ADDIS ABABA UNIVERSITY

COLLEGE OF NATURAL SCIENCES

SCHOOL OF INFORMATION SCIENCE

PREDICTING THE MARKET STATUS OF COFFEE, PEA BEANS AND SESAME: THE CASE OF ETHIOPIA COMMODITY EXCHANGE

BY

TEWABE WORKU ADMAS

October, 2015

ADDIS ABABA UNIVERSITY
SCHOOL OF GRADUATE STUDIES
COLLEGE OF NATURAL SCIENCES
SCHOOL OF INFORMATION SCIENCE

PREDICTING THE MARKET STATUS OF COFFEE, PEA BEANS AND SESAME: THE CASE OF ETHIOPIA COMMODITY EXCHANGE

BY

TEWABE WORKU ADMAS

A Thesis Submitted to the School of Information Science,
College of Natural Sciences, Addis Ababa University;
In Partial Fulfillment of the Requirements of the
Degree of Master of Science in Information Science

Advisor: Solomon Teferra (PhD)

October, 2015

Addis Ababa, Ethiopia

ADDIS ABABA UNIVERSITY
COLLEGE OF NATURAL SCIENCES
SCHOOL OF INFORMATION SCIENCE

PREDICTING THE MARKET STATUS OF COFFEE, PEA BEANS AND SESAME: THE CASE OF ETHIOPIA COMMODITY EXCHANGE

BY

TEWABE WORKU ADMAS

October, 2015

A Thesis Submitted to the School of Information Science,
College of Natural Sciences, Addis Ababa University;
In Partial Fulfillment of the Requirements of the
Degree of Master of Science in Information Science

Name and Signature of Members of the Examining Board

Name	Title	Date	Signature
Million Meshesha (PhD)	Examiner	_____	_____
Martha Yifiru (PhD)	Examiner	_____	_____
Solomon Teferra (PhD)	Advisor	_____	_____

DECLARATION

I declare that this thesis is my original work and has not been submitted as a partial fulfillment of the requirement for a degree in any university.

Tewabe Worku Admas

Signature: _____

Date: _____

This thesis has been submitted for examination with my approval as a university advisor.

Solomon Teferra (PhD)

Signature: _____

Date: _____

DEDICATION

I would like to dedicate this thesis to my parents and families for all their love and support.

ACKNOWLEDGEMENTS

This thesis has been comprehended with the assistance and guidance of keen individuals and organizations. Hence I am so glad when I express my gratitude to those who have played significant role in my graduate career.

First, I would like to express my deepest gratitude to my Advisor Dr. Solomon Teferra, who has been an excellent advisor and mentor. Besides my advisor, I would like to thank a lot Dr. Millon Meshesha & Dr. Martha Yifiru; their feedbacks during our meetings offered a critical contribution to my work and really helped to increase the quality of my work.

Next, I would like to express my heartfelt gratitude to my respected parents, Ato Worku Admas and W/ro Yeshitila Ambaye, for their unreserved all rounded support and love. Also my beloved wife, Fasika Lemma, deserves lots of credits for the care she gave me while I was in this sometimes stressful period. Moreover, the good times I have and encouragements from all my families, friends and classmates are also great.

Finally, I am grateful to the organization (Ethiopia Commodity Exchange /ECX/) market data experts, especially to Ato Hailemichael Hailu and Ato Ermias Assefa who offered me all the necessary information, data and documents I needed for my thesis paper.

TABLE OF CONTENTS

LISTS OF FIGURES

LISTS OF TABLES

ACRONYMS AND ABBREVIATIONS

ARFF	Attribute-Relation File Format
ARM	Association Rules Mining
BPNN	Back Propagation Neural Network
CART	Classification and Regression Trees
CRM	Customer Relation Management
CRISP-DM	CRoss-Industry Standard Process -Data Mining
CSV	Comma Separated Values
DT	Decision Trees
ECX	Ethiopia Commodity Exchange
EMD	ECX Market Data
GA	Genetic Algorithm
GUI	Graphical User Interface
OLAP	On-Line Analytical Processing
KDD	Knowledge Discovery in Database
KDP	Knowledge Discovery Process
KNN	K-Nearest Neighbor
MAP	Moving Average Price
ML	Machine Learning
MLP	Multi-Layer Perceptron
MLR	Multi-Linear Regression
MVRVM	Multivariate Relevance Vector Machine
PDM	Predictive Data Mining
SPSS	Statistical Packages for Social Sciences
SVM	Support Vector Machine
WAP	Weighted Average Price
WEKA	Waikato Environment for Knowledge Analysis

ABSTRACT

Many commodity exchange organizations; including the Ethiopia Commodity Exchange (ECX) have been trying to find a way to predict future market status (in terms of price and sales volume) for commodities that are being traded within the exchange's trading platform. Data mining helps to find predictive information from large databases and data warehouses. Companies use predictive modeling tools for strategic decision-makings and by analyzing the company's historical information we can anticipate the changes in the future. Thus, in this study, by analyzing the historical ECX market data (EMD) obtained from ECX; we have discussed on different data mining methods which are helpful in building a predictive data mining model.

The Hybrid Knowledge Discovery Process for Data Mining is followed to build the predictive model that analyzes and predicts the price and sales volume. This methodology was developed, by adopting the CRoss-Industry Standard Processes for Data Mining (CRISP-DM) model to the needs of academic research community. Our work is based on finding suitable data sets as well as best predictive model that helps in achieving high accuracy and generality for Weighted Average Price (WAP) in Ethiopian Birr and Sales Volume in Tons predictions for selected agricultural commodities (Coffee, Pea Beans and Sesame) traded at ECX. For solving this problem, different data mining classification techniques (Eight WEKA classifiers: Gaussian Processes, Linear Regression, Multilayer Percepton, SMOreg, Decision Table, M5Rules, M5P and REPTree) were evaluated on different data sets.

The experimental results obtained from this study shows; the Decision Table Classifier has the highest optimal accuracy score with maximum optimal Correlation Coefficient Percentage (CCP) of 97.8%, minimum optimal Root Mean Square Percentage Error (RMSPE) of 3.9% and an optimal Time of 0.02 seconds to build the Price Predictive model. While, the M5Rules Classifier has the highest optimal accuracy score with maximum optimal CCP (80.5%), minimum optimal RMSPE (9.4%) and an optimal Time (0.11 seconds) to build the Sales Volume Predictive model.

Finally, by extending WEKA software source code, an application (predictive-model-prototype) which is termed as "ECX Price and Sales Volume Predictive System" with a user-friendly GUI is developed and deployed for the usage of domain experts (end users). Therefore, the results obtained from this research indicate that data mining classification models are very useful in predicting price and sales volume of commodities traded in commodity exchange organizations like ECX for the effective and efficient utilization of massive amount of market data to support experts and stockholders (traders) in making strategic planning as well as proactive and knowledge-driven decisions.

CHAPTER ONE

INTRODUCTION

1.1 Background

Financial market is a chaotic, complex, non-stationary, noisy, nonlinear and dynamic system but it does not follow random walk process. The commodity exchange trading is essentially a non-linear, non-parametric system that is extremely hard to model with any reasonable accuracy [34, 36]. Many commodity exchange organizations; including the Ethiopia Commodity Exchange (ECX) have been trying to find a way to predict future market status for commodities that are being traded within the Exchange's trading platform and forecast the commodities future trade value (price) and volume (sales volume) performance, which in turn supports the exchange's strategic planning and decision making activities. To achieve those objectives, some research used the techniques of fundamental analysis, where trading rules are developed based on the information associated with macroeconomics, productions, traders, and trading platform. As noted by the authors of [34, 35 and 36]; the fundamental analysis assumes that the price and volume of a commodity depends on its intrinsic value and expected return on investment. Analyzing the exchange's operations and the market trend in which the exchange is operating can do this. Consequently, the future market status can be predicted reasonably well. Most people believe that fundamental analysis is a good method only on a long-term basis. However, for short and medium term speculations, fundamental analysis is generally not suitable.

Some other researchers used the techniques of technical analysis, in which trading (commodity exchange) rules were developed based on the historical data of commodity trading price and volume.

Technical analysis refers to the various methods that aim to predict future market status using past commodity prices and sales volume information. It is based on the assumption that history repeats itself and that future market directions can be determined by examining historical price and sales volume data. Thus, it is assumed that price and sales volume trends (patterns) exist that can be identified and utilized for profit. The traditional method of turning data into knowledge relies on manual analysis and interpretation to evaluate and analyze data stored in large databases, new techniques and methods are needed to search large quantities of data, to discover new patterns and relationships hidden in the data [34].

Recently, data mining and machine learning techniques have been applied to this area [5]. Data mining refers to extracting or mining knowledge from large data stores or sets. Following the assumption of technical analysis that patterns exist in market data, it is possible in principle to use data mining techniques to discover these patterns in an automated manner. Once these patterns have been discovered, future market status (in terms of price and sales volume) can be predicted.

Data mining is one among the most important steps in the knowledge discovery process. It can be considered the heart of the KDD process [6]. This is the area, which deals with the application of intelligent algorithms to get useful patterns from the data. It is a new generation of computerized methods for extracting previously unknown, valid, and actionable information from large volume of database and then using this information to make critical decision. Hence, it is better to explore the relevance of this technology to any organizations (like ECX) that perform market analysis activity.

There are various techniques and tools of data mining that help researcher in exploring the benefit of data mining in various sectors. Data mining techniques can be broadly classified based on what they can do [4], namely description and visualization; association and clustering; as well as classification and estimation. Classification and estimation are also known as predictive modeling. Predictive modeling can be used to identify patterns, which can then be used to predict the odds of a particular outcome based upon the observed data.

1.2 Statement of the Problem

Today, predicting commodities future market status (predicting price and/or sales volume) has been one of the biggest challenges for most commodity exchange organizations including ECX. The grand challenge of using a database/data warehouse is to generate useful rules from raw data in a database/data warehouse for users to make decisions, and these rules may be hidden deeply in the raw data of the database/data warehouse. Traditionally, the method of turning data into knowledge relies on manual analysis; this is becoming impractical in many domains as data volumes grow exponentially. The problem with predicting future market status is that the volume of data is too large, huge and complex [21].

Even though there is a data warehousing system at ECX, the application of data mining is not yet practiced. So, there is a problem to predict the future market status/trend (predicting price and/or sales volume) of the commodities traded at ECX, which is very helpful for the organization to make

a strategic plan and decisions to maximize the profit of ECX. Thus, this study defines the problems (challenges) on predicting future market status and proposes solutions by consulting domain experts. Moreover, to address this problem, the concepts of data mining tasks, techniques and applications on the available ECX Market Data (EMD) obtained from the ECX data warehouse will be introduced and implemented.

To the knowledge of the researcher, no attempt has been made using data mining classification technique to predict the price and sales volume of agricultural commodities (Coffee, Pea Beans and Sesame) at the ECX. Moreover, this work has been the first attempt to evaluate the applicability of DM for predicting both the price and sales volume of agricultural commodities (Coffee, Pea Beans and Sesame) at the ECX. Last not least, by extending WEKA software source code, an application (predictive-model-prototype) with a user-friendly GUI have been developed and deployed for the usage of domain experts (end users); which was not used in most of the previous studies.

Therefore, the purpose of this study is to investigate the underlying attributes of price and sales volume from the available historical ECX Market Data (EMD) data using data mining techniques and develop a predictive DM model. To this end, it has been attempted to obtain answer for the following research questions:

- What are the main determinate attributes of EMD records that are used in building Price and Sales Volume Predictive Model
- What are the most appropriate DM techniques (by comparing different DM classification algorithms using different EMD datasets) for building the price and sales volume predictive model
- To what degree (accuracy level) the predicted Price and Sales Volume Predictive Model be determined by applying data mining
- To what extent can we customize DM tools (such as WEKA) for developing a Predictive Model Prototype/Application with a user-friendly GUI for the usage of domain experts (end users)

1.3 Objective of the Research

1.3.1 General Objective

The main objective of this study is to analyze the historical market data (EMD) available on the Ethiopia Commodity Exchange organization using data mining methods and build a predictive data mining model which enable to predict the price and sales volume for the selected commodities (Coffee, Pea Beans and Sesame) traded at the exchange.

1.3.2 Specific Objectives

To achieve the general objective, the following specific objectives are aimed to be achieved in this research:

- To conduct a thorough review of literature that can support the study in the area of applying data mining technology on predicting price and sales volume of any commodity in general and commodities at the Ethiopia Commodity Exchange in particular.
- To select, collect and generate good quality market data that can be used for data mining task.
- To select suitable data mining tool, technique and algorithms for building predictive model from obtained historical ECX market data.
- To evaluate various data mining models and select the best model that is more appropriate to the problem domain.
- To construct an interactive & integrated predictive model application/prototype (i.e. ECX Price and Sales Volume Predictive System) with a user friendly Graphical User Interfaces (GUI).

1.4 Scope and Limitation of the Study

As noted by [34], analyzing commodity exchange's market data of all the commodities over several years may involve a few hundreds or thousands of records, but these must be selected from millions.

The main aim of this research is to apply data mining for determining and predicting Price and Sales Volume traded in the ECX. Thus, the scope of the data/dataset used in this study is to build the future market status (price and sales volume) predictive model will be on specific selected commodities (Coffee, Pea Beans and Sesame) obtained from the historical market data of the Ethiopia Commodity Exchange over some specified time durations. Moreover, due to difficulties in

obtaining other external market data sources, the international market (Reference Market) records are not considered in this study.

1.5 Significance of the Study

With the increase of economic globalization and evolution of information technology, commodity exchange's financial market data are being generated and accumulated at an unprecedented pace. As a result, there has been a critical need for automated approaches to effective and efficient utilization of massive amount of market data to support companies and individuals in strategic planning and investment decision makings. Data mining techniques have been used to uncover hidden patterns and predict future trends and behaviors in financial markets. The competitive advantages achieved by data mining include increased revenue, reduced cost, and much improved market place responsiveness and awareness [37].

Thus, the significance and applicability of this research (Predicting Price and Sales Volume by building a Predictive Data Mining Model) is very high for a commodity exchange organizations like ECX that will have a massive impact on individual, organizational and national economy growth. Utilizing this model helps the exchange to build a strategic plans and decisions in order to achieve a better commodity exchange trading platform with an optimal profit easily.

Moreover, in our setting where the problems are intertwined and there is an ever increasing demand of most commodity exchange (in particular) and financial (in general) organizations. Thus, as an academic research, this study could also invite interested researchers to explore more in related and similar areas.

1.6 Ethical Considerations

Data mining and data warehousing raise ethical and legal issues, combining information via data warehousing/data mining could violate privacy act, thus must tell people/organization how their information will be used when the data is obtained.

The study is conducted after getting approval from the Addis Ababa University-School of Information Science and accepted by the Ethiopia Commodity Exchange-Market Data unit. The confidentiality of any information and data obtained from the client organization is maintained

confidentially. That is, there is no information and data transferred to any third party and using the information and data for any other purpose/s beyond to this research work.

1.7 Organization of the Thesis

This thesis paper is organized into six chapters. The *First Chapter* introduces by giving an overview of the research: briefly discusses background to the problem area and states the problem, the general and specific objectives of the study as well as the scope, limitation, significance and ethical considerations of the research.

The *Second Chapter* gives a literature review about the concepts pertaining to the data mining technology and its application in the problem are reviewed and some related works in that subject is shown in this section. This chapter also discusses on various predictive data mining techniques used in order to try to build a predictive data mining model in general, as well as specific to predicting future market/trade status of various commodities at financial organizations. Moreover, the previous researches in which a researchable gap was left and become the concern of this research, is reviewed in this chapter.

Chapter Three is all about the Research Design and Methodology followed in this study; by discussing. the Methods of Problem Domain Understanding, Methods of Data Understanding & Data Preparation, Method of Data Mining (Predictive Model Building), Methods of Evaluating of the Discovered Knowledge and Methods of Using the Discovered Knowledge.

Chapter Four talks about the exploration and procedures performed for the preparations of the dataset/s (Data Exploration and Preprocessing) detail that is consumed in building the proposed "ECX Price and Sales Volume Predictive Model".

Then, *Chapter Five* presents the experimentation and result analysis phase of the study at hand. Experimental findings & results of the proposed Price and Sales Volume Predictive Model will be also discussed here.

Finally, a conclusion of the research and recommendations for future works about the topic is depicted in *Chapter Six*.

CHAPTER TWO

LITERATURE REVIEW

This chapter discusses the literature reviews conducted by refereeing of books, journals, articles, conference paper and the internet get more insight to the concept of data mining and its application, especially in related areas (i.e. Price and/or Sale Volume Predictive Models).

2.1. Overview of Data Mining

Data, Information, and Knowledge

Data: are any facts, numbers, or text that can be processed by a computer. Today, organizations are accumulating vast and growing amounts of data in different formats and different databases. This includes: Operational or transactional data (such as, sales, cost, inventory, payroll, and accounting), Non-Operational data (such as industry sales, forecast data, and macro-economic data) and Meta data (a data about the data itself, such as logical database design or data dictionary definitions) [3].

Information: The patterns, associations, or relationships among all this *data* can provide *information*. For example, analysis of retail point of sale transaction data can yield information on which products are selling and when [3].

Knowledge: Information can be converted into *knowledge* about historical patterns and future trends. For example, summary information on retail supermarket sales can be analyzed in light of promotional efforts to provide knowledge of consumer buying behavior. Thus, a manufacturer or retailer could determine which items are most susceptible to promotional efforts [3].

