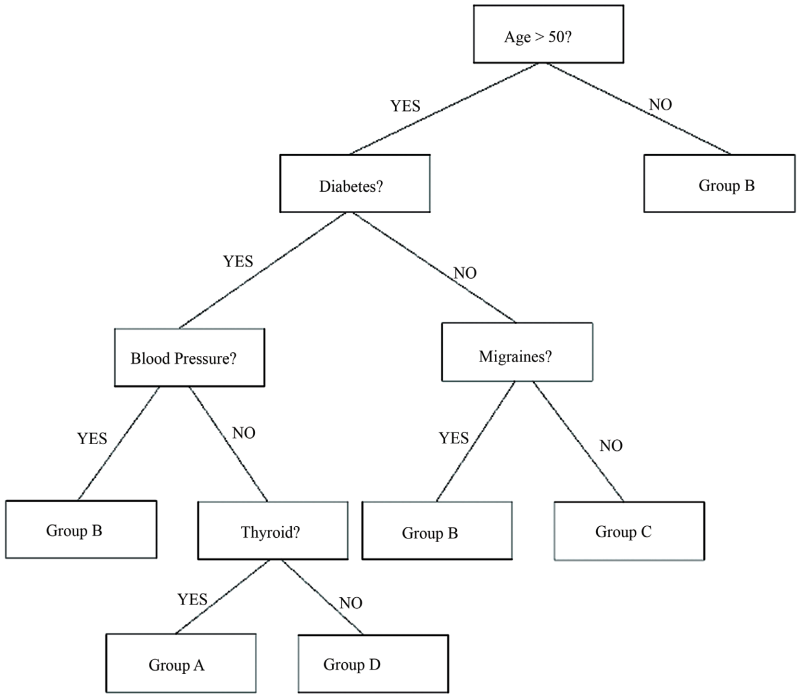
For example, in a large hospital that has several attending doctors, a large amount of patients are treated by different doctors. If one doctor has a patient with a disease and his treatment is not working he is sometimes stuck for alternative courses. Sometimes he can benefit by consulting one of his colleagues for help. But in case of large hospitals with several attending doctors, he may not know the right person to contact. In such a case, if he has access to all patient profiles and their respective treatment courses, he can easily look for other treatment plans that have shown better results. He can then choose the profile with the most similarity to his current profile and then try contacting the doctor who worked on the patient to get more insight. This technique can help doctors all over.

A patient profile can be developed based on the type of ailment, their age, their treatment plan, etc.

A simple classification that groups customers into different classes is shown below:



The above decision tree can be refined to be more detailed as follows:

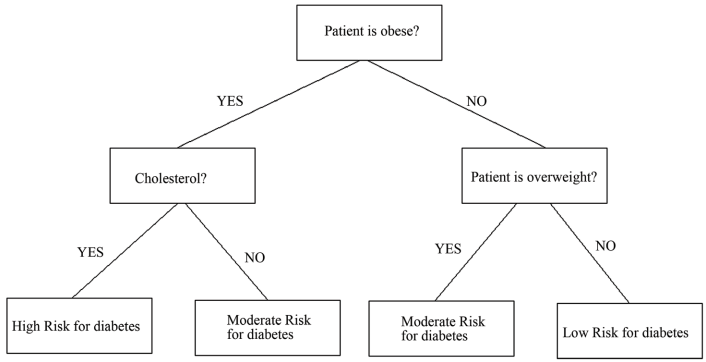


Different profiles can be built based on age, patient history, ailment, treatment plan, geographical data, etc.

b. *Prediction of diseases:* Doctors and healthcare professionals all over the world are always faced by dilemmas regarding the correct diagnosis of patients with accuracy. It is always possible that a professional can make mistakes in their diagnosis as they too after all are human. The cost and impact of such mistakes are sometimes enormous as lives of people are at stake. A minuscule error in diagnosis or prognosis can lead to many fatalities. Doctors are under a lot of pressure to deliver better results but they may not have enough resources at hand.

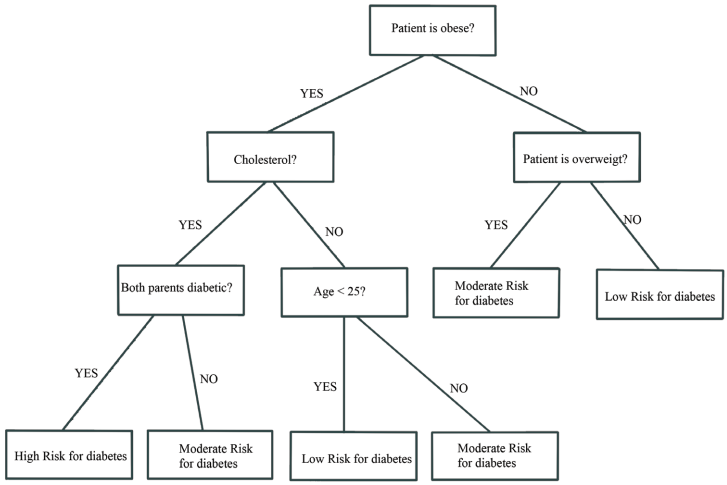
This is where data mining can come in handy. Data mining is capable of analyzing large volumes of data and predicting whether he is at risk or not. Data mining has been successfully been applied in the medicine domain to predict coronary heart diseases, predict possibility of psychosis, etc.

Below is a simple example of prediction patients as high risk or low risk based on their medical history:

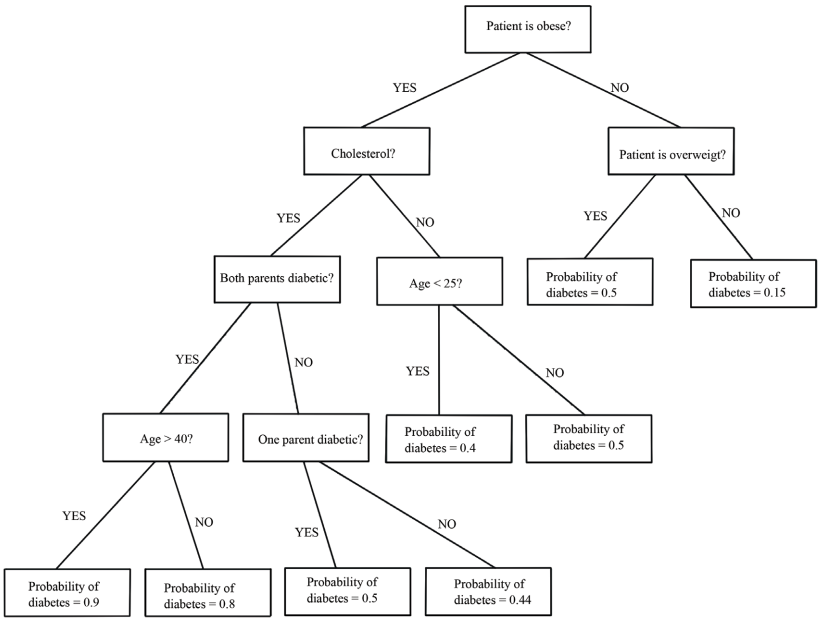


The above tree can be refined to add additional parameters such as family history and age as well.

This is shown below:



Another possibility is the use of decision tree and classification and prediction purposes. Building upon the same example previously shown above, we predict the probability of a patient developing diabetes.



Data mining can also be used to predict other diseases such as heart problems, hypertension and many other diseases using different techniques such neural networks, classification, regression analysis, etc.

Association rule mining is a very well used technique in the medical industry. Based on patient profiles and the tons of medical data that is captured and stored, certain rules can be developed that help the process of analysis and help professionals in this domain make better decisions.

Based on the patient profile and his history many rules can be formed.