Data Warehouse: Dramatic advances in data capture, processing power, data transmission, and storage capabilities are enabling organizations to integrate their various databases into *data warehouses*. Data warehousing is defined as a process of centralized data management and retrieval. Data warehousing, like data mining, is a relatively new term although the concept itself has been around for years. Data warehousing represents an ideal vision of maintaining a central repository of all organizational data. Centralization of data is needed to maximize user access and analysis. Now days, an intense technological advances in data analysis software are allowing users to access this data easily. The data analysis software is what supports data mining [3].

What is Data Mining?

Generally, Data Mining (DM) (sometimes called Data or Knowledge Discovery) is the process of analyzing data from different perspectives and summarizing it into useful information - information that can be used to increase revenue, cuts costs, or both. Data mining (DM) and knowledge discovery are intelligent tools that help to accumulate and process data and make use of it. DM bridges many technical areas, including databases, statistics, machine learning, and human-computer interaction. The set of DM processes used to extract and verify patterns in data is the core of the knowledge discovery process These processes include data cleaning, feature transformation, algorithm and parameter selection, and evaluation, interpretation and validation [2].

Data mining software is one of a number of analytical tools for analyzing data. It allows users to analyze data from many different dimensions or angles, categorize it, and summarize the relationships identified. Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational databases.

Advantages of Using Data Warehouse as Part of Data Mining

Data warehousing is a process by which an enterprise collects data from the whole enterprise to build a single version of the truth. This information is useful for decision makers and may also be used for data mining. As stated by many authors, including [7, 23, 27], data warehouse can be of real help in data mining since data cleaning and other problems of collecting data would have already been overcome. Otherwise this task can be very resource intensive; perhaps more than 50% of effort in a data mining project is spent on this step. Figure 2.1, describes the main procedures in using Data Warehouse as part of Data Mining [59].

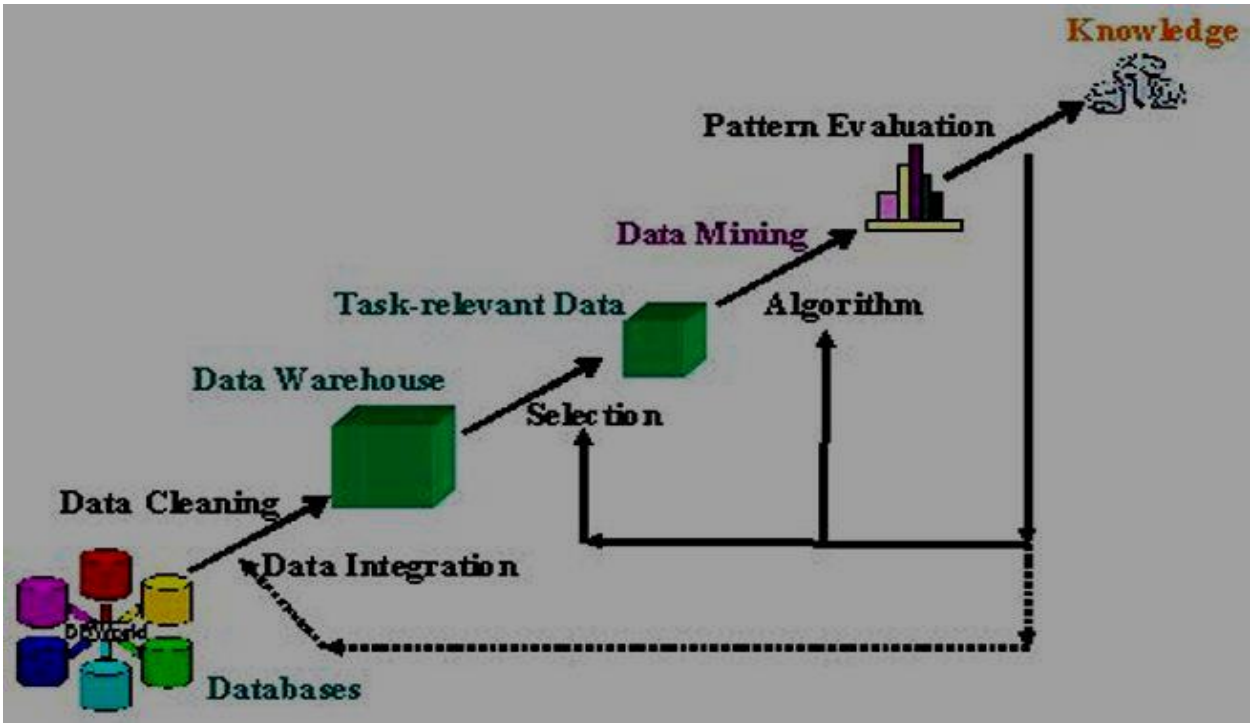

Figure 2.1 Using Data Warehouse as part of Data Mining

2.2. Data Mining Main Tasks and Applications

2.2.1. Data Mining Main Tasks

As many data mining researches depicts, the goal of any data mining effort can be divided in one of the following two types [5, 6, 7, 8]:

- Using data mining to generate ***descriptive models*** to solve problems.
- Using data mining to generate ***predictive models*** to solve problems.

The descriptive data mining tasks characterize the general properties of the data in the database, while predictive data mining tasks perform inference of the current data in order to make prediction. Descriptive data mining focuses on finding patterns describing the data that can be interpreted by humans, and produces new, nontrivial information based on the available data set. Predictive data mining involves using some variables or fields in the data set to predict unknown or future values of other variables of interest, and produces the model of the system described by the given data set.

The goal of predictive data mining is to produce a model that can be used to perform tasks such as classification, prediction or estimation, while the goal of descriptive data mining is to gain an understanding of the analyzed system by uncovering patterns and relationships in large data sets.

The goal of a descriptive data mining model is therefore to discover patterns in the data and to understand the relationships between attributes represented by the data, while the goal of a predictive data mining model is to predict the future outcomes based on passed records with known answers.

Furthermore, we can divide the data mining task of generating models into the following two approaches [5, 8]:

- ***Supervised*** *or* ***directed*** data mining modeling.
- ***Unsupervised*** *or* ***undirected*** data mining modeling.

The goal in supervised or directed data mining is to use the available data to build a model that describes one particular variable of interest in terms of the rest of the available data. The task is to explain the values of some particular field. The user selects the target field and directs the computer to determine how to estimate, classify or predict its value.

In unsupervised or undirected data mining however variable is singled out as the target. The goals of predictive and descriptive data mining are achieved by using specific data mining techniques that fall within certain primary data mining tasks. The goal is rather to establish some relationship among all the variables in the data. The user asks the computer to identify patterns in the data that may be significant. Undirected modeling is used to explain those patters and relationships one they have been found.

Thus, several data mining problem types or analysis tasks are typically encountered during a data mining project. Depending on the desired outcome, several data analysis techniques with different goals may be applied successively to achieve a desired result. The major data analysis tasks and techniques used in data mining are discussed as follows.

2.2.1.1. Classification

Classification assumes that a set of objects characterized by some attributes or features belong to different classes. The class label is a discrete qualitative identifier; for example, large, medium, or small. The objective is to build classification models that assign the correct class to previously unseen and unlabeled objects. Classification models are mostly used for predictive modeling. Discriminant analysis, decision tree, rule induction methods, and genetic algorithms generally apply [1].

Classification involves the discovery of a predictive learning function that classifies a data item into

one of several predefines classes. It involves examining the features of a newly presented object and assigning to it a predefined class. Classification has a two-step process. First a model is built describing a predetermined set of data classes or concepts and secondly, the model is used for classification [7].

2.2.1.2. Prediction

Prediction is very similar to classification. The difference is that in prediction, the class is not a qualitative discrete attribute but a continuous one. The goal of prediction is to find the numerical value of the target attribute for unseen objects; this problem type is also known as regression, and if the prediction deals with time series data, then it is often called forecasting. Regression analysis, decision trees, and neural nets generally apply [1].

Prediction can be viewed as the construction and use of a model to assess the class of a unlabeled sample, or to assess the value or value range of an attribute that a given sample is likely to have. According to, any of the techniques used for classification can be adapted for use in prediction by using training examples where the value of the variable to be predicted is already known, along with historical data for those examples.
Typical business related questions that can be answered using classification or prediction tasks are [7]:

- Which customers will buy?
- Which products will customer buy?
- How much will customer buy?

Prediction tasks in finance typically are posed in one of two forms: (1) straight prediction of the market numeric characteristic, e.g., stock return or exchange rate, and (2) the prediction whether the market characteristic will increase or decrease. Having in mind that we need to take into account the trading cost and significance of the trading return in the second case we need to forecast whether the market characteristic will increase or decrease no less than some threshold. Thus, the difference between data mining methods for (1) or (2) can be less obvious, because (2) may require some kind of numeric forecast.

Examples of the use of data mining in financial applications (mainly related to trading) are Stock Market Returns and Foreign Currency Exchange Rates including: *portfolio management, trading futures* and *Interpreting Trading Rules* [19, 20, 21, 22].

2.2.1.3. Estimation

While classification deals with discrete outcomes such as yes or no, debit card, home loan or vehicle financing, estimation deals with continuously valued outcomes. If some input data is available, estimation can be used to come up with some unknown continuous variable such as income or height. In estimation, one wants to come up with a plausible value or a range of plausible values for the unknown parameters of a system. Classification and estimation are often used together, as when data mining is used or predict who is likely to respond to a credit card balance transfer offer and also to estimate the size of the balance to be transferred. Estimation is grouped under predictive data mining tasks [7].

Typical examples of estimation and the business related questions that can be addressed by making use it include the following [7]:

- How many children are in a family?
- Estimating a family's total household income.
- Estimating the value of a piece of property.

2.2.1.4. Segmentation

Segmentation separates the data into interesting and meaningful sub-groups or classes. In this case, the analyst can hypothesize certain subgroups as relevant for the business question based on prior knowledge or based on the outcome of data description and summarization. Automatic clustering techniques can detect previously unsuspected and hidden structures in data that allow segmentation. Clustering techniques, visualization and neural nets generally apply.

Segmentation simply means making different offers to different markets segments; groups of people defined by some combination of demographic variables such as age, gender or income. Define the segmentation as a form of analysis used to for instance break down the visitors to a website into unique groups with individual behaviors. The grouping can then be used to make statistical projections, such as the potential amount of purchases they are likely to make. Segmentation is grouped under descriptive data mining tasks.

Typical business questions that can be answered using segmentation are [7]:

- What are the different types of visitors attracted to our website?
- In which age groups do the listeners of a certain radio station fall into?

2.2.1.5. Clustering

Clustering is the task of segmenting a diverse group into a number of similar subgroups or clusters.

Clusters of objects are formed so that objects within a cluster have high similarity in comparison to one another, but are very dissimilar to objects in other clusters. Clustering is commonly used to search for unique groupings within a data set. The distinguishing factor between clustering and classification is that in clustering there are no predefined classes and no examples. The objects are grouped together based on self-similarity. Clustering is grouped under descriptive data mining tasks. Typical business question that can be answered using clustering are [7]:

- What are the groupings hidden in your data.
- Which customer should be grouped together for target marketing purposes?

2.2.1.6. Summarization

Data Summarization gives the user an overview of the structure of the data and is generally carried out in the early stages of a project. This type of initial exploratory data analysis can help to understand the nature of the data and to find potential hypotheses for hidden information. Simple descriptive statistical and visualization techniques generally apply [1].

Summarization is the generalization or abstraction of data. A set of relevant data is abstracted and summarized, resulting a smaller set which gives a general overview of loan borrowers (Name, Age, and Income) who serve as training set, a classification model can be built, which concludes bank loan application as either safe or risky. (If age = Youth then Loan decision = risky) [5].

2.2.1.7. Association

Association is looking for relationship between variables or objects. It aims to extract interesting association, correlations or casual structures among the objects i.e. the appearance of another set of objects. The association rules can be useful for marketing, commodity management, advertising etc. Association rule learning is a popular and well researched method for discovering interesting relations between variables in large databases. It is intended to identify strong rules discovered in databases using different measures of interestingness and based on the concept of strong rules introduced association rules for discovering regularities between products in large-scale transaction data recorded by point-of-sale (POS) systems in supermarkets [5].

The association rules can be useful for marketing, commodity management, advertising, etc. For example, the rule {Onions, potatoes}{burger} found in the sales data of a supermarket would indicate that if a customer buys onions and potatoes together, he or she is likely to also buy hamburger meat. Such information can be used as the basis for decisions about marketing activities

such as, e.g., promotional pricing or product placements. In addition to the above example from market basket analysis association rules are employed today in many application areas including Web usage mining, intrusion detection, Continuous production, and bioinformatics [6].

2.2.1.8. Dependency Analysis

Dependency Analysis deals with finding a model that describes significant dependencies (or associations) between data items or events. Dependencies can be used to predict the value of an item given information on other data items. Dependency analysis has close connections with classification and prediction because the dependencies are implicitly used for the formulation of predictive models. Correlation analysis, regression analysis, association rules, case-based reasoning and visualization techniques generally apply [1].

2.2.1.9. Time Series and Trend Analysis

Analyzing time series is the process of using statistical techniques to model and explain a time-dependent series of data points, while time series forecasting is the process of using a model to generate predictions (forecasts) for future events based on known past events. Examples of time series applications include: capacity planning, inventory replenishment, sales forecasting and future staffing levels. Time series data has a natural temporal ordering, which differs from typical data mining learning applications where each data point is an independent example of the concept to be learned, and the ordering of data points within a data set does not matter [5].

Trend Analysis discovers interesting patterns in the evolution of history of the objects. One topic in trend analysis is the identification of an objects evolution, such as up, down, pick, valley, etc. A model or function is constructed to simulate the behaviors of the object, which can be used to predict the future behaviors. Typical tasks are forecasting of future values, finding typical patterns in a time series or finding similar time series by clustering. The analysis of time series plays an import role in many different applications, including prediction of stock markets, forecasting of energy consumption and other markets, and quality supervision in production, and is also supported by most data mining tools [6, 7].

2.2.1.10. Description and Visualization

The purpose of data mining is sometimes simply to describe what is going on in a complicated database in a way that increased our understanding of the people, products or processes that produced the data in the first place. They state that a good enough description of behavior will often

suggest an explanation for it as well [1, 6, 7]. One of the most powerful forms of descriptive data mining is data visualization. Although visualization is not always easy, the right picture can truly speak a thousand words since human beings are extremely practiced at extracting meaning from visual scenes. Visualization can be useful in providing a visual representation of the location and distribution of a company's major clients on a map of a city or a province or even a country. Allowing the visualization of discovered patterns in various forms can help users with different backgrounds to analyze and interpret easily.

2.2.2. Data Mining Applications

Various fields use data mining technologies because of fast access of data and valuable information from vast amount of data. Data mining technologies have been applied successfully in many areas like marketing, telecommunication, fraud detection, and finance, medical and so on. Some of the application is discussed below.

2.2.2.1. Financial Data Analysis

The financial data in banking and financial industry is generally reliable and of high quality which facilitates the systematic data analysis and data mining. Here are the few typical cases: Design and construction of data warehouses for multidimensional data analysis and data mining. Loan payment prediction and customer credit policy analysis. Classification and clustering of customers for targeted marketing, Detection of money laundering and other financial crimes [23].

Forecasting stock market, currency exchange rate, bank bankruptcies, understanding and managing financial risk, trading futures, credit rating, loan management, bank customer profiling, and money laundering analyses are core financial tasks for data mining [19, 20, 21, 22].

Stock market forecasting includes uncovering market trends, planning investment strategies, identifying the best time to purchase the stocks and what stocks to purchase. Financial institutions produce huge datasets that build a foundation for approaching these enormously complex and dynamic problems with data mining tools. Potential significant benefits of solving these problems motivated extensive research for years [31, 32, 34, 35].

Generally speaking, [19, 20, 23] essentials of data mining in finance are coming from the need to:

- Forecast *multidimensional time series* with high level of *noise;*

- Accommodate specific efficiency criteria (e.g., the maximum of *trading profit*) in addition to prediction accuracy;
- Make coordinated *multi-resolution forecast* (minutes, days, weeks, months, and years);
- Incorporate a *stream of text signals* as input data for forecasting models (e.g., Enron case, September 11 and others);
- Be able to *explain the forecast* and the *forecasting model* ("black box" models have limited interest and future for significant investment decisions);
- Be able to benefit from very *subtle patterns* with a *short life time;* and
- Incorporate the impact of market players on market regularities.

2.2.2.2. Retail Industry

Data Mining has its great application in Retail Industry, because it collects large amount of data from on sales, customer purchasing history, goods transportation, consumption and services. It is natural that the quantity of data collected will continue to expand rapidly because of increasing ease, availability and popularity of web. The Data Mining in Retail Industry helps in identifying customer buying patterns and trends. That leads to improved quality of customer service and good customer retention and satisfaction. Here is the list of examples of data mining in retail industry [5]:

- Design and Construction of data warehouses based on benefits of data mining.
- Multidimensional analysis of sales, customers, products, time and region.
- Analysis of effectiveness of sales campaigns.
- Customer Retention.
- Product recommendation and cross-referencing of items.

Moreover, sales forecasting looks at when customers bought, and tries to predict when they will buy again. We could use this type of analysis to determine a strategy of planned obsolescence or figure out complimentary products to sell. It also looks at the number of customers in the market and predicts how many will actually buy. When it comes to forecasting sales, create three cash flow projections: *realistic*, *optimistic* and *pessimistic*. This way we can plan to have the right amount of capital on hand to endure the worst situation possible if sales don't go as planned [18].

2.2.2.3. Telecommunication Industry

Today the Telecommunication industry is one of the most emerging industries providing various services such as fax, pager, cellular phone, Internet messenger, images, email, web data transmission etc. Due to the development of new computer and communication technologies, the telecommunication industry is rapidly expanding. This is the reason why data mining is become very important to help and understand the business. Data Mining in Telecommunication industry helps in identifying the telecommunication patterns, catch fraudulent activities, make better use of resource, and improve quality of service. Here is the list examples for which data mining improve telecommunication services [5]:

- Multidimensional Analysis of Telecommunication data.
- Fraudulent pattern analysis.
- Identification of unusual patterns
- Multidimensional association and sequential patterns analysis.
- Mobile Telecommunication services.
- Use of visualization tools in telecommunication data analysis.

2.2.2.4. Biological Data Analysis

Now a days we see that there is vast growth in field of biology such as genomics, proteomics, functional Genomics and biomedical research. Biological data mining is very important part of Bioinformatics. Following are the aspects in which Data mining contribute for biological data analysis:

- Semantic integration of heterogeneous, distributed genomic and proteomic databases.
- Alignment, indexing, similarity search and comparative analysis multiple nucleotide sequences.
- Discovery of structural patterns and analysis of genetic networks and protein pathways.

2.2.2.5. Other Scientific Applications

The applications discussed above tend to handle relatively small and homogeneous data sets for which the statistical techniques are appropriate. Huge amount of data have been collected from scientific domains such as geosciences, astronomy etc. There is large amount of data sets being generated because of the fast numerical simulations in various fields such as climate, and ecosystem modeling, chemical engineering, fluid dynamics etc.

2.3. Predictive Data Mining, Modeling and Analytics

Data mining techniques are the result of a long process of research and product development. This evolution began when business data was first stored on computers, continued with improvements in data access, and more recently, generated technologies that allow users to navigate through their data in real time. Data mining takes this evolutionary process beyond retrospective data access and navigation to prospective and proactive information delivery. Data mining is ready for application in the business community because it is supported by three technologies Massive data collection, Powerful multiprocessor computers and Data mining algorithms. Data mining tools are used to predict future trends and behaviors. In order to make fact based decisions that have a positive effect on their business performance companies use Predictive Analytics and period Predictive Analytics uses analysis of current and historical data to predict future outcomes. The basic application of predictive analytics is to capture relationships in historical data to predict future outcomes, to guide decision making with minimum risks, searching the databases for hidden patterns, allowing businesses to make proactive, knowledge-driven decisions and to answer business questions that traditionally were too time consuming [33]. Most companies already collect and refine massive quantities of data. Business intelligence organizations, financial analysts, healthcare management and medical diagnosis use data mining to extract information from the enormous data sets generated by modern experimental and observational methods.

In recent years, Data Mining (DM) has become one of the most valuable tools for extracting and manipulating data and for establishing patterns in order to produce useful information for decision-making [21]. Nearly all areas of life activities demonstrate a similar pattern. Whether the activity is finance, banking, marketing, retail sales, production, population study, employment, human migration, health sector, monitoring of human or machines, science or education, all have ways to record known information but are handicapped by not having the right tools to use this known information to tackle the uncertainties of the future. Planning for the future is very important in business. Estimates of future values of business variables are needed. The commodities industry needs prediction or forecasting of supply, sales, and demand for production planning, sales, marketing and financial decisions [22, 23].

2.3.1. Predictive Modeling

Predictive data mining (PDM) works the same way as does a human in handling data analysis for a small data set; however, PDM can be used for a large data set without the constraints that a human analyst has. PDM "learns" from past experience and applies this knowledge to present or future situations. As [26] Illustrates Predictive data-mining tools are designed to help us understand what the “gold,” or useful information looks like and what has happened during past “gold-mining” procedures. Therefore, the tools can use the description of the “gold” to find similar examples of hidden information in the database and use the information learned from the past to develop a predictive model of what will happen in the future.

In a broader sense, the ultimate aim of data mining is prediction; therefore, predictive data mining is the most common type of data mining and is the one that has the most application to businesses or life concerns.DM starts with the collection and storage of data in the data warehouse. Data collection and warehousing is a whole topic of its own, consisting of identifying relevant features in a business and setting a storage file to document them. It also involves cleaning and securing the data to avoid its corruption. A data ware house is a copy of transactional or non-transactional data specifically structured for querying, analyzing, and reporting. Data exploration, which follows, may include the preliminary analysis done to data to get it prepared for mining. The next step involves feature selection and or reduction. Mining or model building for prediction is the third main stage, and finally come the data post-processing, interpretation, and/or deployment. The following diagram represents the main stages of Predictive Data Mining [25].

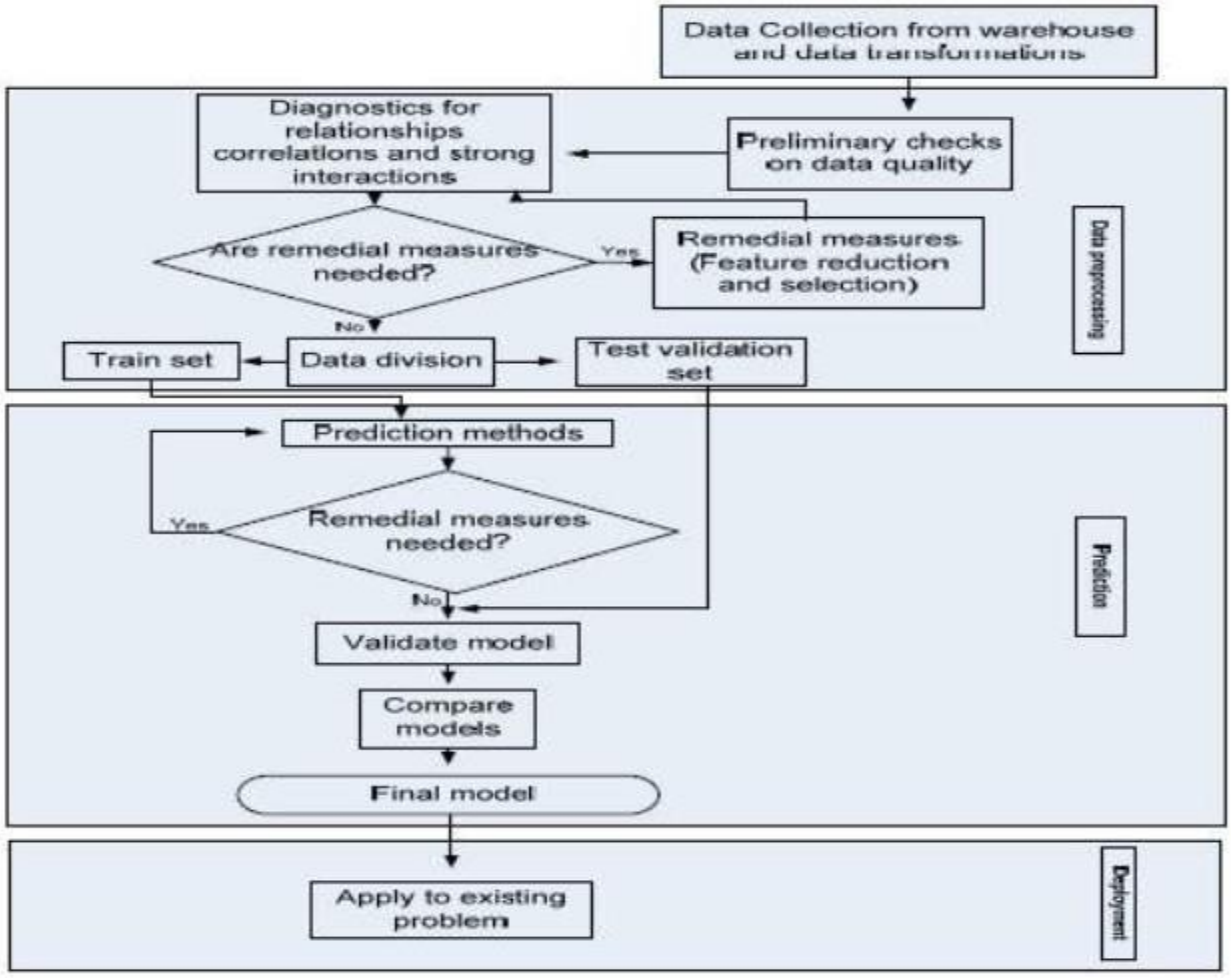

Figure 2.2 The Stages of Predictive Data Mining

Usually, data mining tasks can be categorized into either prediction or description. Descriptive mining techniques are Clustering, Association Rule Mining (ARM) and Sequential pattern mining [9]. The predictive mining techniques are Classification, Regression and Deviation detection [10]. One of the most-used subfield of data mining is predictive modeling which is a combination of statistics, machine learning, database techniques, pattern recognition, and optimization techniques.

Predictive modeling is a process used in predictive analytics to create a statistical model of future behavior. Predictive analytics is the branch of data mining concerned with the prediction of future probabilities and trends. The central element of predictive analytics is the predictor, a variable that can be measured for an individual or other entity to predict future behavior. Multiple predictors are combined into a predictive model, which, when subjected to analysis, can be used to forecast future probabilities with an acceptable level of reliability.

In predictive modeling, data is collected for the relevant predictors, a statistical model is formulated, predictions are made and the model is validated (or revised) as additional data becomes available. The model may employ a simple linear equation or a complex neural network, mapped out by sophisticated software.

Models are created using data from the past in order for the model to make predictions about the future. This process is called training the model. In this step, the data mining algorithms find patterns that are of predictive value. Next, the model is refined using the test set. The model needs to be refined to prevent it from memorizing the training set. This step ensures that the model is more general (i.e. stable) and will perform well on unseen data. Next, the performance of the model is estimated using the evaluation set. The evaluation set is entirely separate and distinct from the training and test sets. The evaluation set (or hold out set) is used to assess the expected accuracy of the model when it is applied to data outside the model set. Finally, the model is applied to the score set. The score set is not pre-classified and is not part of the model set used to create the data model. The outcomes for the score set are not known in advance. The final model is applied to the score set to make predictions.

The predictive sores will, presumably, be used to make more informed business decisions. The process is summarized in Figure 2.3: below [1].

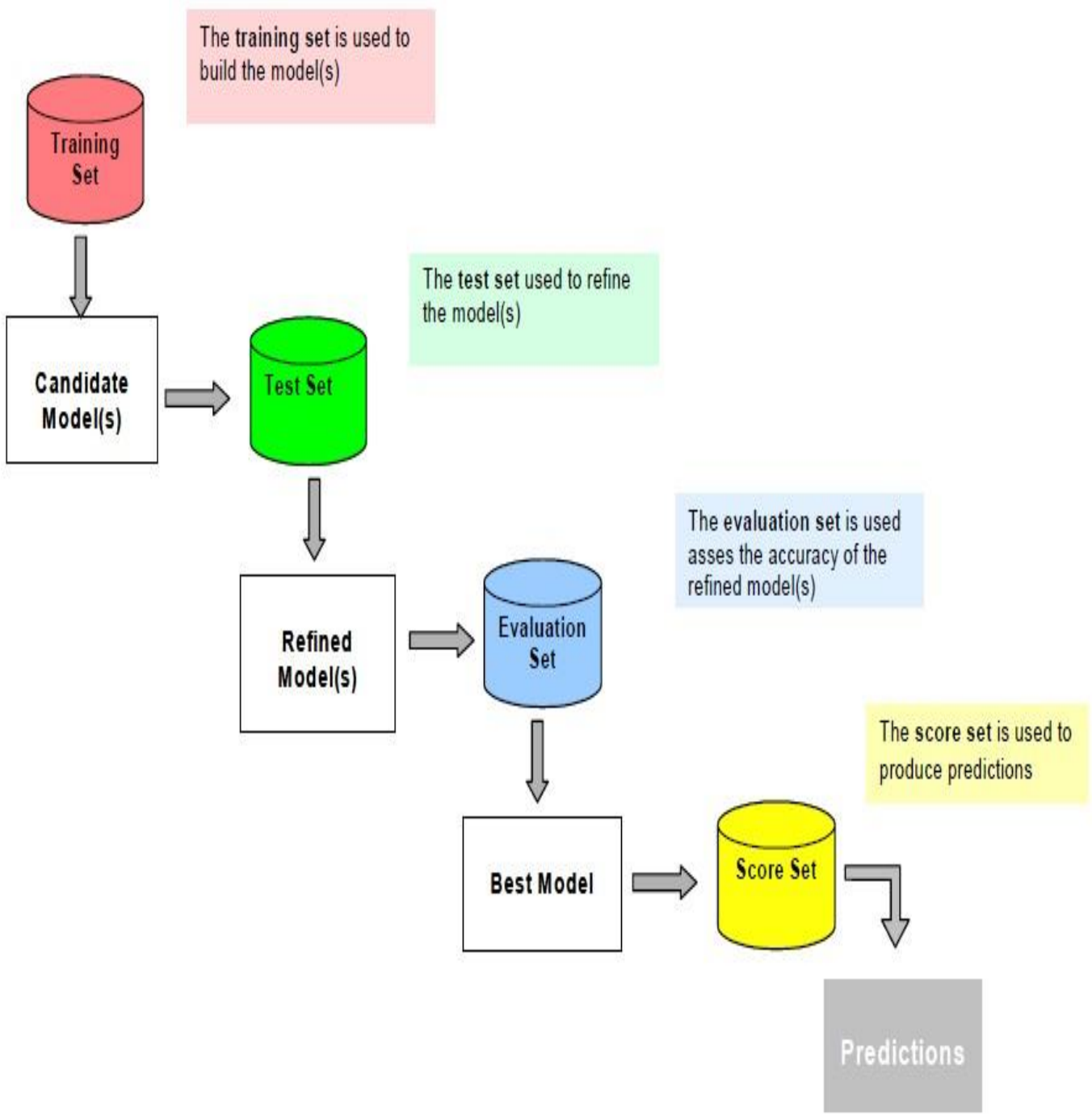

Figure 2.3 The Process of Building a Predictive Model

2.3.2. Predictive Analytics

Predictive analytics is the branch of data mining concerned with the prediction of future probabilities and trends. The central element of predictive analytics is the predictor, a variable that can be measured for an individual or other entity to predict future behavior. Multiple predictors are combined into a predictive model, which, when subjected to analysis, can be used to forecast future probabilities with an acceptable level of reliability [33].

Predictive analytics are applied to many research areas, including meteorology, security, genetics, economics, and marketing. Companies use predictive modeling tools for strategic decision-making. It helps companies identify and account for the key assumptions that drive business value-enabling good decision-making that leads to predictable results. In order to make the right decision we have to anticipate and plan for possible changes in the future. By analyzing the company's historical information we can anticipate these changes [25, 26, 33]. As [33] suggested that, to anticipate possible changes in the future, we must start addressing questions about the future possible outcomes, like Which ones are most likely?, Which ones matter most?, Which ones are best for the company?

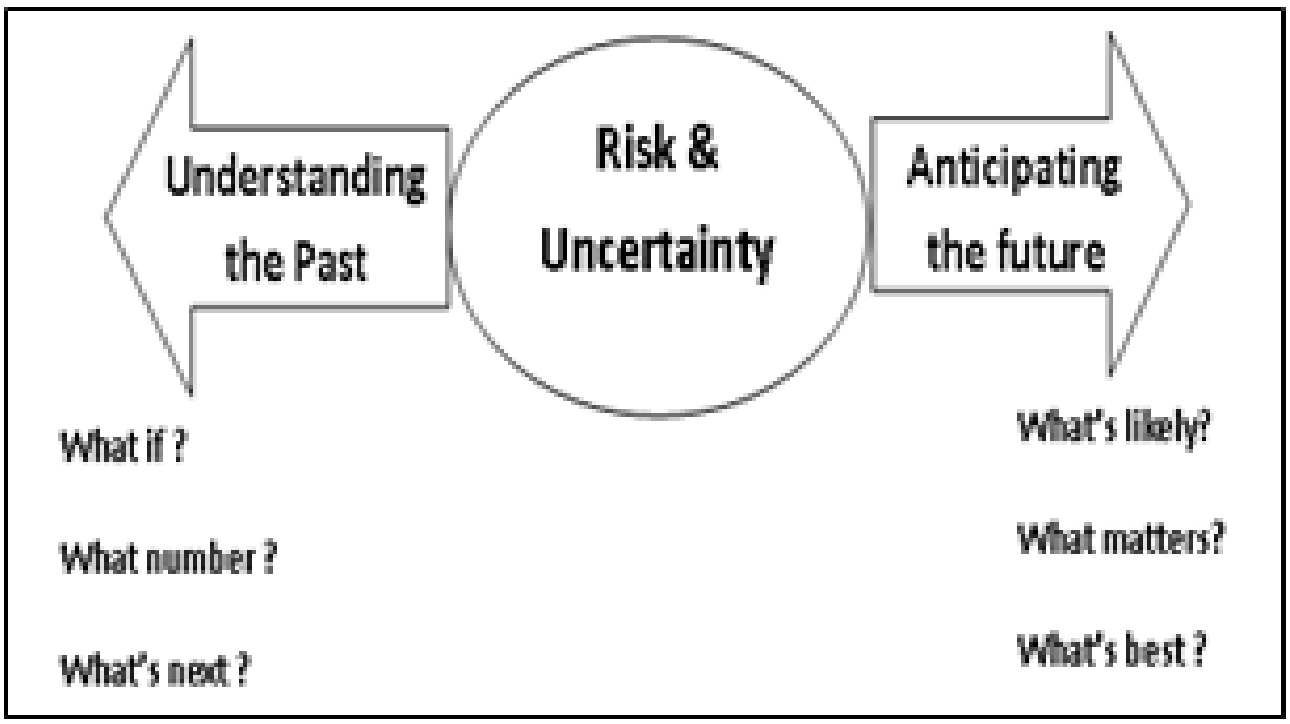

Figure 2.4 Understanding the past can help model the future

2.3.2.1. Predictive Analytics Techniques

As discussed on [26] Following are some of the basic predictive analytics techniques so far.