Some examples of rules are shown below:

IF Age >45

AND

IF Blood Pressure > 160/80

THEN Heart Problem = YES

IF Age=65

AND Heart rate>75

AND

Blood pressure>150/80
THEN Heart problem=YES

IF (Age=45 AND Heart rate>75)
IF Age>70
AND
IF DIABETES = YES
AND
IF Heart Rate> 75
THEN Heart Problem = YES

IF Age > 40
AND IF DIABETES = YES
AND
IF Heart Rate > 80
AND
IF Blood Pressure > 180/100
THEN Heart problem=YES

IF (Age=85 AND Heart rate>75)
OR
IF (Age = 75 AND Heart Rate> 77)
THEN Heart problem=YES

IF AGE > 75
AND
IF Blood pressure > 139/65
AND
IF Cholesterol = YES
AND
IF Heart Rate > 75

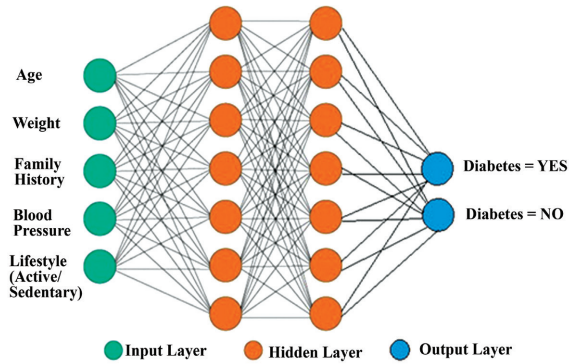
AND
IF Previous history of hypertension = YES
THEN Heart problem=YES

IF AGE < 35
AND
IF Blood pressure > 139/65
THEN Heart problem=NO

IF AGE <40
AND
IF Blood pressure > 170/80
AND
IF Cholesterol = YES
AND
IF Previous history of hypertension = YES
THEN Heart problem=YES

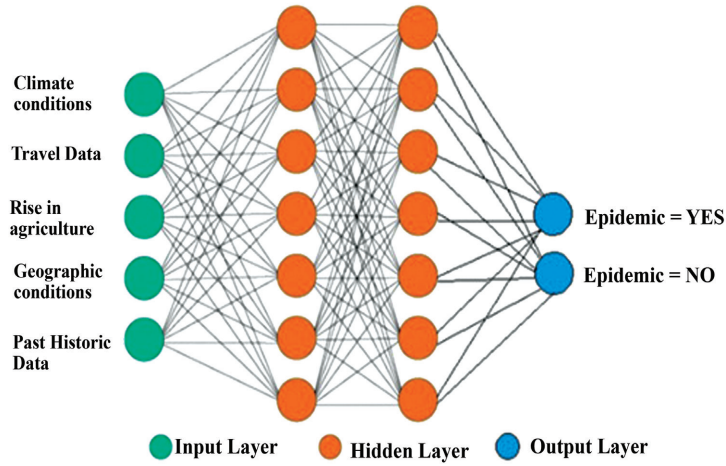
Neural networks can also be used to predict the possibility of diseases based on several parameters such as age, weight, family history, cholesterol, etc.

One such neural network that predicts the possibility of diabetes giving a YES/NO answer is shown below:



Neural networks can also be used to predict epidemics. Sometimes there exist some contagious diseases that are transmitted by the air or by touch and as people are in close proximity to one another in public places, there are high chances of many people being infected. Sometimes, the number of cases that are reported needs to be constantly monitored in order to predict if there will be an epidemic or not. Neural networks can be used to this effect.

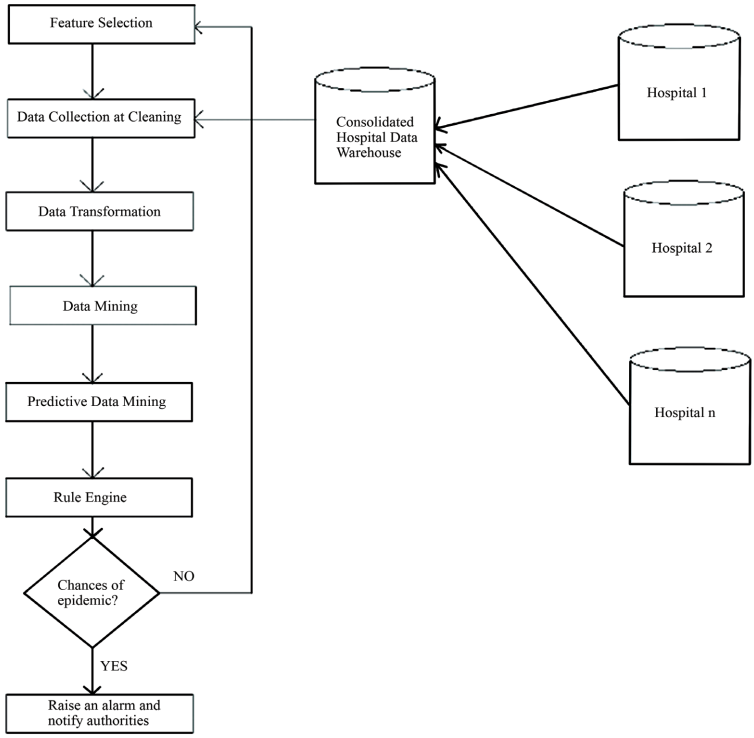
A sample is shown below:



The data model developed can be run once there is a significant rise in the number of cases of a contagious disease. Such models can be extremely used to government agencies such the CDC which can employ preventive measures in order to avoid a potential outbreak of the disease and thereby save a lot of lives and costs as well.

Additionally, a monitoring system can be put into place that constantly monitors records from all the hospitals and analyses the data for diseases of a specific type and if a certain threshold is reached, an alert is sent to the respective authorities who can then take appropriate action.

The steps followed are shown below:



c. Treatment Effectiveness Calculation: The medical healthcare professionals are continuously looking for ways and means to improve their system of diagnosis so as to achieve better results. They continuously look for ways to improve upon their current treatment plans.

Earlier, due to lack of advanced databases, patient data was stored in the form of hard copies such as paper or floppy disks. Hence, the analysis of different treatment plans for different patients was not possible as the data was not consolidated and even if it was, it was done manually. But, due to the development of advanced database systems and multi-dimensional systems, the issue of storage is no longer a problem. All the patient data is now stored in databases and can be accessed easily.

But, even though the data is readily available, it is hard to keep track of the treatment plans due to the vast nature and range of diseases and doctors. Even within a hospital, there are often many doctors that maybe use similar treatment plans. Additionally, the hospitals also may need to evaluate the use of different treatment plans for different groups of patients and decide which one is more effective. Further, the medical institutions may want information regarding the effectiveness of particular drugs over others so as

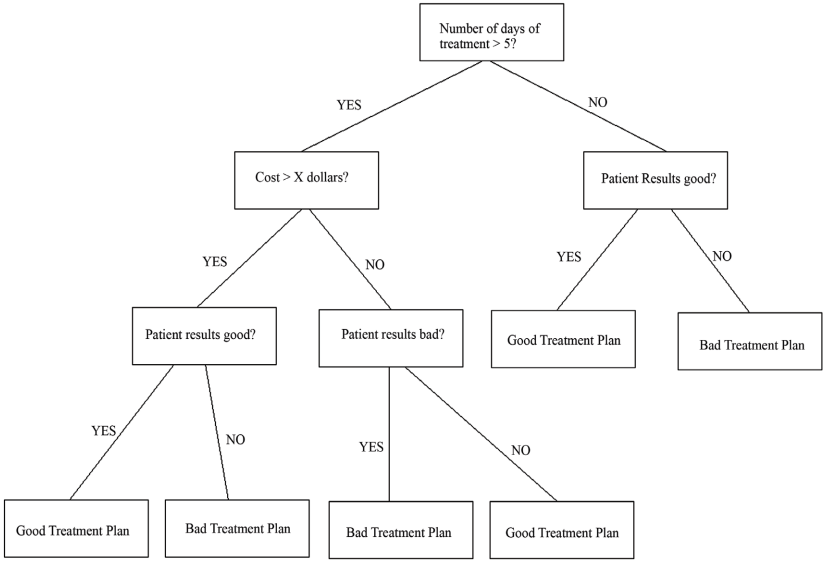
to provide better yet cheaper healthcare to its patients.

This is where data mining comes in handy. Data mining can be used efficiently and effectively to analyze data from different patients with or without selective treatment plans, their symptoms, causes etc. to determine which treatment option that was followed yielded better results under what type of circumstances. Data mining can also be used to define a set of treatments that are standardized for a particular set of diseases.

We shall now see a few examples of the application of data mining for calculation of treatment effectiveness.