- ***Linear Regression*** is the first kind of regression analysis to be studied and applied in various practical applications like epidemiology, environmental science and finance. In a very lucid term, linear regression otherwise called as straight line regression analysis is a regression to estimate the unknown effect of changing one variable over another. Specifically, it models Y as a linear function of X i.e. how much Y changes when X changes one unit. So it is expressed as a straight line equation: $Y = b + wX$ (1) Here b and w are the regression coefficients where b is the Y intercept and w is the slope of the line. In cases, the coefficients can be assumed to be weights, where: $Y=w_0+w1X$ *(2)* This can be solved for the coefficients by method of least squares so as to minimize the error between the actual data and the estimated data. A training data set D, consisting of several predictor variable X and response variable Y, $|D| = \{ (X1 , Y1), (X2 , Y2)....(X|D| , Y|D|) \}$ (3) The estimated regression coefficient is given as: $w_1 = \Sigma_i (X_i – Mean\ X) (Yi – Mean\ Y) / \Sigma i\ (Xi – Mean\ x)^2$ *where* $i = 1$ *to* $|D|$, (4) and $w_0=Mean_Y–w_1\text{-}Mean_x$ *(5)*Linear regression model identifies the relationship between a single predictor variable Xi and the response variable Y when all other predictor variables in the model are “held fixed”. This is called as the *unique effect Xi* on *Y*.
- ***Multiple Linear Regression (MLR)*** is a mathematical technique that uses a number of variables to predict some unknown variable. It is a study on the relationship between a single dependent variable and one or more independent variables. This model describes a dependent variable Y by independent variable X1, X2…Xp (p>1) is expressed by the equation as , $Y = \alpha + \Sigma k\ \beta k\ Xk + \epsilon$ (6) Where $\alpha, \beta k\ (k = 1,2...p)$ are the parameters and ϵ is the error term. MLR combines the idea of correlation and linear regression.
- ***Logistic Regression*** is a type of predictive model that can be used when the target variable is categorical variable that has exactly two categories like, win game/doesn’t win, live/die. Technically it can be said as logistic regression is used for binomial regression. Simultaneously it also applies to continuous target variable that models the probability of some event occurring as a linear function of a set of predictor variable. Due to this it has extensively applied in the field of medical sciences, marketing application and social sciences. Mathematically, $f(Z) = eZ / (eZ + 1) = 1 / (1+e^{-Z})$ (7) Z is called as the logit, exposure to some set of independent variable is probability of a particular outcome. Logistic regression takes Z as input and outputs $f(Z)$ i.e. it can take input as any value from negative to positive infinity and give output between 0 and 1.

- ***Time Series Forecasting*** predicts the future value of a measure based on past values. Time series forecasting uses a model to forecast future events based on known past events. Examples include stock prices and sales revenue.
- ***Data Profiling and Transformation*** uses functions that analyze row and column attributes and dependencies, change data formats, merge fields, aggregate records, and join rows and columns.
- ***Bayesian Analytics*** capture the concepts used in probability forecasting. It is a statistical procedure which estimate parameters of an underlying distribution based on the observed distribution.
- ***Regression Analysis*** is a statistical tool for the investigation of relationships between variables. Usually, the investigator seeks to ascertain the causal effect of one variable upon another-the effect of a price increase upon demand, for example, or the effect of changes in the money supply upon the inflation rate.
- ***Classification*** used attributes in data to assign an object to a predefined class or predict the value of a numeric variable of interest. Examples include credit risk analysis, likelihood to purchase. Examples include acquisition, cross-sell, attrition, credit scoring and collections.
- ***Clustering or Segmentation*** separates data into homogeneous subgroups based on attributes. Clustering assigns a set of observations into subsets (clusters) so that observations in the same cluster are similar. An example is customer demographic segmentation.

2.4. Review on Existing Methodologies of Data Mining Researches

It this section, we reviewed basic existing Foundations (Theory)-Oriented Frameworks for DM based on different paradigms, originating from statistics, data compression, machine learning, databases, granular computing and philosophy of science, and the most well-known Process-Oriented frameworks for DM, including *Fayyad's (KDD)* , *SEMMA*, *CRISP- DM* , and *Reinartz's/Hybrid* frameworks.

2.4.1. Foundations (Theory)-Oriented Frameworks for Data Mining

Frameworks of this type are based mainly on one of the following paradigms, which will be discussed subsequently.

2.4.1.1. Data Mining and Statistics

The Statistical Paradigms

Generally, it is possible to consider the task of DM from the statistical point of view, emphasizing the fact that DM techniques are applied to larger datasets than it is commonly done in applied statistics [2]. Thus the analysis of appropriate statistical literature, where strong analytical background is accumulated, would solve most DM problems. Many DM tasks naturally may be formulated in the statistical terms, and many statistical contributions may be used in DM in a quite straightforward manner [15].

According to [2] there exist two basic statistical paradigms that are used in theoretical support for DM. The first paradigm is so-called "Statistical experiment". The second one is, an evolved version of the "Statistical experiment" paradigm is the "Statistical learning from empirical process" paradigm. Generally, many DM tasks can be seen as the task of finding the underlying joint distribution of variables in the data. Good examples of this approach would be a Bayesian network or a hierarchical Bayesian model, which give a short and understandable representation of the joint distribution. DM tasks dealing with clustering and/or classification fit easily into this approach. The second statistical paradigm is called "Structural data analysis" and can be associated with singular value decomposition methods, which are broadly used, for example, in text mining applications [2].

Some of the basic Predictive Statistical Analysis techniques are [23]:

- *Cluster Analysis ,Correlation Analysis, Discriminant Analysis, Factor Analysis, Regression Analysis* and *Logistic Regression*

Moreover there are also Descriptive and Visualization Techniques that include simple descriptive statistics described in such as:

- *Averages and measures of variation, Counts and percentages, and Cross-tabs* and *simple correlations*

They are useful for understanding the structure of the data. Visualization is primarily a discovery technique and is useful for interpreting large amounts of data; visualization tools include histograms, box plots, scatter diagrams, and multi-dimensional surface plots.

2.4.1.2. Data Mining and Data Warehousing

The Data Compression Paradigms

According to [2], the data compression approach to DM can be stated in the following way:

compress the dataset by finding some structure or knowledge within it, where knowledge is interpreted as a representation that allows coding the data using a fewer amount of bits. For example, the minimum description length (MDL) principle can be used to select among different encodings accounting for both the complexity of a model and its predictive accuracy. On the other hand, simplicity is the remedy for being ad hoc both in the recommendations of the philosophy of science and in the practice of machine learning.

Many (if not all) DM techniques can be viewed in terms of the data compression approach [14]. For example, association rules and pruned decision trees can be viewed as ways of providing compression of parts of the data. Clustering can also be considered as a way of compressing the dataset. There is a connection with the Bayesian theory for modeling the joint distribution {any compression scheme can be viewed as providing a distribution on the set of possible instances of the data.

The Data Base Paradigms

A database perspective on DM and knowledge discovery was introduced in [12]. The main postulate of their approach is: "there is no such thing as discovery; it is all in the power of the query language". That is, one can benefit from viewing common DM tasks not as the dynamic operations constructing the new pieces of information, but as operations finding unknown (i.e. not found so far) but existing parts of knowledge.

In an inductive databases framework for the DM and knowledge discovery in databases (KDD) modeling was introduced. The basic idea here is that the data-mining task can be formulated as locating interesting sentences from a given logic that are true in the database. Then knowledge discovery from data can be viewed as querying the set of interesting sentences. Therefore the term "an inductive database" refers to such a type of databases that contains not only data but a theory about the data as well [2]. This approach has some logical connection to the idea of deductive databases, which contain normal database content and additionally a set of rules for deriving new facts from the facts already present in the database.

The Data Warehouse Paradigms

The construction of a data warehouse, which involves data cleaning and data integration, can be viewed as an important pre-processing step for data mining. However, a data warehouse is not a

requirement for data mining. Building a large data warehouse that consolidates data from multiple sources, resolves data integrity problems, and loads the data into a database, can be an enormous task, sometimes taking years and costing millions of dollars [15]. If a data warehouse is not available, the data to be mined can be extracted from one or more operational or transactional databases, or data marts. Alternatively, the data mining database could be a logical or a physical subset of a data warehouse.

Data mining uses the data warehouse as the source of information for knowledge data discovery (KDD) systems through an amalgam of artificial intelligence and statistics-related techniques to find associations, sequences, classifications, clusters, and forecasts [59]. Almost all data enter the warehouse from the operational environment. The data are then "cleaned" and moved into the warehouse. The data continue to reside in the warehouse until they reach an age where one of three actions is taken: the data are purged; the data, together with other information, are summarized; or the data are archived. An aging process inside the warehouse moves current data into old detail data.

Typically the data warehouse architecture has three components [59]:

- *Data acquisition software (back-end)* which extracts data from legacy systems and external sources, consolidates and summarizes the data, and loads them into the data warehouse.
- *The data warehouse* itself contains the data and associated database software. It is often referred to as the "target database."
- *The client (front-end) software*, which allows users and applications (such as DSS and EIS) to access and analyze data in the warehouse.

These three components may reside on different platforms, or two or three of them may be on the same platform. Regardless of the platform combination, all three components are required.

2.4.1.3. Data Mining and Machine Learning

The Machine Learning Paradigms

The machine learning (ML) paradigm, "let the data suggest a model", can be seen as a practical alternative to the statistical paradigm "fit a model to the data". It is certainly reasonable in many situations to fit a small dataset to a parametric model based on a series of assumptions. However, for applications with large volumes of data under analysis the ML paradigm may be beneficial because of its flexibility with a nonparametric, assumption-free nature.

Machine learning is the study of computational methods for improving performance by mechanizing the acquisition of knowledge from experience [2]. Machine learning aims to provide increasing levels of automation in the knowledge engineering process, replacing much time-consuming human activity with automatic techniques that improve accuracy or efficiency by discovering and exploiting regularities in training data1. Although machine learning algorithms are central to the data mining process, it is important to note that the process also involves other important steps, including [39]:

- Building and maintaining the database,
- Data formatting and cleansing,
- Data visualization and summarization,
- The use of human expert knowledge to formulate the inputs to the learning algorithm and to evaluate the empirical regularities it discovers, and
- Determining how to deploy the results.

Following are the basic learning algorithms [39, 57] :

- *Neural Networks (NN), Case-Based Reasoning (CBR), Genetic Algorithms (GA), Decision Trees (DT)* and *Association Rules (AR)*

While these so-called first-generation algorithms are widely used, they have significant limitations. They typically assume the data contains only numeric and textual symbols and do not contain images. They assume the data was carefully collected into a single database with a specific data mining task in mind. Furthermore, these algorithms tend to be fully automatic and therefore fail to allow guidance from knowledgeable users at key stages in the search for data regularities.

2.4.1.4. Data Mining and OLAP

The Granular-Computing Approach

Generally, granular computing is a broad term covering theories, methodologies, and techniques that operate with subsets, classes, and clusters (called granules) of a universe. Granular computing concept is widely used in computer science and mathematics. Recently, [2] reviewed the concepts of fuzzy information granulation and considered it in the context of human reasoning and fuzzy logic. Some others proposed to use the term "granular computing" to label the computational theory of information granulation. In the same paper Lin introduces a view on DM as a "reverse" engineering of database processing. While database processing organizes and stores data according to the given structure, DM is aimed at discovering the structure of stored data. Lin defines automated DM as "a

process of deriving interesting (to human) properties from the underlying mathematical structure of the stored bits and bytes". Assuming that the underlying mathematical structure of a database relation is a set of binary relations or a granular structure, Lin considers DM as a processing of the granules or structure-granular computing. And then if there is no additional semantics, then the binary relations are equivalence relations and granular computing reduces to the rough set theory. However, since in the DM process the goal is to derive also the properties of stored data, additional structures are imposed. To process these additional semantics, Lin introduces the notion of granular computing in DM context .Yao applied the granular computing approach to machine learning tasks focusing on covering and partitioning in the process of data mining and showed how the commonly used ID3 and PRISM algorithms can be extended with the granular computing approach [57].

The OLAP Paradigms

The question of how data warehousing and OLAP relate to data mining is a question that often arises. The relationship can be succinctly captured as follows [1]: “The capability of OLAP to provide multiple and dynamic views of summarized data in a data warehouse sets a solid foundation for successful data mining.” Therefore, data mining and OLAP can be seen as tools than can be used to complement one another. The term OLAP, standing for Online Analytical Processing, is often used to describe the various types of query-driven analysis that are undertaken when analyzing the data in a database or a data warehouse. OLAP provides for the selective extraction and viewing of data from different points of view; these views are generally referred to as dimensions. Each dimension can and generally has many levels of aggregation, i.e. a time dimension can be organized into days, weeks, and years.

The essential distinction between OLAP and data mining is that OLAP is a data summarization/ aggregation tool, while data mining thrives on detail. Data mining allows the automated discovery of implicit patterns and interesting knowledge that’s hidden in large amounts of data.

A powerful paradigm that integrates OLAP with data mining technology is OLAM (Online Analytical Mining) which is sometimes referred to as OLAP mining.

OLAM system are particularly important because most data mining tools need to work on integrated, consistent, and cleaned data, which again, requires costly data cleaning, data transformation, and data integration as pre-processing steps. A data warehouse constructed by such pre-processing serves

as a valuable source of high-quality data for OLAP as well as for OLAM. OLAM provides a multi-dimensional view of its data and creates an interactive data mining environment whereby users can dynamically select data mining and OLAP functions, perform OLAP operations (such as drilling, slicing, dicing and pivoting on the data mining results), as well as perform mining operations on OLAP results, that is, mining different portions of data at multiple levels of abstraction [30].

2.4.1.5. Data Mining and the Philosophy of Science

The categorization of subjectivist and objectivist approaches [2] can be considered in the context of DM. The possibility to compare nominalistic and realistic ontological believes gives us an opportunity to consider data that is under analysis as descriptive facts or constitutive meanings. The analysis of voluntaristics as opposed to deterministic assumptions about the nature of every instance constituting the observed data directs our attitude and understanding of that data. One possibility is to view every instance and its state as determined by the context and/or a law. Another position consists in consideration of each instance as autonomous and independent. An epistemological assumption about how a criterion to validate knowledge discovered (or a model that explains reality and allows making predictions) can be constructed may impact the selection of appropriate DM technique. From the positivistic point of view such a model-building process can be performed by searching for regularities and causal relationships between the constitutive constructs of a model. And anti-positivism suggests analyzing every individual observation trying to understand it and making an interpretation.

2.4.1.6. Data Mining and the Web

With the large amount of information available online, the Web is a fertile area for data mining and knowledge discovery [1]. In Web mining, data can be collected at the: Server-side, Client-side, Proxy servers, or Obtained from an organization's database (which may contain business data or consolidated web data). Each type of data collection differs not only in terms of the location of the data source, but also: the kinds of data available, the segment of population from which the data was collected, and its method of implementation.

A meta-analysis of the web mining literature, categorized web mining into three areas of interest based on which part of the web is to be mined: Web Content Mining, Web Structure Mining and

Web Usage Mining. The three main tasks are performed in web usage mining are: preprocessing, pattern discovery, and pattern analysis Despite being a rich source for data mining, the Web poses challenges for effective resource and knowledge discovery particularly in terms of data collection. The Web seems to be too huge for effective data warehousing and data mining. Also, Web pages are complex and lack a unifying structure. The highly dynamic nature of the Web as an information source poses challenges as well [1].

2.4.2. Processes-Oriented Frameworks for Data Mining

Many data mining process methodologies are available. However, the various steps do not differ much from methodology to methodology. They view DM as a sequence of iterative processes that include data cleaning, feature transformation, algorithm and parameter selection, and evaluation, interpretation and validation. The most well-known Process-Oriented frameworks for DM, include: Fayyad's/KDD, SEMMA, CRISP- DM, and Reinartz's / Hybrid frameworks.

2.4.2.1. KDD: Fayyad's View on the Knowledge Discovery Process

Fayyad [12] define KDD as "the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data". Before focusing on discussion of KDD as a process, we would like to make a note that this definition given by Fayyad is very capacious, it gives an idea what is the goal of KDD and in fact it is cited in many DM related papers in introductory sections. However, in many cases those papers have nothing to do with novelty, interestingness, potential usefulness and validity of patterns which were discovered or could be discovered using proposed in the papers DM techniques.