As discussed above, data mining can be used to study effectiveness of treatment plans. The effectiveness of treatment plan can be quantified on the basis of different parameters such as severity of the disease, patient reports, etc.

An example is shown below:



The decision tree shown above is a random sample based on randomly chosen factors for explaining purposes. But, in reality, the actual decision trees that are used are extremely complicated.

In terms of standardization, different rules can be uncovered with the help of data mining.

A few examples are shown below:

IF drug X works successfully for disease Y > 99% of the time

THEN

"DRUG X IS THE STANDARD DRUG TO BE PRESCRIBED FOR DISEASE Y"

IF drugs (X + Y) works for the disease A 99.99% of the time,

THEN

"THE TREATMENT COURSE (X+Y) IS THE STANDARD TREATMENT TO BE PRESCRIBED FOR DISEASE A"

IF drugs (X+Y) or drugs (X+Z) works for the disease B > 98% of the time,

THEN

"THE TREATMENT COURSE (X+Y) OR (X+Z) IS THE STANDARD TREATMENT TO BE PRESCRIBED FOR DISEASE B".

IF drugs (X+Y+Z) or drugs (X+Z+A) works for the disease C > 99% of the time,

THEN

"THE TREATMENT COURSE (X+Y+Z) OR (X+Z+A) IS THE STANDARD TREATMENT TO BE PRESCRIBED FOR DISEASE C".

d. Patient and Healthcare Management (Hospital Management): Data is being stored on a large-scale in the medicine and healthcare industry. As more and more data gets accumulated it is of essence that this data be stored and represented in a more consolidated view. Hospitals can greatly benefit from the applications of data mining that help better the organization and retrieval of customer/patient data, staff data and medicinal data. Data mining can help provide global views of the data as well as detailed views on demand and this can easily help the concerned people.

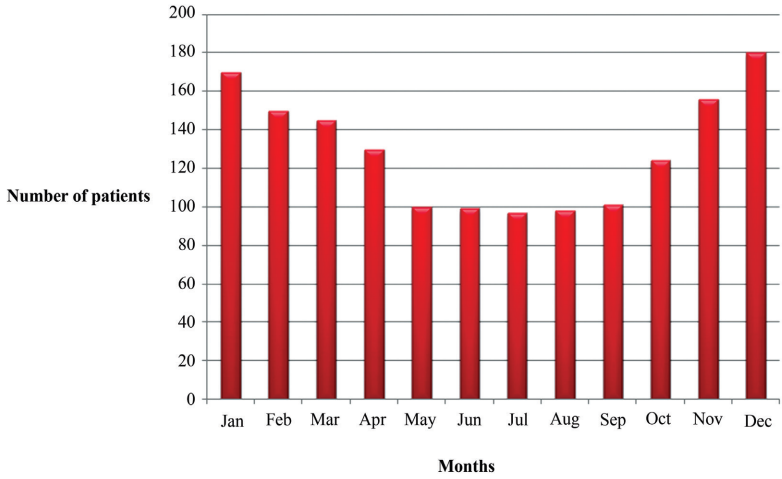
Different views can also be prepared for different type of end users. For instance, a view can be prepared for doctors, another can be done for the medical staff, and one view can be added for the administration and a view for the patient data as well. This gives the possibility of segregating the data based on the end user and also the possibility of consolidation of the data as well.

Another possibility is patient management. Usually doctors are very busy and

they need to stick to fixed schedules when they see their patients. So, when a patient doesn't show up, his schedule is disturbed. Sometimes this is not an issue. But, when the doctor is required to be present at a different hospital and has constraints, he could waste precious time waiting for his patient. Sometimes, it turns out that the patient simply forgot about the appointment and hence didn't show up. Hence, at such times a monitoring system can be put in place to remind the patients that they have an appointment due. Such systems are already put in place.

Data mining can be used successfully in the generation of patient-specific rules that can help the doctor better prepare for his appointment. Data mining can be applied on the data of a particular doctor, to see which time is the busiest for him and when he is likely to have more free time.

Consider the following sample data set:



The above set shows the amount of patients a particular doctor has seen over the past year on a monthly basis. The consolidation functionality of data mining helps gather this information. This information can be used to infer information regarding the schedule of the doctor.

Some inferences that can be extracted from the above data are as follows:

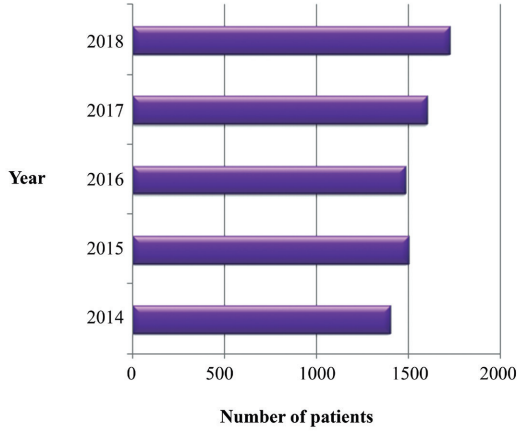
- The doctor is very busy in the month of December.
- During the months from May to September (including) the number of patients stays more or less stable.
- There is an exponential rise in the number of patients from the month of October until February. This inference can further lead

to the inference – 'The number of patients increases over the winter season.'

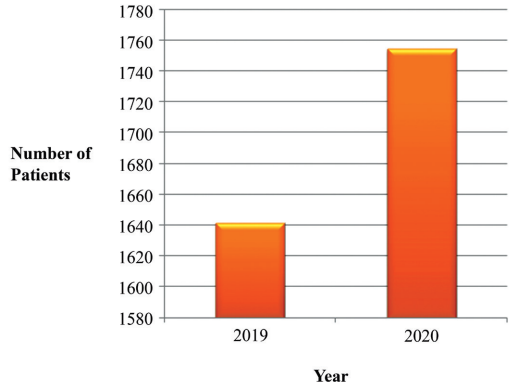
- The doctor sees a maximum of 180 patients each month and a minimum of 97 patients per month.

The doctors data stored over a range of years can be used to perform predictions regarding the number of patients a doctor is likely to see the following year.

Consider the following sample data set for a doctor:



Based on this data, predictions can be made for the upcoming years as shown below:



Based on the predictions, the doctor can develop a better scheduling system for himself and his patients so as to achieve the most optimal results.

e. Fraud Detection: As previously discussed fraud detection is a

common application in many sectors such as engineering, management, etc. The same goes for the medicine and healthcare industry. There are different types of fraud that are possible in this sector and the organizations involved can lose a lot of revenue and time due to cases of fraud and abuse of the medical system.

For example, detection of fraudulent practices is a definite application of data mining in the healthcare industry. Another area that needs analysis is the medical claims. The number of cases of medical fraud has been on the rise where people abuse government benefits and claim medical insurance fraudulently.

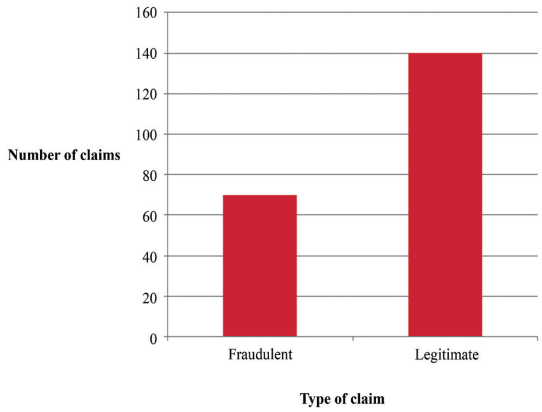
These practices cause the medical industry to lose lot of money and funds. Hence, this industry is now showing more and more interest in data mining to help implement methods to detect fraudulent claims. Usually outlier detection combined with support vector machines can be used in such cases. Other algorithms such as Classification, clustering, k-means have also been successfully implemented in the formation of clusters.

We shall now take a look at a few sample examples.

Consider the following data set that contains the past records of fraudulent claims for Paralysis disease for a year.