KDD process comprises many steps, which involve data selection, data pre-processing, data transformation, DM (search for patterns), and interpretation and evaluation of patterns (Figure 2.5) [12]. The steps depicted start with the raw data and finish with the extracted knowledge, which was acquired as a result of the KDD process. The set of DM tasks used to extract and verify patterns in data is the core of the process. DM consists of applying data analysis and discovery algorithms for producing a particular enumeration of patterns (or models) over the data. Most of current KDD research is dedicated to the DM step. [12] would like to clarify that according to this scheme, and some other research literature, DM is commonly referred to as a particular phase of the entire process of turning raw data into valuable knowledge, and covers the application of modeling and

discovery algorithms. In industry, however, both knowledge discovery and DM terms are often used as synonyms to the entire process of getting valuable knowledge.

Nevertheless, this core process of search for potentially useful patterns typically takes only a small part estimated at 15% - 25%) of the effort of the overall KDD process. The additional steps of the KDD process, such as data preparation, data selection, data cleaning, incorporating appropriate prior knowledge, and proper interpretation of the results of mining, are also essential to derive useful knowledge from data.

As of [2, 12] opinion the main problem of the framework presented in Figure 2.5 is that all KDD activities are seen from "inside" of DM having nothing to do with the relevance of these activities to practice (business).

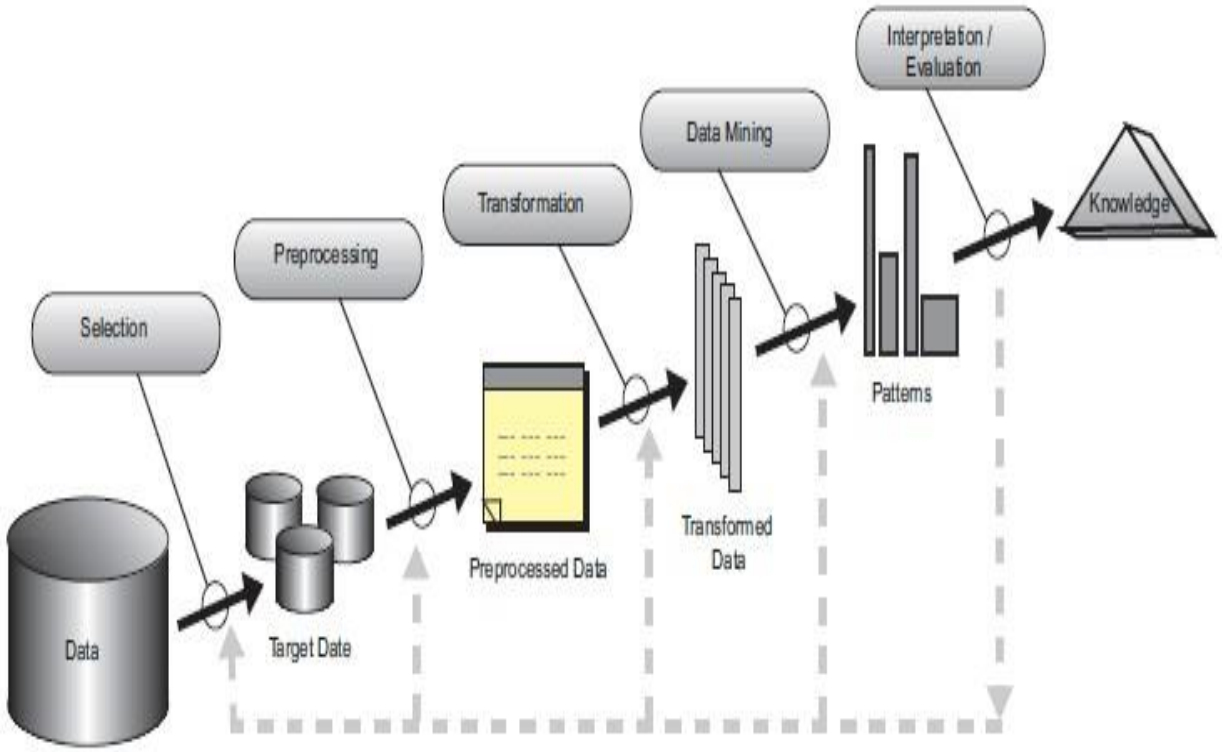

Figure 2.5 Basic steps of the KDD process

2.4.2.2. SEMMA

The two popular methodologies used by data mining tools are the SEMMA process for SAS Enterprise Miner and the 5 A's process for SPSS.

SAS - THE SEMMA ANALYSIS CYCLE

SAS developed a data mining analysis cycle known by the acronym SEMMA. This acronym stands for the five steps of the analyses that are generally a part of a data mining project: 1. **S**ample 2. **E**xplore 3. **M**odify 4. **M**odel 5. **A**ssess

As illustrated in Figure 2.6 [1]. The SEMMA analysis cycle guides the analyst through the process of exploring the data using visual and statistical techniques, transforming data to uncover the most significant predictive variables, modeling the variables to predict outcomes, and assessing the model by testing it with new data.

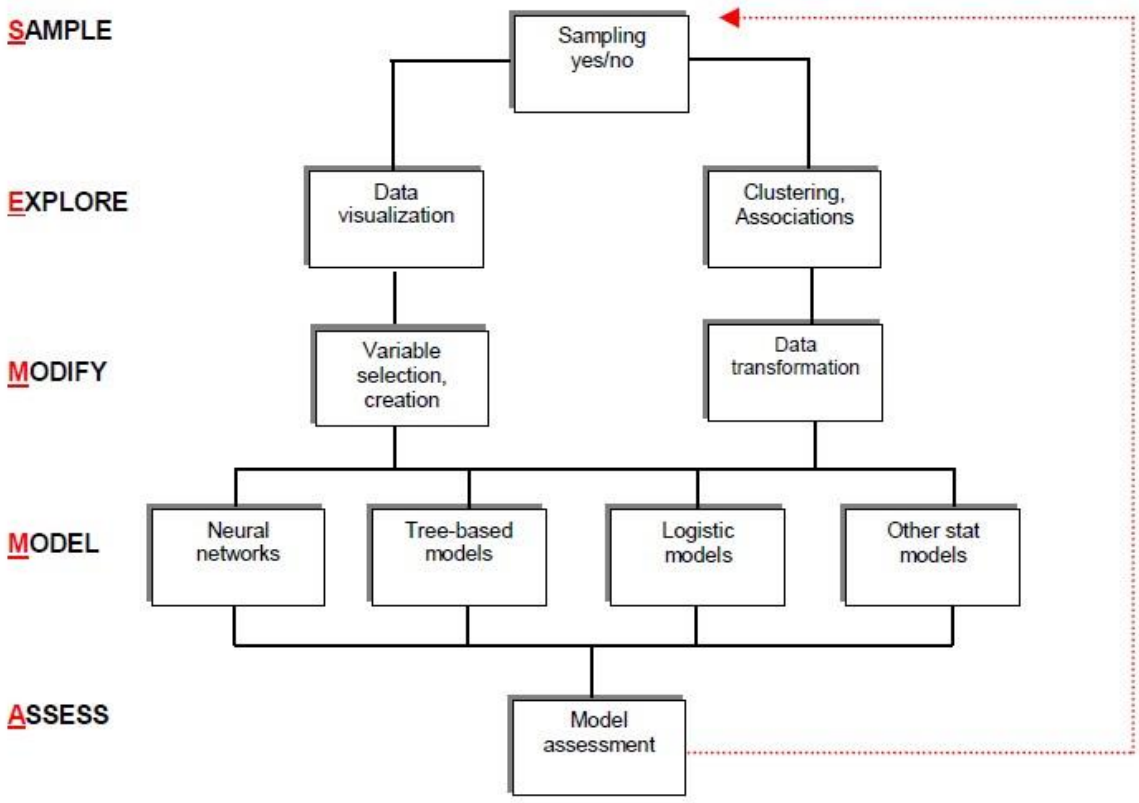

Figure 2.6 The SEMMA Analysis Cycle

SPSS - THE 5 A'S PROCESS

SPSS originally developed a data mining analysis cycle called the 5 A's Process5. The five steps in the process are: • **A**ssess · **A**ccess · **A**utomate · **A**nalyze · **A**ct

As illustrated in Figure 2.7 [1]. The 5 A's process methodology is similar to that of the SEMMA analysis cycle.

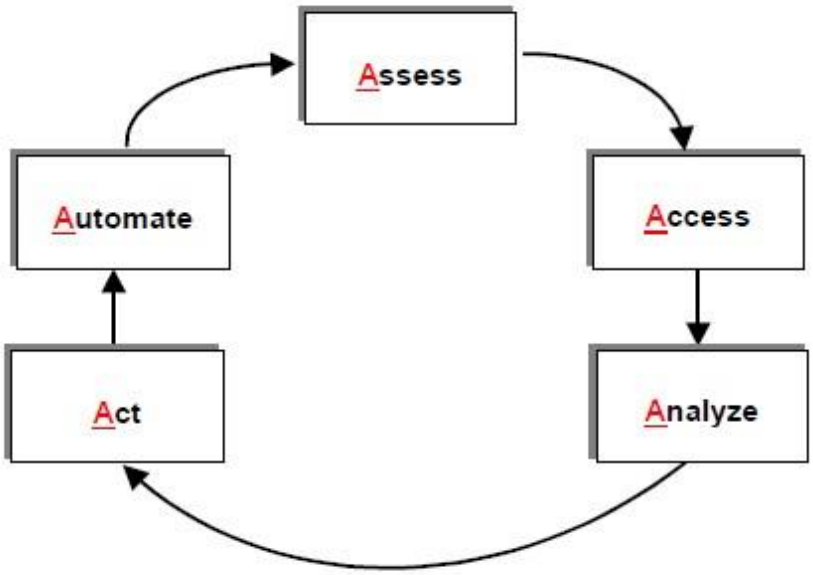

Figure 2.7 The 5A's Process

2.4.2.3. CRISP-DM: CRoss Industry Standard Process for Data Mining

The CRISP-DM (The De Facto Standard for Industry) project began in mid-1997 and was funded in part by the European commission [10]. The leading sponsors were: NCR, DaimlerChrysler, Integral Solutions Limited (ISL) (now a part of SPSS), and OHRA, a Netherlands' independent insurance company. The goal of the project was to define and validate an industry- and tool-neutral data mining process model that which would make the development of large as well as small data mining projects faster, cheaper, more reliable and more manageable. The project started in July 1997 and was planned to be completed within 18 months. However, the work of the CRISP-DM received substantial international interest, which caused the project to put emphasis on disseminating its work. As a result, the project end date was pushed back to and completed on April 30, 1999. The CRISP-DM model is illustrated in Figure 2.8 [1].

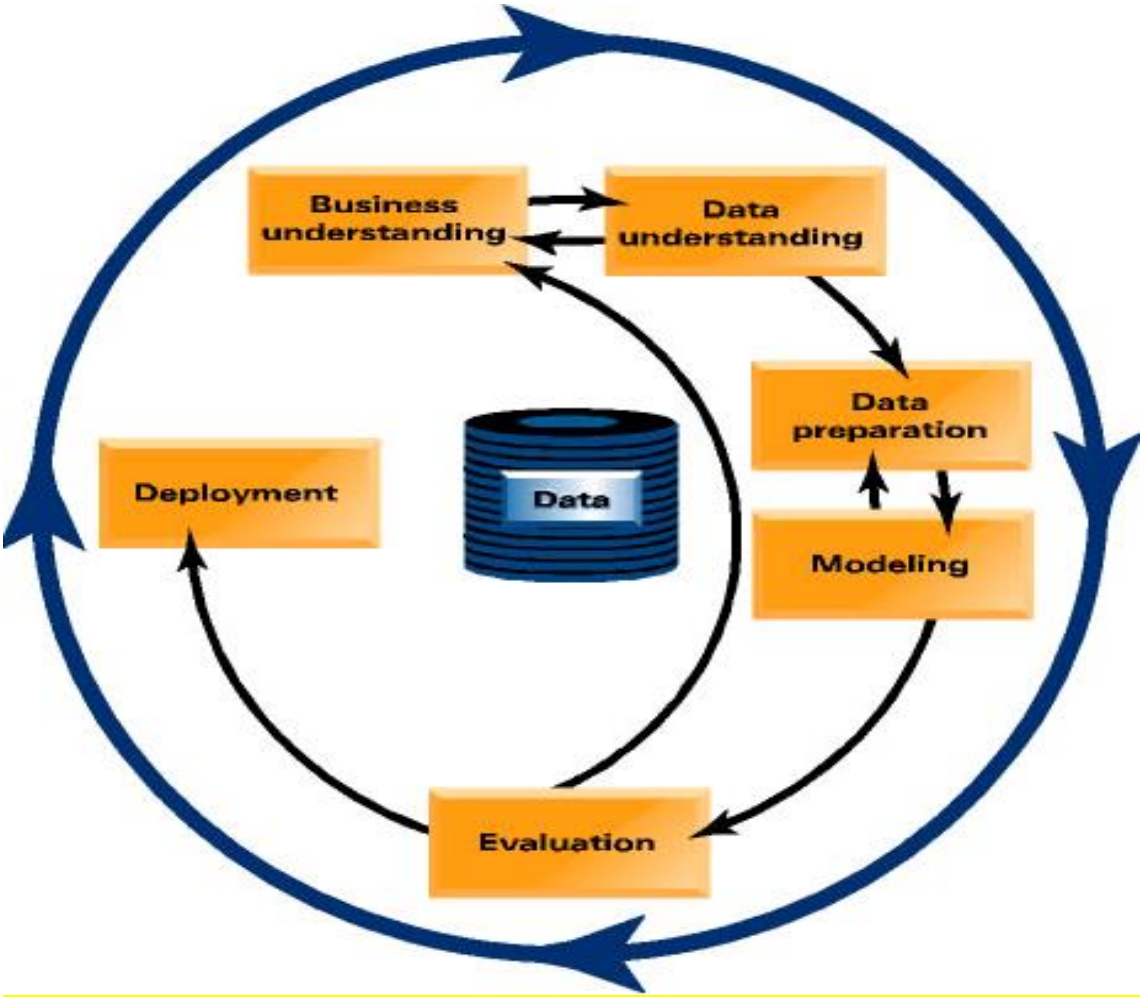

Figure 2.8 The CRISP-DM Model

2.4.2.4. Reunartz's view: Hybrid Knowledge Discovery Processes for Data Mining

Reinartz's framework [28] follows CRISP-DM with some modifications (Figure 2.9), introducing a data exploration phase and explicitly showing the accumulation of the experience achieved during the DM/KDD processes [28].

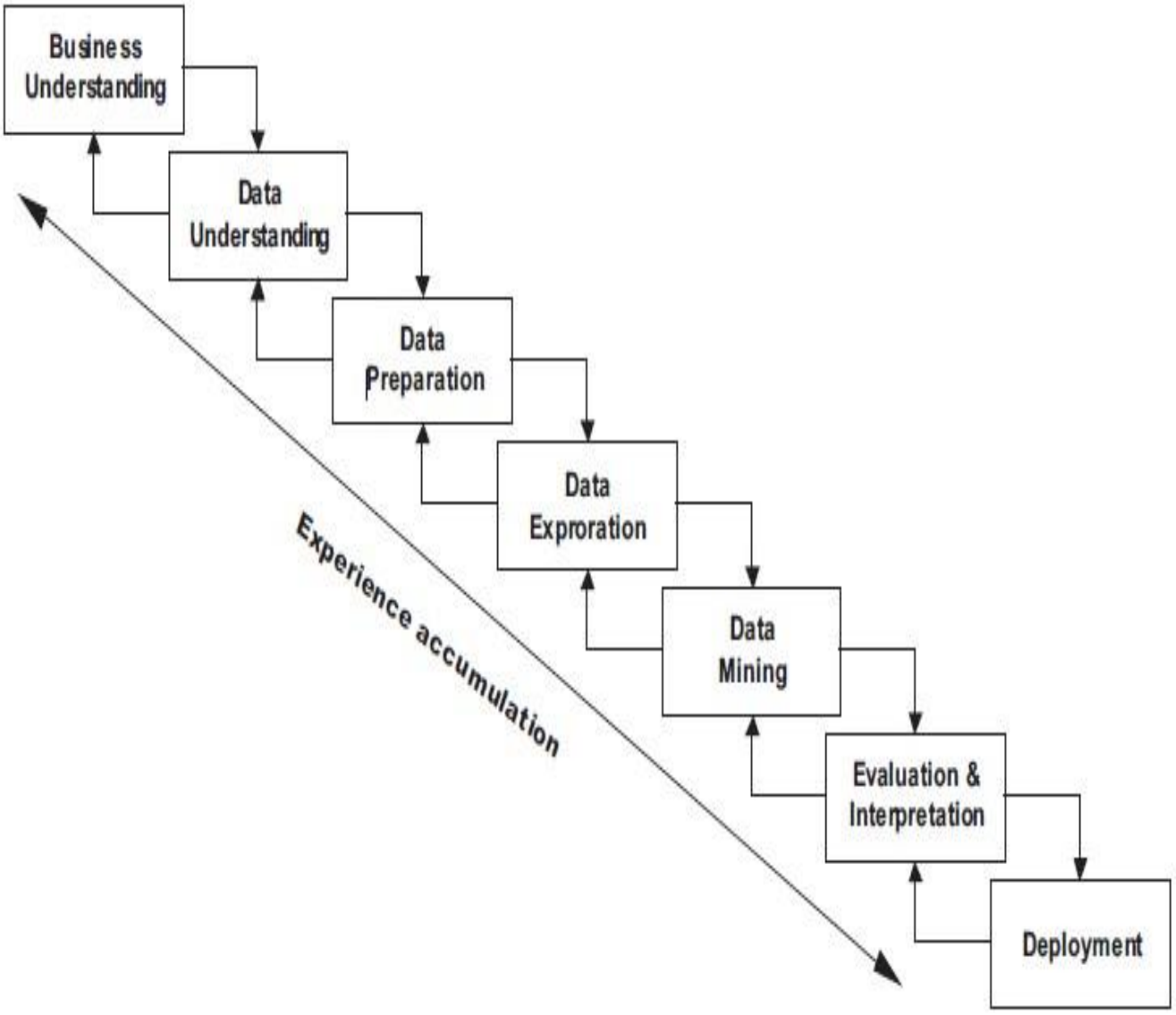

Figure 2.9 Reinartz's Framework (Knowledge Discovery Process: from problem understanding to deployment)

It is considered that, the main problem with CRISP-DM and Reinartz's frameworks is that they assume that the DM artifact is ready to be applied and easy to be deployed and used [2, 28]. Therefore, the development and use processes are almost disregarded in these frameworks though being embedded in an implicit way. Consequently, it is hard to see what the most crucial success factors of a DM project are.

2.5. Review of Related Works

Even though it is very difficult to get directly-related researches/articles, i.e. studies on "Predicting Future Market Status (in terms of price & sales volume) of Commodity Exchange/Trading Organizations", as most of the studies that we found were dealing with *Stock* Exchange/Trading companies; here are some of the researches/articles that we reviewed (relevant to my thesis paper) and depicted in a summarized way on predicting future market (trade) status (trend).

First let's begin by discussing on literatures made on agricultural commodities future market status/trend predictions.

The paper [39] applies a Multivariate Relevance Vector Machine (MVRVM) model to develop multiple-time-ahead predictions with confidence intervals of monthly agricultural commodity prices. The predictions are one, two and three months ahead of prices of cattle, hogs and corn. The MVRVM is a regression tool extension of the RVM model to produce multivariate outputs. The statistical test results indicate an overall good performance of the model for one and two month's prediction for all the commodity prices. The model performance decreased in early 2008 for the corn price predictions. The performance also decreased for the three-month prediction of the three commodity prices. The MVRVM model outperforms the Artificial Neural Network (ANN) most of the time with the exception of corn price prediction two and three months ahead. However, the bootstrap histograms of the MVRVM model show narrow confidence bounds in comparison to the histograms of the ANN model for the three commodity price forecasts. Therefore, the MVRVM is more robust.

The results presented in their paper [39] have demonstrated the overall good performance and robustness of MVRVM for simultaneous multiple-time-ahead predictions of agricultural commodity prices. The potential benefit of these predictions lies in assisting producers in making better-informed decisions and managing price risk.

A study on [40] uses the Back Propagation Neural Network (BPNN) prediction model of vegetable market price is established. Three years and three months Coimbatore market price of tomato as an example and simulated the result using Matlab and predict the result. The prediction results of weekly are discussed. Their result shows that the larger dataset produced more accuracy result than

the smaller data set. The result shows that neural network is one way of predicting the market price of vegetable with the non-linear time series. They also proposed, in future, the Genetic Algorithm based neural network will be constructed for price prediction to increase the accuracy percentage.

A study [42] made on one of the main agricultural commodities in Indonesia, known as, *Dried Eucheumacottonii Seaweed* (DES) noted that, farmer's plant, harvest, and dry the seaweed, and subsequently sell it to traders. The selling price of seaweed depends on internal and external factors. To maximize their profit, farmers have to estimate the future price development by means of these factors. Their paper presents a novel data-driven method reports on our attempts to develop methods to aid farmers in making such predictions. More specifically, they apply data mining to the task of predicting the seaweed price eight weeks ahead, at the time of selling. The data mining algorithm, i.e., classifier, requires attributes as its inputs. In their experiments, they selected three internal and three external factors as input attributes. The internal-factor attributes used are: current price, dirty content, and moisture content. The three external-factor attributes all relate to the weather and are minimum temperature, maximum temperature, and precipitation. The data is collected from two publicly available sources and consists of 275 measurements of six attributes each. They evaluated the performances of four classifiers on our data set using 10-fold cross-validation. The results obtained from their report paper revealed that they were able to predict the selling price of seaweed with an accuracy of 64%. Moreover, they presented that, from analysis of the trained classifier revealed that the main attribute used for predicting the future price is the current price. In conclusion, they suggested that it is feasible to predict the future price development of seaweed selling price in Indonesia with accuracy above chance level.

Another study made on "Forecasting Commodity Prices with Nonlinear Models" [36], stated, the problem of export dependency on primary commodities in developing countries is exacerbated by the erratic behavior of commodity prices. The author noted that, rational expectations competitive storage theory conjectures that commodity prices should follow a nonlinear process. Motivated by theoretical considerations and armed with a veritable explosion in nonlinear model developments important subgroups of primary commodities, namely industrial base metals are examined using state of the art nonlinear models. Their simulated out-of-sample results indicate that with respect to weekly and monthly data nonlinear models in all cases produce the lowest forecast errors. However, in some cases there is a doubt on whether the observed differences in forecast performance between nonlinear and linear models are statistically significant. Moreover, it is missioned that, a commodity trading framework is constructed to measure whether the forecasts generated using nonlinear models

are economically meaningful. And the author concluded that, the trading exercise reveals significant economic gains from implementing a simple trading rule based on nonlinear models.
A survey, entitled "A survey on Data Mining Techniques for Crop Yield Prediction" [51], presents the various crop yield prediction methods using data mining techniques. In their survey, they noted that, agricultural system is very complex since it deals with large data situation which comes from a number of factors. Thus, crop yield prediction has been a topic of interest for producers, consultants, and agricultural related organizations. Their paper focus on the applications of data mining techniques in agricultural field and different Data Mining techniques such as K-Means, K-Nearest Neighbor (KNN), Artificial Neural Networks (ANN) and Support Vector Machines (SVM) for very recent applications of data mining techniques in agriculture field. They also concluded that, the problem of yield prediction can be solved by employing data mining techniques.

And, following are some of the literatures that we reviewed and focused on stock trend prediction.

In a study [50], they presented regression analysis as a data mining technique and developed tool for exploiting especially time series data in financial institution. A prediction system has been built that uses data mining technique to produce periodically forecasts about stock market prices. The authors noted that, the technique complement proven numeric forecasting method using regression analysis with technology taking as input the financial information obtained from the daily activity summary (equities) published by Nigerian Stock Exchange. In the paper, they demonstrated that, they are able to use regression analysis as a data mining technique to describe the trends of stock market prices and predict the future stock market prices of three banks as a case study from banking sector of Nigerian economy.

A study on [41], as the authors declared, their paper was a continuation of their research work on the Nigerian Stock Exchange Market (NSEM) uncertainties, In their previous work they presented the Naive Bayes and SVM-SMO algorithms as a tools for predicting the Nigerian Stock Exchange Market; subsequently they used the same transformed data of the NSEM and explored the implementation of the Logistic function on Back-propagation algorithm on the WEKA platform, and results obtained, made them to also conclude that the Back-propagation model of Artificial Neural Network (ANN) performed very well and thus it is another algorithm that can effectively and efficiently be used for predicting the Nigerian Stock Exchange Market. They also indicate further works will involve research work on possibilities of using other Machines Learning algorithms to predict the NSE.

From the “A Review on Data Mining Applications to the Performance of Stock Marketing” [31], their paper revels that; data mining is the extraction of hidden predictive information from large data bases. Data mining is used to increase revenues and reduce costs. Data mining finds patterns and correlations or relationships in data by using sophisticated techniques. In today’s society of high demanding customers, increasing competitiveness and in order to avoid the initial investment of gathering a data without being certain that it will be successful. A customer-focused approach will discover customer’s preferences, needs and behavioral characteristics. The current approach is described for data development in share market to find the correlations between the customer feelings or perceptions and the physical characteristics of a data. Stock Market is an organized set-up with a regulatory body and the members who trade in shares are registered with the stock market and regulatory body SEBI. The stock market is also called the secondary market as it involves trading between two investors and it is often considered the primary indicator of a country's economic strength and development. The NSE is the largest stock exchange in India in terms of daily turnover and number of trades, for both equities and derivative trading.

Another paper [35], made on “Romanian Stock Exchange”, they discussed that, the general results show that using time series neural networks for forecasting on Romanian Stock Exchange is feasible but requires specific implementation methodology for best results. As a general observation, it is important to observe the minimum particular time span for selecting historical training data sets in order to ensure better results. Future development will focus on determining this optimum timespan for each analyzed stock in or-der to improve final results. Because a pre-learn stage is needed to find this optimum time span for each learning process, the actual model evaluation is time consuming and requires a well-organized set of experiments. The best prediction method will help end users in determining the actual buy-sell strategy. The system can be integrated with existing pre-diction tools using a stream processing layer as a bind. The actual limitations of the model are still yet to be found and solved. Many successful trading systems are limited by environmental and unexpected variables but continue to be successfully used by traders on different stock exchange markets.

In a paper [45], they used neural networks model to predict the value of stock share in the next day using the previous data about stock market value. For this purpose two different well known types of neural networks were applied to the problem. The obtained results show that for predicting the direction of changes of the values in the next day none of these methods are better than simple linear regression model. But the error of the prediction of the amount of value changes using MLP neural

network is less than both Elman network and linear regression method. In addition to this, when the feed forward MLP neural network predicts the direction of the changes correctly, the amount of change is completely close to the real one in comparison to the other two mentioned methods. Moreover they indicates, in their future works of the study, they are going to apply other recently proposed regression methods such as Support Vector Regression models which is newer in the field of machine learning researches and claimed to have good generalization ability due to application of large margin concept.

Another study made on "Predicting Stock Prices Using Polynomial Classifiers: The Case of Dubai Financial Market" [44], two prediction models were developed in their study. The first model was developed with the well-known back propagation feed forward neural network. The second model used here is based on polynomial classifiers which are being used for the first time in stock prices prediction. The inputs to both models were identical, and both models were trained and tested on the same data in three different training scenarios and two prediction modes. The data used here is the historical prices for two of the leading stocks in Dubai Financial Market. From their findings, both models achieved outstanding results in terms of mean absolute error percentage (MEAP). Both models achieved around 1.5% MEAP in predicting the next day, 2.5% MEAP in predicting the second day, and around 4% MEAP in predicting the third day. The pre-diction accuracy of the two models was certainly remarkable, where around 60% of the predicted prices of the first day, 50% of the predicted prices of the second day, and 35% of the predicted prices of the third day, were all within –1% to 1% of the actual prices of the three days. When comparing the neural network and polynomial classifiers prediction models, it was found that first order polynomial classifier performed comparable to or slightly better than the neural network. Whereas the second order polynomial classifier could barely achieve similar results on the stocks used in this study. They also indicate further work can be done using other stocks in similar emerging markets and mature markets, to verify this conclusion. On the other hand they noted that PC is a lot more computationally efficient than ANN since its weights can be obtained directly and non-iteratively from a closed formula.

From a study in [43], the author discussed, determining the Stock market forecasts has always been challenging work for business analysts. In the paper, the author attempted to make use of these huge chaotic in nature data to predict the stock market indices. Moreover, if we combine both these chaotic data and numeric time series analysis the accuracy in predictions can be achieved. Investors can use this prediction model to take trading decision by observing market behavior. Enhancements

of this system are focused to help in improving more accurate predictability in stock market regardless how chaotic the stock market data can be.

Another study [37] made on Finance Mining, "Analysis Of Stock Market Exchange For Foreign Using Classification Techniques", the authors depict that, an increase of economic globalization information technology financial data are being generated and accumulated at an unprecedented pace. Thus, a critical need for automated approaches to effective and efficient utilization of massive amount of financial data to support organizations in strategic investment decisions. They also noted that, there has been a large body of research and practice focusing on exploring data mining techniques to solve financial problems. To be successful, a data mining project should be driven by the application needs and results should be tested quickly. Their paper proposes a data mining in finance stock market and specific requirements for data mining methods including in making interpretations, incorporating relations and probabilistic learning for exchange.

And a study [46] proposes to implement a "Hybrid Clustering Method for Stock Price and Commodity Price" for a better predictive model. They find out that, combination between K-Means Clustering and Principal Component Analysis can give better analysis and classification. The results of using those two methods are a dimension reduced cluster (compact cluster). The result is supported by some findings in the reality through having some observations in the news and daily reports of the company conditions. In their summary, the present study has shown that it is an effective method to use a hybrid method to cluster the stock price and commodity price using K-Means Clustering and Principal Component Analysis. It is supported; the result of that Principal Component Analysis can have a direct impact on the number and type of dietary patterns revealed in the data. They also missioned that, reducing the dimension of the cluster is an important task and it can be implemented by using Principal Component Analysis. Moreover, for the next study, they suggested to combine the K-Means algorithm, Principal Component Analysis, and Neural Network to have better solutions. By conducting the Neural Network to the established clusters, it will give the exact information on how to identify the information in every cluster.

The paper [47] studies on "Predicting Australian Stock Market Index Using Neural Networks Exploiting Dynamical Swings and Inter-market Influences" presents a computational approach for predicting the Australian stock market index–AORD using multi-layer feed-forward neural networks from the time series data of AORD and various interrelated markets. This effort aims to discover an effective neural network or a set of adaptive neural networks for this prediction purpose, which can

exploit or model various dynamical swings and inter-market influences discovered from professional technical analysis and quantitative analysis. Within a limited range defined by their empirical knowledge, three aspects of effectiveness on data selection are considered: effective inputs from the target market (AORD) itself, a sufficient set of interrelated markets, and effective inputs from the interrelated markets. Two traditional dimensions of the neural network architecture are also considered: the optimal number of hidden layers, and the optimal number of hidden neurons for each hidden layer. Three important results were obtained: A 6-day cycle was discovered in the Australian stock market during the studied period; the time signature used as additional inputs provides useful information; and a basic neural network using six daily returns of AORD and one daily returns of SP500 plus the day of the week as inputs exhibits up to 80% directional prediction correctness.

Another study entitled "Predicting the Future of Car Manufacturing Industry using Data Mining Techniques" [33] introduces the application of data mining technology in the car manufacturing unit and obtains an analysis result from small data, may expand the sample capacity in the practical application, to obtain more accurate conclusion. The noted, some of the modifications that can be made to simple linear regression are rather than one predictor variable more predictor variables can be used. Moreover, transformations can be applied to the predictors, Predictors can be multiplied together and used as terms in the equation and modifications can be made to accommodate response predictions that just have yes/no or 0/1 values. They also indicate, further work is under progress to develop an algorithm which can handle multiple predictor variables.

CHAPTER THREE

RESEARCH DESIGN AND METHODOLOGY

Data mining methodology is designed to ensure that the data mining effort leads to a stable model that successfully addresses the problem it is designed to solve. Various data mining methodologies have been proposed to serve as blueprints for how to organize the process of gathering data, analyzing data, disseminating results, implementing results, and monitoring improvements.

The research design is aimed to build a predictive model that predicts the price and sales volume of commodities. For this particular study, a hybrid of Experimental and Design based research technique is applied.

To build the model that analyses and predicts the market trends (in terms of price and sales volume) using predictive data mining techniques, the Hybrid Knowledge Discovery Process for Data Mining is used. This methodology was developed, by adopting the CRISP-DM (Cross-Industry Standard Processes for Data Mining) model to the needs of academic research community [10, 28].