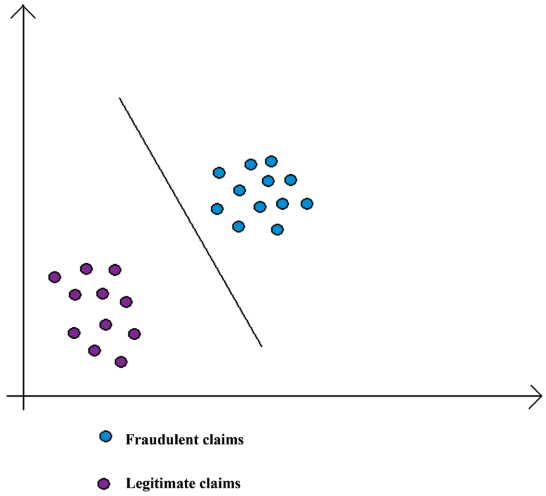
Fraudulent	**Legitimate**
70	140

This data is represented as follows:



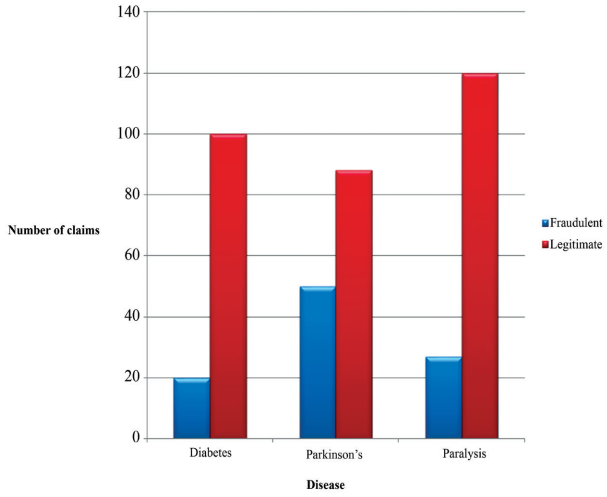
Each and every one of these claims can be studied and classified into different groups based on their nature using clustering techniques. This offers a more detailed view.

This is shown below:

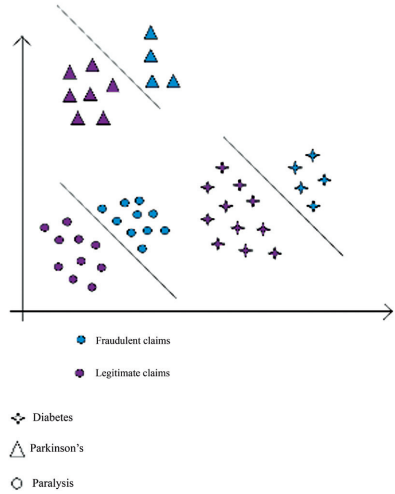


Different techniques can be used to detect claims that are not legitimate such as clustering and k-means. Sometimes only one technique is not enough to detect such claims and hence, a combination of different methods is used so that the accuracy level is high.

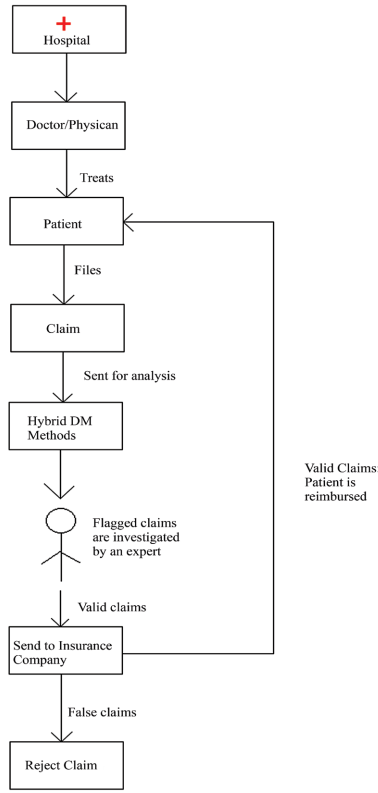
The above clustering can be done on different diseases at the same time. They can be presented as follows:



The data can be analysis using clustering and represented as shown below:



The process followed by the hospital in the detection of frauds can be summarized as shown below:



Data mining can be used to detect prescription fraud as well. Sometimes, it is possible that patients and physicians commit fraud for different reasons, primary reason being monetary gain.

Data mining can help detect instances of fraud.

For instance, if the treatment for a non-serious disease is generally drug (X+Y) in 99% of cases, then a treatment of drug (X+Y+Z) may be flagged. Several such rules and standards can be developed using the knowledge stored in databases and building training sets that help build rules and standards. Once the rules and standards are built, each prescription that will occur in the future is passed through the monitoring system. The monitoring system then flags the prescriptions that are suspicious. These prescriptions are then further analyzed by an expert to verify if the prescription is fraudulent or not. Association rule mining can be used for generation of rules. Some samples of rules that can be developed are shown below:

IF A DOCTOR PRESCRIBES DRUG X FOR DISEASE A,

THEN HE WILL MOST LIKELY PRESCRIBE DRUG Y WITH A PROBABILITY OF 99%

IF A DOCTOR PRESCRIBES DRUG (X+Y) FOR DISEASE B,

THEN HE WILL MOST LIKELY PRESCRIBE DRUG Z WITH A PROBABILITY OF 98%

IF A DOCTOR PRESCRIBES DRUG (X+Y+Z) FOR DISEASE C

AND

IF PATIENT AGE> 60

THEN HE WILL MOST LIKELY PRESCRIBE DRUG I WITH A PROBABILITY OF 99%

IF A DOCTOR PRESCRIBES DRUG (X+I) FOR DISEASE D

AND

IF PATIENT AGE > 45

AND

IF BLOOD PRESSURE > 150/80

THEN HE WILL MOST LIKELY PRESCRIBE DRUG K WITH A PROBABILITY OF 98.9%

IF A DOCTOR PRESCRIBES DRUG (X+K) FOR DISEASE D

AND

IF PATIENT AGE>65

AND

IF BLOOD PRESSURE > 150/80

AND

IF HEART RATE > 75

AND

IF HYPERTENSION = YES

THEN HE WILL MOST LIKELY PRESCRIBE DRUG M WITH A PROBABILITY OF 98%

IF A DOCTOR PRESCRIBES DRUG (Y+Z+I) FOR DISEASE E

AND

IF PATIENT > 69

AND

IF BLOOD PRESSURE > 170/80

AND

IF HEART RATE > 77

AND

IF DIABETES = YES

THEN HE WILL MOST LIKELY PRESCRIBE DRUG X WITH A PROBABILITY OF 97%

IF A DOCTOR PRESCRIBES DRUG (X+I) FOR DISEASE E

AND

IF BLOOD PRESSURE < 130/60

AND

IF PATIENT < 30

AND

IF HEART RATE = 72

THEN HE WILL MOST LIKELY NOT PRESCRIBE ANY OTHER DRUG WITH A PROBABILITY OF 99%

IF A DOCTOR PRESCRIBES DRUG X FOR DISEASE A

AND

IF BLOOD PRESSURE = 110/70

AND

IF PATIENT < 30

AND

IF HEART RATE = 72

AND

IF WEIGHT < 70 kg

AND

IF DIABETES = NO

AND

IF MIGRANES = NO

AND

IF PARKINSON'S = NO

THEN HE WILL MOST LIKELY NOT PRESCRIBE ANY OTHER DRUG WITH A PROBABILITY OF 98%

IF A DOCTOR PRESCRIBES DRUG Y FOR DISEASE B

AND

IF BLOOD PRESSURE = 110/70

AND

IF PATIENT < 25

AND

IF HEART RATE = 72

AND

IF WEIGHT < 70 kg

AND

IF DIABETES = NO

AND

IF MIGRANES = NO

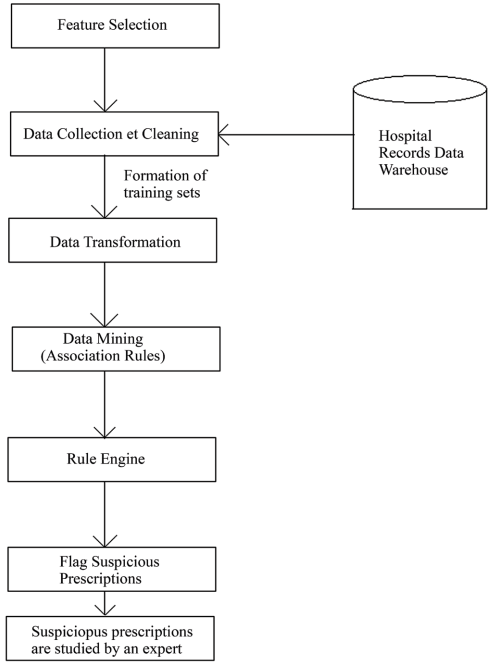
AND

IF PARKINSON'S = NO

THEN HE WILL MOST LIKELY NOT PRESCRIBE ANY OTHER DRUG WITH A PROBABILITY OF 99.5%

Several rules of this type can be generated and built into a rule engine that will constantly monitor all the prescriptions in a hospital to check for suspicious behavior and flag the behavior that seems suspicious and doesn't fit into the prescription standards that were previously defined.