The model is also termed as The Six Step Cios et al. [24] Process Model which consists of the following six steps as depicted below (in Figure 3.1 and subsections):

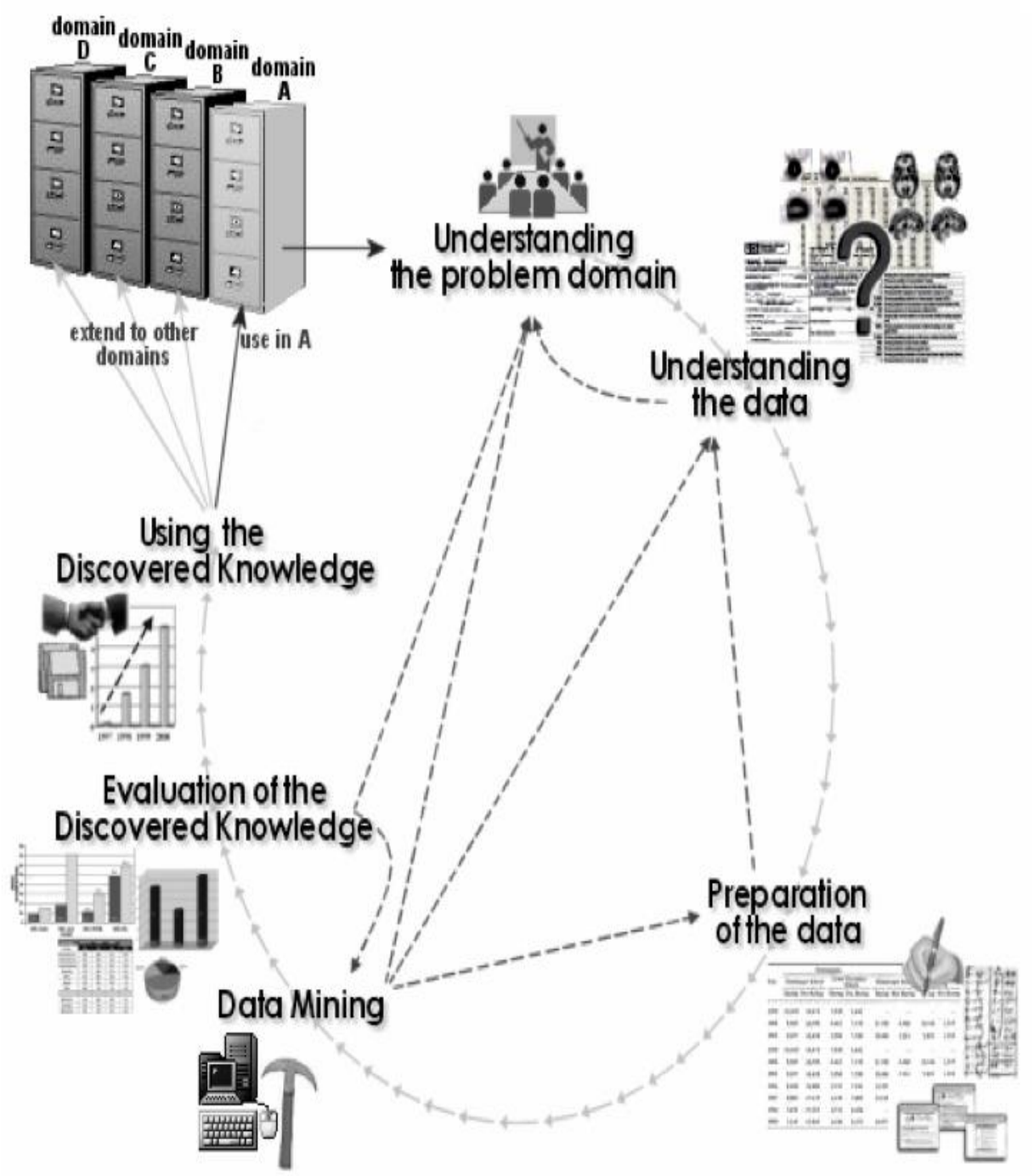

Figure 3.1 The Six Step Cios et al. (2005) Process Model: Hybrid Knowledge Discovery Process for Data Mining

In *Understanding the Problem Domain* step, the researcher works closely with domain experts to define the problem and determine the research goals, identifies key people, and learns about current solutions to the problem. On the basis of the insights gained from this phase, a description of the problem including its restrictions is done. To clearly identify, understand, and analyze the business problems, the primary (observation and interview), and secondary (data analysis) data collection methods are employed. Interview is employed to define features selection with the domain experts

while observation is conducted to understand some complex business processes. The research goals then need to be translated into the data mining goals, and include initial selection of the data mining tools.

Understanding the Data step includes collection of sample data, and deciding which data will be needed including its format and size. If background knowledge does exist some attributes may be ranked as more important. Next, we need to verify usefulness of the data in respect to the data mining goals. Data needs to be checked for completeness, redundancy, missing values, plausibility of attribute values, etc.

Preparation of the Data is the key step upon which the success of the entire knowledge discovery process depends; it usually consumes about half of the entire research effort. In this step, which data is used as input for data mining tools; is decided. It may involve sampling of data, data cleaning like checking completeness of data records, removing or correcting for noise, etc. The cleaned data can be, further processed by feature selection and extraction algorithms (to reduce dimensionality), and by derivation of new attributes (say by discretization and/or normalization). The result would be new data records, meeting specific input requirements for the planned to be used data mining tools.

Data Mining is another key step in the knowledge discovery process. Although it is the data mining tools that discover new information, their application usually takes less time than data preparation. This step involves usage of the planned data mining tools and selection of the new ones. Data mining tools include many types of algorithms; this step involves the use of several data mining tools on data prepared so far. First, the training and testing procedures are designed and the data model is constructed using one of the chosen data mining tools; the generated data model is verified by using testing procedures.

In *Evaluation of the Discovered Knowledge* step includes understanding the results, checking whether the new information is novel and interesting, interpretation of the results by domain experts, and checking the impact of the discovered knowledge. Only the approved models are retained. The entire data mining process may be revisited to identify which alternative actions could have been taken to improve the results.

The final step is *Use of the Discovered Knowledge*. This step is entirely in the hands of the owner of the database/data warehouse. It consists of planning where and how the discovered knowledge will be used. Moreover, the application area in the current domain will be extended to other domains.

The researcher discusses in brief the tasks performed (specific to this study) under each of The Six Step Cios et al. [24] Process Model as follows.

3.1. Methods of Problem Domain Understanding

After conducting a critical survey of the existing market/trade related documents, it has been attempted to thoroughly understand what ECX Market Data Unit need to accomplish. Various data mining concepts, theories, tools and techniques are also explored so as to use them as a means of solving the problem. The main objectives of the ECX Market Data business unit are identified and tried to be mapped to data mining problem.

To understand the ECX Market Data business unit business processes: document analysis, observation and discussion with domain experts are used. Based on understanding of the ECX Market Data business unit services problem, the research problem (i.e. predicting price and sales volume) has been formulated and the objectives of the research (i.e. building price & sales volume predictive model) in relation with how the data mining techniques can address the problem are identified.

3.2. Methods of Data Understanding

Data Source

For any data mining task, the primary requirement is availability of data in usable format. In this study, the data source is the ECX's historical market data obtained from the ECX Data warehouse, which uses MS SQL Server 2008 R2 Data Base Management Studio.

The initial EMD data obtained from ECX Market Data business unit have 20 attributes. The EMD is obtained in Excel Format which was directly extracted from the ECX Market Data warehouse. This initial dataset is described and visualized using Microsoft Excel as well as Microsoft SQL Server 2008 R2 to examine the properties of the dataset in detail. Each attribute values are summarized with frequency distribution using Microsoft Excel 2010, so as to see the distribution of the data. Organizing the attribute using frequency distribution has helped to easily detect missing and error values.

Depending on the task at hand, it may be the case that to focus on subsets of the available attributes. Attributes from EMD dataset is selected based on relevance to the data mining objectives. Moreover,

since attributes with too much missing values, attributes with single value and attributes with unique values are not relevant for data mining purpose they are not considered in the dataset.

3.3. Methods of Data Preparation

The EMD Data is easily pre-processed including (data reduction, integration and transformation) without any problem using in MS Excel 2010 and WEKA 3.6.12 data mining tool. Finally, the pre-processed EMD dataset is converted to CSV and then to ARFF for further exploration and experimentation on WEKA 3.6.12 data mining tool.

Description of the Initial EMD Dataset

For the purpose of this study, the data is obtained directly from ECX historical market data warehouse. The data was first in Microsoft Excel format and it is converted to ARFF file format for further preparation and ready for as input to WEKA 3.6.12 data mining tool. This initial dataset is described and visualized using Microsoft Excel as well as Microsoft SQL Server 2008 R2 to examine the properties of the dataset in detail. Moreover, simple statistical analysis is performed to verify the quality of the dataset such as: missing values, error values and outliers to obtain the maximal level information regarding the data mining questions.

Table 3.1 High-Level Description of the Initial EMD Dataset

Commodities	Duration	Total No. of Attributes	Total No. of Instances	Data File Size
Coffee **Pea Beans** **Sesame**	January, 2010 – March 2015	20	77,045	12.4 MB

As shown in Table 3.1. The initial data are collected from ECX market data (for the three major agricultural commodities traded at ECX: Coffee, Pea Beans and Sesame) for more than five years, with an Excel file data size of 12.4 MB. The total numbers of attributes were 20, but there are many attributes which are not relevant for this particular research (which is discussed in detail on the next chapter, on: “Selection and Statistical Summary of Attribute” part).

Berry & Linoff [9] have pointed out that data mining expects data to be in a particular format. First, all the data should be in a single table. Second, each row should correspond to an entity that is

relevant to the business. Third, columns with a single value should be ignored. Fourth, columns with a different value for every column should be ignored. Last, for predictive modeling, the target column should be identified and all synonymous columns removed.

According to Kantardzic [10], the two standard tasks are associated with producing a reduced set of features. These are based on the knowledge of the application domain and the goals of the mining effort. The human analyst may select a subset of the features found in the initial dataset. The other one is feature composition by transformations of data that can have a surprisingly strong impact on the results of data mining methods. Berry & Linoff [9] have added on these ideas that it is impossible to do a good job of selecting input variables without knowledge of the business problem being addressed. This is true even when using data mining tools that claim the ability to accept all the data and figure out automatically which fields are important.

As a result of EMD business process analysis relevant attributes are selected in the subsequent sections. Moreover, data mining techniques such as data integration and transformation are used to summarize the low level too many distinct values in to high level values. .

The detail is further presented on "*Exploration and Preparation of EMD Datasets*" section of this document.

3.4. Method of Data Mining (Predictive Model Building)

Data Mining Tools and Algorithms

For conducting this research the WEKA Version 3.6.12 data mining software is chosen. WEKA stands for Waikato Environment for Knowledge Learning. It was developed by the University of Waikato, New Zealand. WEKA supports many data mining tasks such as data re-processing, classification, clustering, regression and feature selection to name a few. WEKA is chosen because of its free-open-source widespread application in different data mining researches [4, 8, 14] and also the familiarity of the researcher with the software/tool.

A large different number of classifiers are used in Weka. However, choosing an algorithm that best suits for building a predictive model is very challenging. Thus in this study, for creating WAP and

Total Volume predictive model, the following eight WEKA 3.6.12 classifiers algorithms such as: GaussianProcesses, LinearRegression, MultilayerPerceptron, SMOreg, DecisionTable, M5Rules, M5P and REPTree; which are applicable to the datasets specific to this study are used.

The details on the DM Tasks, Algorithms and Techniques used in building, evaluating and using the predictive model is discussed on the "*Experimentation and Result Analysis*" section of this document.

CHAPTER FOUR

EXPLORATION AND DATASET PREPARATION

This chapter deals with understanding the problem domain as well as investigating the initial EMD dataset and attributes used during these processes in detail. Moreover, pre-processing of the initial EMD data obtained from ECX are performed.

4.1. Problem Domain Understanding

The Context of ECX Market Data (EMD)

As noted by [64, 65, 66, 67], some people find data mining techniques interesting from a technical perspective; however, for most people, the techniques are interesting as means to an end. They further said that the techniques do not exist in a vacuum; data mining techniques should be applied in the business context.

In this study, to build the proposed Price and Sales Volume Predictive Model, we need to really understand and define the context of EMD. Thus, as EMD is obtained from historical market data obtained from ECX, we need to first know about ECX's business processes.

About ECX (Ethiopia Commodity Exchange)

The ***Ethiopia Commodity Exchange (ECX)*** is a commodities exchange established April 2008 in Ethiopia. In Proclamation 2007-551, which created the ECX, its stated objective was "to ensure the development of an efficient modern trading system" that would "protect the rights and benefits of sellers, buyers, intermediaries, and the general public" [68].

ECX is a marketplace, where buyers and sellers come together to trade, assured of quality, delivery and payment. The vision of ECX is to transform the Ethiopian economy by becoming a global commodity market of choice. ECX's mission is to connect all buyers and sellers in an efficient, reliable, and transparent market by harnessing innovation and technology, and based on continuous learning, fairness, and commitment to excellence. The first of its kind in Ethiopia, ECX is a national multi-commodity exchange that provides low-cost, secure marketplace services to benefit all agricultural market stakeholders and invites industry professionals to seek membership enabling them to participate in trading. ECX promotes and enables the following market services [68, 69]:

- *Market Integrity*: by guaranteeing the product grade and quantity and operating a system of daily clearing and settling of contracts.
- *Market Efficiency*: by operating a trading system where buyers and sellers can coordinate in a seamless way on the basis of standardized contracts.
- *Market Transparency*: by disseminating market information in real time to all market players.
- *Risk Management*: by offering contracts for future delivery, providing sellers and buyers a way to hedge against price risk.

By the time of this paper is conducted, ECX provides the following main services: *Membership and Compliance Monitoring , Quality Certification and Warehousing Services, Central Depository, Trading Operations, Clearing & Settlement and Market Data*.

Coffee, Sesame, Haricot Beans (Pea Beanss), Maize and Wheat are the major agricultural commodities which are traded/exchanged via ECX Trading Platform [68, 69, 70].

About EMD (ECX Market Data)

ECX Market Data is a data obtained from trading activities of the exchange's data warehouse. The main objective of EMD is to provide market information in a *neutral***,** *timely*, and *accurate* manner to all stakeholders [69].

Why Market Data?

- *Power and Participation*: Sellers and buyers are better able to negotiate in a given transparency
- *Quality*: Sellers are able to get premium for value added to the product
- *Market Access*: Farmers and other market actors are not restricted or captive to local market, can access national or international market
- *Planning*: Farmers can use prices for planting decisions, sellers/buyers also decide on when to sell/buy
- *Standards*: Equal playing field for farmers and other market actors by providing market standard trading principles
- *Market Integrity*: Creates an efficient and transparent market for ALL

The EMD Data Flow at ECX is shown in Figure 4.1 [69] :

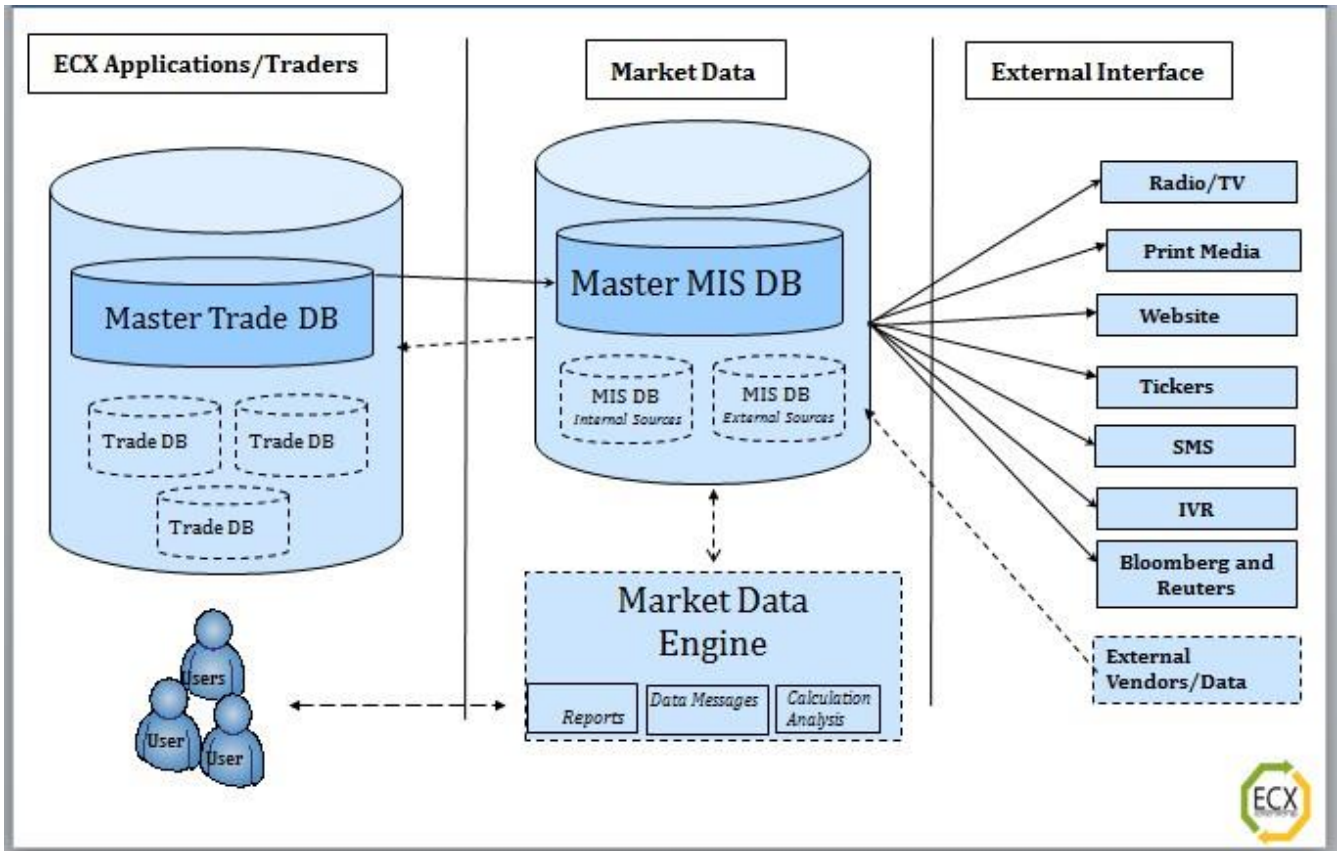

Figure 4.1 ECX Market Data Flow Diagram

EMD Data Dissemination Tools at ECX

The components of the ECX Market Information System are [69]:

- *Electronic Price Tickers (EPT)*: Transmit daily market prices from ECX trading floor to 80 strategically selected regional market sites across Ethiopia. Electronic Price Tickers will disseminate data in "real time", < 4 seconds using SMS/broadband technology. It is Fully integrated with ECX Trading systems, so that a real-time information will be provided during trading hours for each commodity and summary information on all commodities will be disseminated after trading hours
- *Mass media (Radio, TV, Print)*: Users can get EMD information from TV, Radio and print media on regular basis
- *The website*: The website content at *www.ecx.com.et* will provide real-time and detailed market information by commodity grade, warehouse location and production year. The

website will also provide other value added market information, such as: Archive of daily market information, Commodity profiles and associated research, Analysis and graphs, Stock Positions and Index Calculation

- *Mobile Phone Short Messaging Service (SMS)*: Available as long as there is a phone reception and provides real time price of ECX traded commodities. To use SMS, users need to text entry codes & key words and send to 934**.**
- *Interactive Voice Response (IVR) Service*: It is automated system used for dissemination of immediate update of market information in the local language of the recipient. It allows users to access data via the telephone network using their telephone set. Market players interact with the system by pressing digits on their telephone keypad. To use the IVR, users only need to dial **929** and follow on the prompt instructions to get the desired information.
- *Information Center*: Different reports are generated and disseminated to trading members, regulators (Ethiopia Commodity Exchange Authority/ECEA and Ministry of Trade/MOT), electronic and print media, various regional offices and international data vendors and others through printed paper, email or fax.