The process is as shown below:



There exist several implementations of data mining in the medicine and healthcare sectors that have been extremely successful. One successful

application of data mining was implemented by the Utah Bureau of Medicaid Fraud that mined large volumes of data that was generated by a lot of prescriptions, treatment courses and operations that detected unusual and different patterns and uncovered fraud (Milley, 2000). Yet another method that was successfully implemented was done by the Texas Medicaid Fraud and Abuse Detection System, who were able to recover $2.2 million and were able to identify 1400 suspects that were investigated for fraud in 1998 after being in operation for a time period that was less than a year. This system also went on to win a national award for achievement for innovative use of technology with honors (Anonymous, 1999).

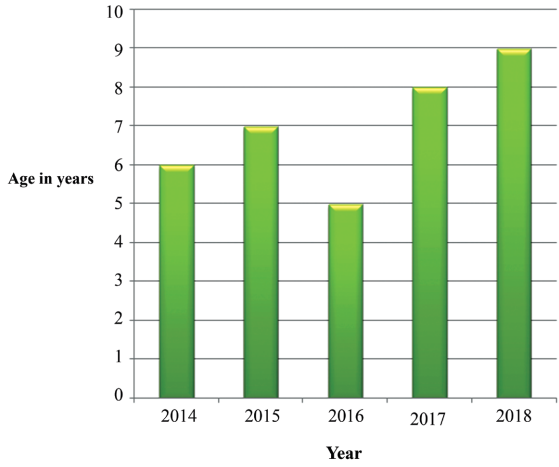
f. Medical Device Analysis: Medical advancements have led to rise in the number of medical devices. As more and more devices are being created for different diseases such as blood glucose machine for diabetes, pacemakers for heart problems, pedometers, etc. The quality of critical devices needs to be tested from so as to ensure the heath of the patients. No margin of error must be present in devices that work closely with the heart. Hence, the devices in the medical industry must be constantly checked for effectiveness.

Moreover, the medical community has seen a rise in wireless technology including wireless sensors that are new mobile healthcare options. The strength of these devices is always under question and needs to be check from time to time.

In order to check the life and validity of the machines different data mining techniques can be used. Different prediction algorithms can be used that successfully predict the life and quality of the machine and battery respectively.

As it has been already said before, the medical and healthcare industry has a dearth of data and the data accumulated over the last few years can be used to form trainings set, feed them to the data mining model and use predictive analysis methods to predict the life/quality of particular medical device.

For example consider the following data set that consists of the life of pacemakers developed by a particular company for over the last 5 years.



The above information compounded over decades, along with other parameters such as battery data, past data such as problems encountered in its use, etc. can be used to predict the life of the device.

With this information, the patient and the doctor can prepare and put into place steps for replacement of the existing pacemaker.

Many such applications of data mining are being used currently in the domain of medical devices.

g. Pharmaceutical Industry Applications: The pharmaceutical industry is always developing different drugs and products for a wide variety of diseases. This industry is always faced with a lot of completion as there are several companies in this sector that compete to sell medicines for the same illnesses. Pharmaceutical companies are now forced to provide drugs with minimum side-effects and low cost. Hence, these companies need to understand the medical data and predict what kind of drugs are most in need and what strategies can be put into place.

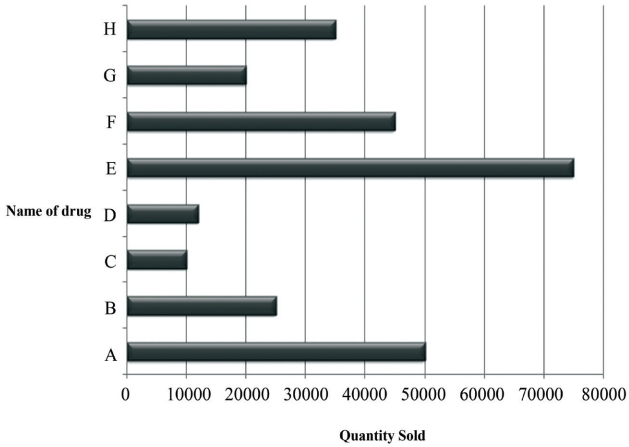
Pharmaceutical companies have a lot of data that is stored about their invoices, purchases, inventories, drug trails etc. All this data is stored in giant warehouses and not put to use. This sector is also 'rich' in data but 'poor' in the information it has. Hence more and more pharmaceutical companies and organizations are looking to Data mining to help them improve their sales and better manage their processes.

Data mining techniques such as clustering, naïve Bayes classification, regression analysis can be used to predict the risks/side effects of drugs, predict the most popular drug, predict the need for more production, help

classify drugs into different categories, predict the potential sales of certain drugs and so on.

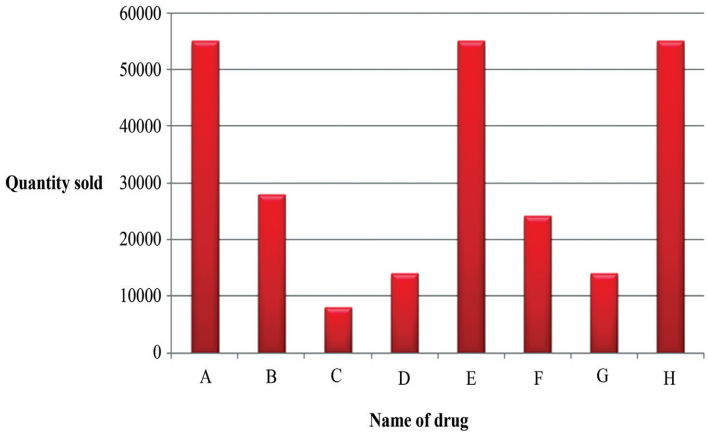
We shall now take a look at a few samples of application of data mining techniques on the data stored by pharmaceutical companies.

Consider the following data set showing the drugs that are most purchased over a year:



The above data can be used further to draw inferences and then predict the sales for the next year based on this data and other factors that has been accumulated over the years.

A sample predict with respect to the above sample data set is shown below:



The above prediction data can then be used to prepare the orders and plan for the future sale of these drugs as per the predictions so as to not have a scarcity of drugs or an overflow. Different managerial decisions can be made regarding the drugs as well.

For example, the drug C is most likely to not sell as well. The managers can decide to monitor sales of this drug and if it drops down to a particular threshold level, they can choose to recall the drug and introduce a new drug instead. Many such decisions can be taken owing to the predictive analysis capabilities of data mining.

Data mining can also be used to study existing medicines/drugs that are currently being used by patients all over and detect the most common side effects that are faced by patients.

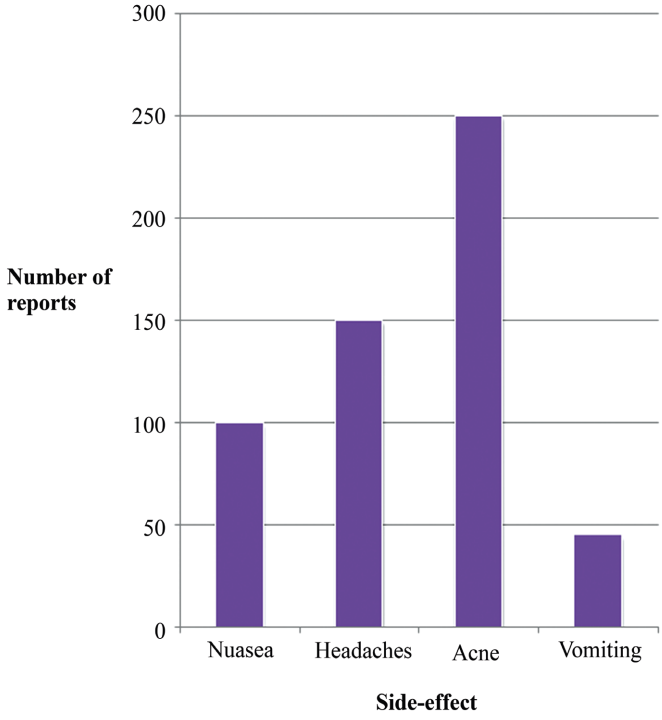
The reporting of potential side effects to the pharmaceutical companies can be done in several ways such as through online forums, doctors get this information from the patients during medical visits and many other sources online. The doctors report the symptoms back to the pharmaceutical companies and then pharmaceutical companies usually store this information in the database.

As the rules regarding the medicines are getting stricter, pharmaceutical companies need to ensure that the drugs they sell are not posing long-term problems to the patients taking it. They need to take smart business decisions regarding their drugs and keep monitoring the market.

Text mining of online forums combined with traditional data mining methods can be used to analyze the side effects of drugs and monitor them as well.

For example, a hybrid data mining approach can be applied on the data related to drugs and different analysis can be performed.

Consider the following data set that gathers the most common side effects that were reported by patients for a particular drug X for a period of 6 months:

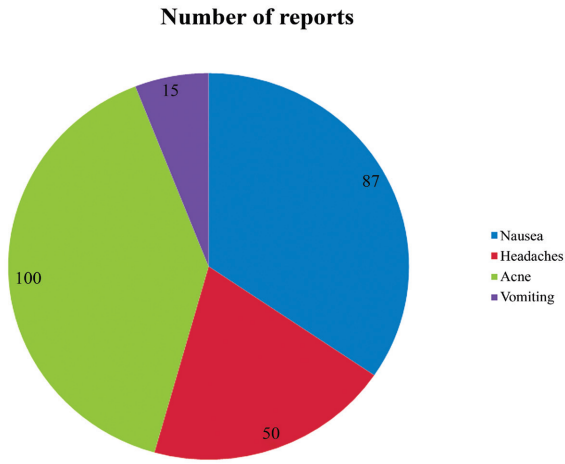


This data can further be used to make predictions regarding the possibility of rise in of potential side-effects or to gather data so as to further analyze the drug so that changes can be made to it to reduce the side-effects.

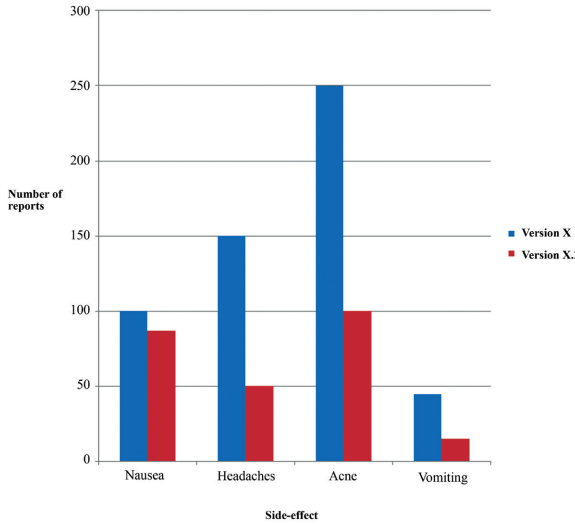
For instance, if the pharmaceutical company introduced another version for drug X called version X.2.

After the introduction and sale of this new version in the market for some time, a new round of data mining can be applied to the data related to this drug (the new version) and the number of reported cases can be monitored to see if the number of patients that report the symptoms have reduced or not.

For instance, consider the following data that shows the side effects post the release of version X.2 showing the reported side-effects for a period of 6 months:



A comparison of the can also be done and represented in a visual manner. This is shown below:



Additionally, the calculation of the percentage of improvement can be done as well.

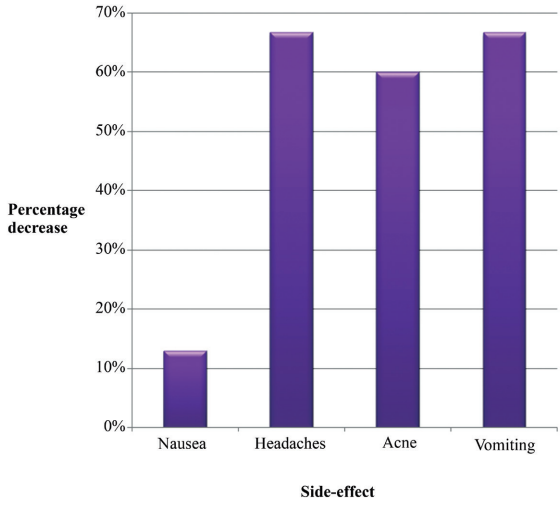
This is shown in the table below:

Symptom/Drug Version	Version X	Version X2	Percentage Decrease
Nausea	100	87	13%

Headaches	150	50	66.66%
Acne	250	100	60%
Vomiting	45	15	66.66%

Based on the above calculations, the pharmaceutical company can infer that they have successfully managed to reduce 3 out of the 4 reported side-effects with a success of over 60%.

This information can be represented visually as shown below:



Another possible application of data mining for pharmaceutical companies is the marketing of drugs. For instance, if the pharmaceutical companies are able to discover trends in the purchase data, they can propose different kinds of deals to the hospitals and medical institutes for bulk purchase of certain drugs that are bought in combination of one another.

Due to the dearth of data that is available in the pharmaceutical sector as well, their data can be mined so as to discover different patterns and rules regarding different drugs that are purchased by different organizations.

Different data mining techniques can be used to mine the data such as classification, clustering, decision trees, CART, association rule mining, etc.

For instance, consider the following sample rules that can be mined from the pharmaceutical data:

IF HOSPITAL A buys drug X

THEN THEY ALSO BUY DRUG Y WITH A PROBABILITY OF 99%

IF HOSPITAL A buys drugs X and Y
THEN THEY ALSO BUY DRUG Z WITH A PROBABILITY OF 98%

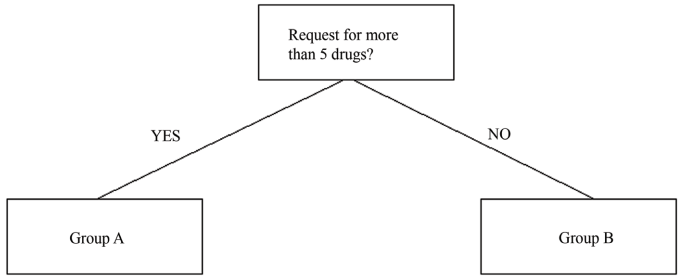
IF HOSPITAL B buys drugs X and Z
THEN THEY ALSO BUY DRUG Y WITH A PROBABILITY OF 98%

IF HOSPITAL B buys drug X, Y and Z
AND
IF HOSPITAL B previously bough drug Y
THEN
THEY WILL ALSO BUY DRUGS I and K WITH A PROBABILITY OF 99.9%

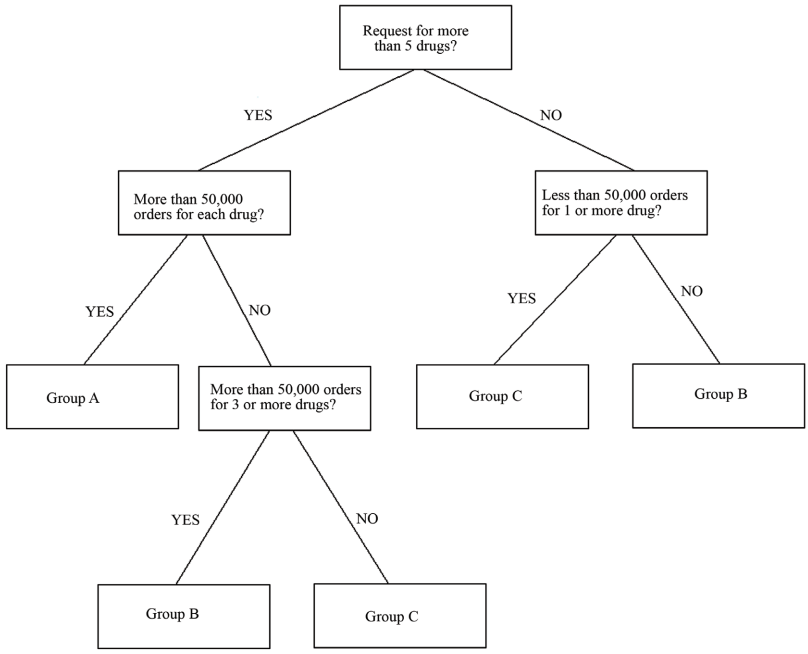
IF HOSPITAL C buys drug X, Y and I
AND
IF HOSPITAL C previously bough drug X and Y
THEN
AND IF EDPIDEMIC = TRUE
THEN
THEY WILL ALSO BUY DRUGS J, M and K WITH A PROBABILITY OF 99.8%

Many such rules can be extracted and developed from the data. These rules can further be developed so as to develop different strategies to sell the drugs that are often bought together. Different discounts could be offered for bulk orders of drugs depending on various factors such as the demand, the institute that is asking, the price and so on.

Such an example of a decision tree that classifies clients into two types is shown below:

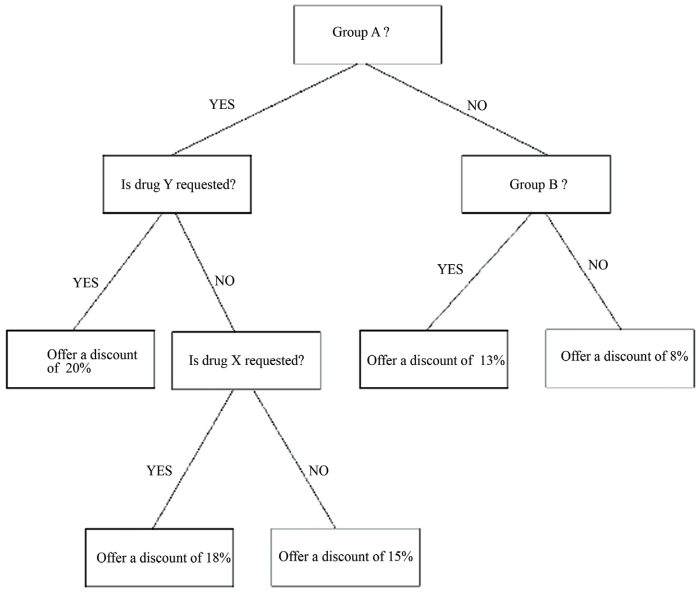


Further, this tree can be improved upon so as to include other factors as shown below:



Based on the groups formed for the request of drugs by different medical institutions, different discount plans can be proposed.

One such sample plan is shown below:



Several marketing decisions can be made based on information retrieved from application of data mining algorithms on pharmaceutical drug data.

h. Marketing: Marketing is a phenomenon that is common to almost all the sectors in the world. Each and every industry surely markets most of its products using different means and mediums of marketing. The medicine and/or the healthcare industry are not an exception to this phenomenon. On the contrary, this industry is one of the industries that do a lot of direct and indirect marketing to consumers all over the world.

Different mass campaigns such as door-to-door campaigns, advertisements, billboards, television, using the Web, social media are constantly being done by the medicinal domain in order to advertise and sell different drugs, promote hospitals, their facilities, infrastructure, staff, etc.

But, many hospitals and pharmaceutical companies tend to lose a lot of funds due to these types of campaigns. Additionally, research done on the usefulness of these campaigns suggested that these campaigns were not extremely effective.

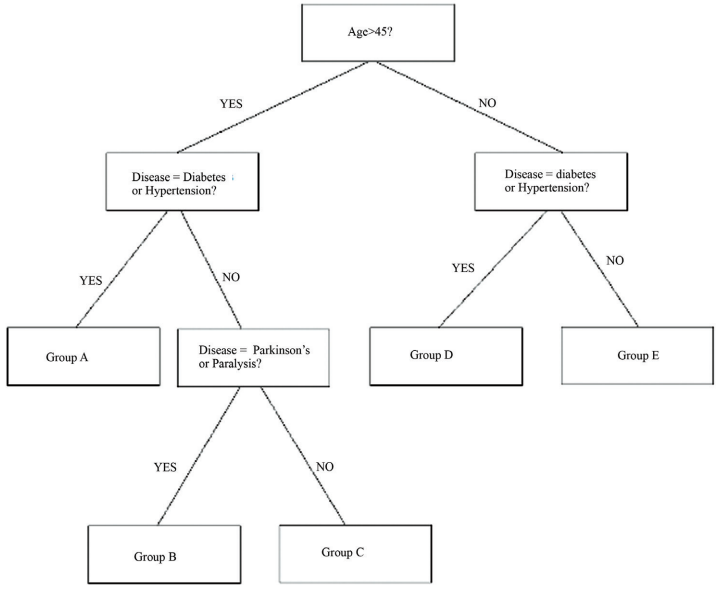
This is where data mining can play a part. Traditionally, the origin of data mining began with the need to do targeted marketing and this industry is in dire need of similar types of implementations. Data mining can be used effectively so as to target specific type of customers that have more chances of purchasing a product or a drug. The presence of advanced infrastructure

may place a key role in attracting new clients to hospitals as well. Similar to other commercial sectors, data mining can play a lead role in the healthcare domain to find out customer preferences, their patterns of usage, current trends and future trends as well. It can be used to assess and analyze current and future needs of customers so as to increase their level of satisfaction and contentment with the processes. The data mining applications are able to easily predict different healthcare products a customer is more likely to buy at a later date. Additionally, these mining techniques can also be used to mine rules for preventive care and predict whether such type of care is useful in the long term or not, etc.

Hospitals have now started applying data mining techniques by taking the health and demographic patient data and mining it for information so as to form targeted customer groups. There exist a plethora of applications of data mining in marketing and customer relationship management.

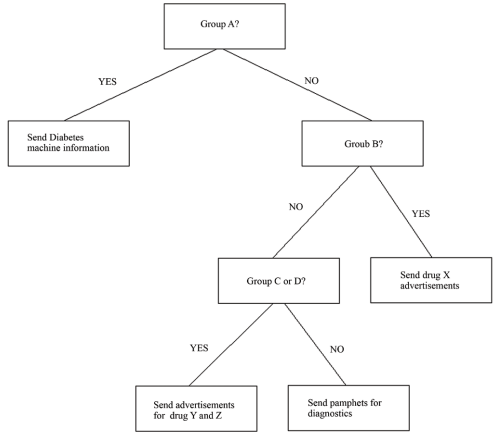
We shall examples of a few of the vast array of applications.

A sample that classifies patients based on their age, disease and sex is shown below:



Based on the groups formed above, the patients belonging to each group are target with particular mailings and messages.

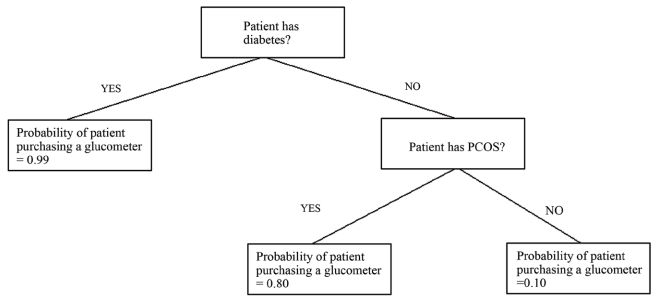
One such example is shown below:



The above decision tree shows that based on customer profiles and group classification of the customers, different types of marketing strategies are employed for customers of each group.

Data mining techniques such as decision trees can also be used to predict the probability of a customer purchasing a particular product.

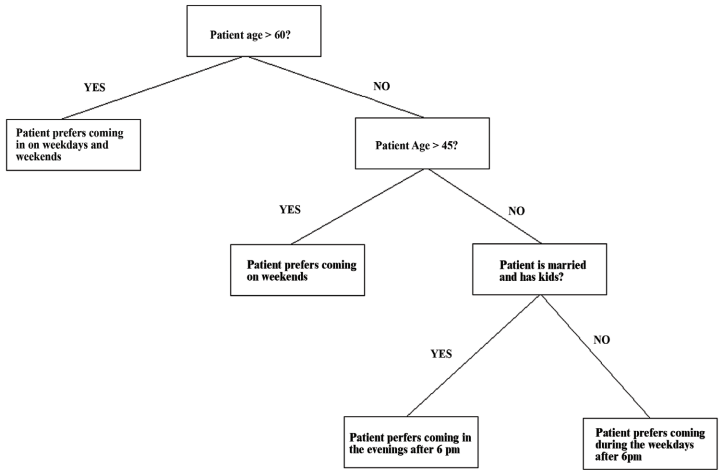
The following decision tree is used to predict whether a customer will most likely purchase a glucometer (machine to check blood sugar levels) or not.



Data mining can also be used to analyze specific customer behaviors. Different patterns in their usage can me mined from their profile and put to use for providing the customers with better services. For instance, if data mining uncovers patterns regarding the time period when customers more likely prefer picking up medication, a reminder service can be generated.

Additionally, based on the customer preferences, different decisions can be taken.

For instance, if we imagine the total number of patients of a particular doctor, different kinds of information regarding the appointment times of the patients can be uncovered. Based on the information extracted, different inferences can be made and then later used.
Consider the following decision tree that describes different rules that can be formed based on patient appointment times:



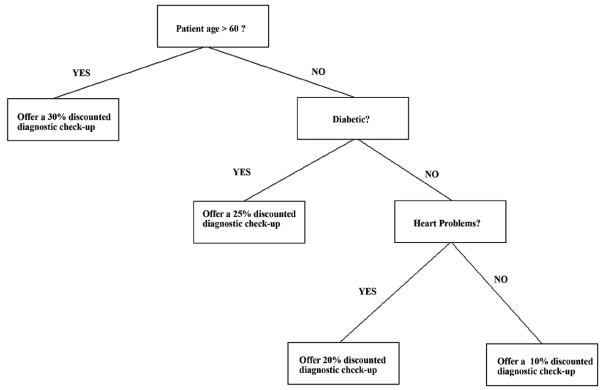
As we can see, different patients have different preferences of appointment times. This mined data can be used further to suggest different time slots to different patients based on their preferences.

Hence, in this way doctors can make sure a patient will most likely keep his appointment and the patient will be satisfied with the hospital for providing him a suitable appointment time based on his preferences.

Additionally, different patient information can be used to promote different products that can help the patient in the long run.

For example, if a person is a long-term diabetic patient, the may benefit from a diagnostic check for the heart or from different vitamin supplements. Based on the patient's situation different products can be marketed to the patient by the doctor and by the hospital as well.

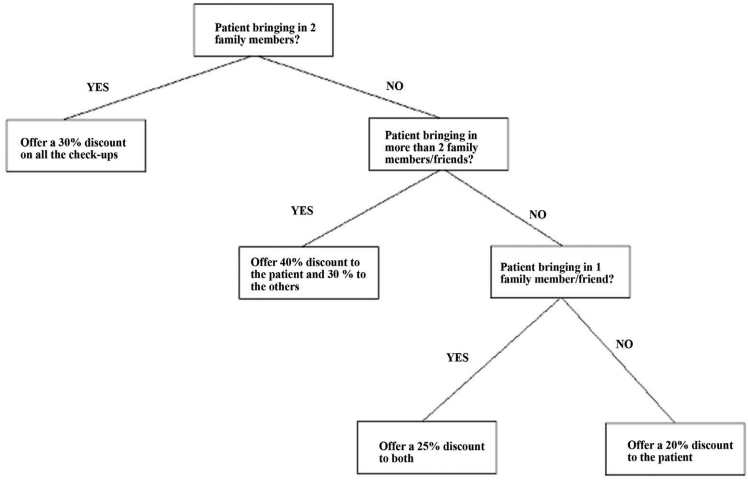
A sample is shown below:



As shown above, based on the patient information different offers are made to the patients so as help patients and to increase profits of the diagnostic department of the hospital as well. This can be done so as to promote the quality of the hospital.

For instance, if a patient also brings along a family member for a discounted family check-up he is given an additional discount. By doing this, the patient is benefited by a discount, so is his family and the hospital has a new customer that has been exposed to the hospital's services. If the patient is happy with the diagnostic service, he is more likely to get treatment from the same hospital in case he ever needs medical attention. This helps the hospital invest and market to potential new clients and also help the clients get access to cheaper diagnostics.

An example of a promotional plan is shown below:

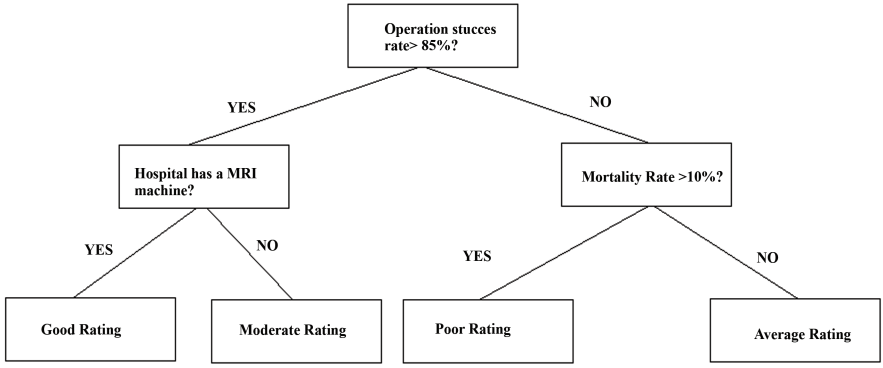


The above classification offers different discount plans based on the number of members a particular patient brings in. This helps encourage the patient to do a diagnostic check-up to check his health and also introduces the hospital to potential new clients.

Another application of data mining is in the ranking of hospitals. Hospitals world-over are ranked for their surgery team, quality, infrastructure, facilities, etc. Different titles are awarded to hospitals based on these input parameters. Data mining is capable of analyzing these different parameters and providing detailed information regarding the hospital that fared best.

Hospitals are ranked based on their quality, service, infrastructure, mortality rate and operation success rate. Based on such features, hospitals can be classified into different groups.

One such classification is shown below:



Many such classifications can be done in order to discover which service is the best in which hospital and which hospital performs the best on a global scale.

Data mining has successfully been implemented in the medicine industry for different purposes. There exist several tools in the market that are being used to predict the levels of accuracy for different diseases.

Many tools that have been developed and that are used for predictive purposes are developed on the basis of the volumes of data recorded in the medical sector. There are tools that are built to predict the success rates or accuracy rates/level of different diseases in this industry.

Various studies have been conducted to this effect. One such comparative study performed by on the following list of diseases is shown below (Durairaj & Ranjani, 2013):

- Heart Disease
- Cancer
- HIV/AIDS
- Blood
- Brain Cancer
- Tuberculosis
- Diabetes Mellitus
- Kidney dialysis
- Dengue
- IVF
- Hepatitis C

S.No	Type of disease	Data mining tool	Technique	Algorithm	Traditional Method	Accuracy level(%) from DM application
1	Heart Disease	ODND, NCC2	Classification	Naive	Probability	60
2	Cancer	WEKA	Classification	Rules. Decision Table		97.77
3	HIV/AIDS	WEKA 3.6	Classification, Association Rule Mining	J48	Statistics	81.8
4	Blood Bank Sector	WEKA	Classification	J48		89.9
5	Brain Cancer	K-means Clustering	Clustering	MAFIA		85
6	Tuberculosis	WEKA	Naive Bayes Classifier	KNN	Probability, Statistics	78
7	Diabetes Mellitus	ANN	Classification	C4.5 algorithm	Neural Network	82.6
8	Kidney dialysis	RST	Classification	Decision Making	Statistics	75.97
9	Dengue	SPSS Modeler		C5.0	Statistics	80
10	IVF	ANN, RST	Classification			91
11	Hepatitis C	SNP	Information Gain	Decision rule		73.20

(Durairaj & Ranjani, 2013).

INDEX