Market Analysis at ECX

Based on ECX Market Data, there are market analyses performed by EMD experts at ECX, such as [69]:

- *Correlation Analysis*: ECX vs. International /Reference Markets (Arabica 'C', ICE New York, ICO Colombia Mild's, ICO Brazilian Naturals, etc…)
- *Periodical Reports*: Weekly Market Data Reports, Monthly Market Outlook Reports. Semi-annual and Annual market outlook reports
- *In-depth Analysis*: Report for ECX/Regulatory Needs, based on Commodity types, prices, volume of trades, trade behaviors etc…
- *Valuation Analysis*: used for current market valuation by financial institutions such as for inventory financing

One of the target of this study is to help the EMD data experts to do more analysis and decisions by fully utilizing the proposed future market status predictive DM Model.

4.2. Data Understanding

4.2.1. Selection and Statistical Summary of Attributes

There are 20 attributes on the initial EMD extracted from the ECX's trade database archive.

These are: "Year", "Month", "Trade Date", "Commodity", "Name", "Symbol", "Session Name", "Origin", "Grade", "Close Price", "High", "Low", "Lot", "Feresula/Quintal", "Kgs","Ton", "Total Value", "Warehouse Code", "Production Year" and "Opening Price".

Attribute "Year" and "Month" are used to represent the commodity trading year and month which are the subjects of this study as the proposed predictive model takes a specific period of previous year/s data and forecasts for some specific future time/year market status (in value/volume). While, attribute "Trade Date" has a unique value and it is not relevant to this particular problem.

Attribute "Commodity" is used to identify the commodity type and has got an importance for this study, while attributes "Name", "Symbol", "Session Name", "Origin", "Grade", "Close Price", "High", "Low", "Warehouse Code", "Production Year" and "Opening Price" are reduced as they do not have significant importance for this particular study.

The attribute "Total Volume", is the total traded amount/volume (in Tons) obtained from a commodity trade, which is the concern of this study, and it's value is obtained from the "Feresula/Quintal", "Kgs","Ton" and "Total Value" attributes. The other crucial attribute is "Total Value" is the total amount (in Ethiopian Birr) obtained from a commodity trade.

Hence out of the 20 attributes obtained from the initial EMD, only five relevant attributes are selected. These are "Year", "Month", "Commodity", "Feresula/Quintal" and "Total Value".
The statistical summary and preprocessing of the selected attributes are presented in the subsequent sections. (*Moreover, See on Annex 1 for full description of some of the main EMD Attributes*)

4.2.2. Duration Attributes

The two major attributes that are related to duration are: "Year" and "Month".
Year & Month: These attributes indicate the trading year & month of a commodity traded at the ECX trading platform. "Year" contains a numeric values of: 2010, 2011, 2012, 2013, 2014 and 2015;

while “Month” contains a string values of: January, February, March, April, May, June, July, August, September, October, November, and December.

Trades made of the different years (January 01, 2010 to April 17, 2015) for the three commodities (Coffee, Pea Beans, and Sesame) obtained from the initial EMD are summarized on table 4.1 below:

Table 4.1 Statistical Summary of Year & Month Attributes of the Initial EMD Dataset

Year	Month	Total No. of Instances	Frequency (%)
2010	Jan 01 – Dec 31	11,494	14.92
2011	Jan 03 – Dec 30	13,360	17.34
2012	Jan 02 – Dec 31	15,139	19.65
2013	Jan 01 – Dec 31	15,140	19.65
2014	Jan 01 – Dec 31	15,978	20.74
2015	Jan 01 – Apr 17	5,934	7.70
	Missing	0	0%
	Invalid	0	0%
	Total	77,045	100%

As shown in table 4.1, 14.92% of the trades are made on year 2010, 17.34% of the trades are made on year 2011, 19.65% of the trades are made on year 2012, 19.65% of the trades are made on year 2013, 20.74% of the trades are made on year 2014, and 7.70% of the trades are made on year 2015. Fortunately, the EMD dataset have 0% Missing & Invalid values. Thus the training module is mainly focuses on the year 2012, 2013 and 2014 data to avoid the imbalance occurrence of instants.

4.2.3. Commodity Attribute

Commodity: This attribute indicates the type of commodity traded at the ECX trading platform. By the time this study was conducted the attribute “Commodity” contains values of: *Coffee*, *Pea Beans*, *Sesame*, *Green Mung Beans*, *Maize* and *Wheat*. The statistical summary details of this attribute values obtained from the initial EMD is presented on Table 4.2:

Table 4.2 Statistical Summary of Commodity Attribute of the Initial EMD Dataset

Year: 2010, Month: Jan 01 – Dec 31	Coffee	Pea Bean	Sesame	Maize	Wheat	Green Mung Beans	Total
No. of Instances	10,676	210	583	24	1	0	11,494
Frequency (%)	13.86	0.27	0.76	0.03	0.00	0	14.92%
Year: 2011, Month: Jan 03 – Dec 30	Coffee	Pea Bean	Sesame	Maize	Wheat	Green Mung Beans	Total
No. of Instances	9,311	951	3,077	21	0	0	13,360
Frequency (%)	12.06	1.23	3.99	0.03	0	0	17.34%
Year: 2012, Month: Jan 02 – Dec 31	Coffee	Pea Bean	Sesame	Maize	Wheat	Green Mung Beans	Total
No. of Instances	10,416	1,571	3,141	11	0	0	15,139
Frequency (%)	13.52	2.04	4.08	0.01	0	0	19.65%
Year: 2013, Month: Jan 01 – Dec 31	Coffee	Pea Bean	Sesame	Maize	Wheat	Green Mung Beans	Total
No. of Instances	10,737	1,458	2,943	0	2	0	15,140
Frequency (%)	13.95	1.89	3.82	0	0.00	0	19.65%
Year: 2014, Month: Jan 01 – Dec 31	Coffee	Pea Bean	Sesame	Maize	Wheat	Green Mung Beans	Total
No. of Instances	10,969	1,167	3,835	0	2	5	15,978
Frequency (%)	14.24	1.51	4.98	0	0.00	0.01	20.74 %
Year: 2015, Month: Jan 01 – Apr 17	Coffee	Pea Bean	Sesame	Maize	Wheat	Green Mung Beans	Total
No. of Instances	3,825	357	1,752	0	0	0	5,934
Frequency (%)	4.96	0.46	0.27	0	0	0	7.70 %
No. Missing Values	0	0	0	0	0	0	0
No. Invalid Values	0	0	0	0	0	0	0
Total No. of Instances	55,934	5,714	15,331	56	5	5	77,045
Total Frequency (%)	72.6 %	7.42 %	19.9 %	0.07 %	0%	0.01 %	100 %

As shown in Table 4.2, there is no missing and invalid values for Commodity attribute. About 72.6% of the commodities are Coffee, 19.9 % are Sesame, 7.42 % are Pea Beans, 0.07 % is Maize, 0.01 % is Green Mung Beans and almost 0% is Wheat. Hence, the first three (Coffee, Sesame and Pea Beans) are selected for this particular study.

4.2.4. Price Related Attributes

Total Value: This attribute indicates the total amount (in Ethiopian Birr) obtained from a commodity trade in a specific period of time at the ECX trading platform. After pre-processing the EMD, Total Value represents the Sum of Total Values of each commodity per month. The statistical summary details of Total Value attribute values obtained from the initial EMD is shown on Table 4.3:

WAP: Weighted Average Price (WAP) is another important price related attribute for this study. In the context of EMD, WAP is described per Feresula (for Coffee) or Quintal (for other ECX Commodities). Thus, WAP attribute is a derived attribute obtained from the Total Value & Feresula/Quintal attributes. How WAP is calculated? How WAP data is integrated & transformed in the EMD dataset is discussed in detail in the "Data Transformation" part of this document.

Table 4.3 Statistical Summary of "Total Value" Attribute of the Initial EMD Dataset

Year: 2010, Month: Jan 01 – Dec 31	**Coffee**	**Pea Bean**	**Sesame**	**Others**	**Total**
No. of Instances	10,676	210	582	25	11,494
Frequency (%)	13.86	0.27	0.76	0.03	14.92%
Year: 2011, Month: Jan 03 – Dec 30	**Coffee**	**Pea Bean**	**Sesame**	**Others**	**Total**
No. of Instances	9,311	951	3,077	21	13,360
Frequency (%)	12.06	1.23	3.99	0.03	17.34%
Year: 2012, Month: Jan 02 – Dec 31	**Coffee**	**Pea Bean**	**Sesame**	**Others**	**Total**
No. of Instances	10,416	1,571	3,141	11	15,139
Frequency (%)	13.52	2.04	4.08	0.01	19.65%

Year: 2013, Month: Jan 01 – Dec 31	Coffee	Pea Bean	Sesame	Others	Total
No. of Instances	10,737	1,458	2,943	2	15,140
Frequency (%)	13.95	1.89	3.82	0.00	19.65%
Year: 2014, Month: Jan 01 – Dec 31	Coffee	Pea Bean	Sesame	Others	Total
No. of Instances	10,969	1,167	3,835	7	15,978
Frequency (%)	14.24	1.51	4.98	0.01	20.74 %
Year: 2015, Month: Jan 01 – Apr 17	Coffee	Pea Bean	Sesame	Others	Total
No. of Instances	3,825	357	1,752	0	5,934
Frequency (%)	4.96	0.46	0.27	0	7.70 %
Total Missing Values	0	0	0	0	0 %
Total Invalid Values	0	0	1	0	0 %
Total No. of Instances	55,934	5,714	15,331	66	77,045
Total Frequency (%)	72.6 %	7.42 %	19.9 %	0.08	100 %

As shown in Table 4.3 Statistical Summary of "Total Value" Attribute of the Initial EMD Dataset

, there is no missing value but there is 1 invalid value for Total Value attribute. The invalid value found was on the Year=2010, Month=December, Commodity=Sesame with Total Value=0, and this invalid value is adjusted manually (by calculating the exact Total Value).

4.2.5. Volume Related Attributes

Total Volume: The attribute represents the total traded amount/volume (in Tons) obtained from a commodity trade, which is the concern of this study, and its value is directly obtained from the values of "Ton". After pre-processing the EMD, Total Volume represents the Sum of Total Volumes of each Commodity per Month with in a Year from January to December. How Total Volume is calculated? How Total Volume data integrated & transformed in the EMD dataset will be discussed in detail in the "Data Integration & Transformation" part of this document. The statistical summary details of Ton attribute values obtained from the initial EMD is summarized on Table 4.4:

Table 4.4 Statistical Summary of "Ton" Attribute of the Initial EMD Dataset

Year: 2010, Month: Jan 01 – Dec 31	Coffee	Pea Bean	Sesame	Others	Total
No. of Instances	10,676	210	582	25	11,494
Frequency (%)	13.86	0.27	0.76	0.03	14.92%
Year: 2011, Month: Jan 03 – Dec 30	**Coffee**	**Pea Bean**	**Sesame**		**Total**
No. of Instances	9,311	951	3,077	21	13,360
Frequency (%)	12.06	1.23	3.99	0.03	17.34%
Year: 2012, Month: Jan 02 – Dec 31	**Coffee**	**Pea Bean**	**Sesame**		**Total**
No. of Instances	10,416	1,571	3,141	11	15,139
Frequency (%)	13.52	2.04	4.08	0.01	19.65%
Year: 2013, Month: Jan 01 – Dec 31	**Coffee**	**Pea Bean**	**Sesame**		**Total**
No. of Instances	10,737	1,458	2,943	2	15,140
Frequency (%)	13.95	1.89	3.82	0.00	19.65%
Year: 2014, Month: Jan 01 – Dec 31	**Coffee**	**Pea Bean**	**Sesame**		**Total**
No. of Instances	10,969	1,167	3,835	7	15,978
Frequency (%)	14.24	1.51	4.98	0.01	20.74 %
Year: 2015, Month: Jan 01 – Apr 17	**Coffee**	**Pea Bean**	**Sesame**		**Total**
No. of Instances	3,825	357	1,752	0	5,934
Frequency (%)	4.96	0.46	0.27	0	7.70 %
Total Missing Values	0	0	0	0	0 %
Total Invalid Values	0	0	0	0	0 %
Total No. of Instances	55,934	5,714	15,331	66	77,045
Total Frequency (%)	72.6 %	7.42 %	19.9 %	0.08	100 %

As shown in Table 4.4, there is no missing and invalid values for Ton attribute.

4.3. Data Preparation

4.3.1. Data Cleaning

In addition to its heterogeneous & distributed nature of data, real world data is low in quality. Some data have problems on their own that needs to be cleaned: ***Outliers*** (misleading data that do not fit to most of the data/facts), ***Missing data*** (attributes values might be absent which needs to be replaced with estimates), ***Irrelevant data*** (attributes in the database that might not be of interest to the DM task being developed) and ***Noisy data*** (attribute values that might be invalid or incorrect. e.g. typographical errors, Inconsistent data, duplicate data, etc.) [15, 23, 26, 27].

In real world application data are rarely complete, but, fortunately, the EMD dataset is free from missing values from the beginning, and only 1 error found on the Total Value attribute. The invalid value found was on the Year=2010, Month=December, Commodity=Sesame with Total Value=0, and this invalid value is adjusted manually (by calculating the exact Total Value i.e. Total Value= WAP * Feresula or Quintal) for that specific tuple. There was also an imbalance of instances problem on the initial EMD dataset. As it was discussed on the EMD data distribution analysis part, the imbalance was mainly observed on the Year and Commodity attributes. These imbalance problem was resolved by reducing those attribute values of data which have very few or insignificant number of instances. Thus, Year = {2010} and Commodities = {Maize, Wheat and Green Mung Beans} attribute values are reduced. Such data are handled separately before integration, and hence will not be considered in the final dataset.

The main reason behind the initial EMD is clean (free from missing values & error) is that, because it was directly obtained from the ECX Market Data Warehouse. And this is one of the main advantages of using data warehouse as part of data mining. Preprocessing is very important although often considered too mundane to be taken seriously [2, 3, 28]. Even though using Data Warehouse for Data Mining handles issues related to Data Cleaning, further preprocessing may also be needed after the data warehouse phase, such as Data Integration, Reduction and Transformation. For this particular study, Data Transformation is done which is discussed in the subsequent sections.

4.3.2. Data Transformation

As Han & Kamber [11] have stated, data mining often requires data integration or the merging of data from multiple data sources. Data reduction may be needed to transform very high dimensional data to a lower dimensional data. Moreover, the data may also need to be transformed into forms appropriate for mining.

Data Transformation is a function that maps the entire set of values of a given attribute to a new set of replacement values such that each old value can be identified with one of the new values. Data is mainly transformed into forms that are appropriate for data mining to improve the performance of the proposed data mining model. The two major methods for data transformation are: *Normalization* (Scaled to fall within a smaller, specified range of values) and *Discretization* (Reduce data size by dividing the range of a continuous attribute into intervals. Interval labels can then be used to replace actual data values)

Here are some of the data transformations implemented on the cleansed EMD dataset. The data integration and transformation are made based on the need for intended predictive model and market data made on the ECX Market Data Expertise. First let's see a sample of the Monthly Market Data Report form of ECX is shown in Table 4.5 [70]:

Table 4.5 ECX Traded Commodity in Volume, Value, and Market Share Performance: October 2014

Commodity	Category	Volume in Ton	Value in ETB	WAP/Feresula or Quintal*	Market Share	
					Volume	Value
Coffee	Export Unwashed	8, 09	627,652,597	1,284	**55.76%**	**71.60%**
	Export Washed	83	6,504,360	1,330		
	Local Unwashed	2,121	106,910,330	857		
	Local Washed	226	15,832,905	1,192		
	Coffee Total	**10,739**	**756,900,192**	**1,198**		
Sesame	Humera Sesame	3,589	144,846,419	4,036	**39.57%**	**27.35%**
	Reddish Sesame	***15***	425,820	***2820***		
	Wollega Sesame	4,016	143,813,380	***3581***		
	Sesame Total	***76 20***	**289,085,619**	3,794		
Pea Bean	Round White Pea	***777***	9,359,420	1,205	**4.67%**	**1.05%**
	Flat White Pea	***123***	1,719,700	1,404		
	Pea Bean Total	***899***	**11,079,120**	1232		
	Grand Total	**19,259**	**1,057,064,931**	**1,475**	**100.00%**	**100.00%**

- ***WAP is weighted Average Price (*For coffee, price is in Birr/Feresula and for other commodity price is in Birr/Quintal)**

Based on the analysis made on results of Table 4.5 and from behavior of the proposed predictive model, there have to be some new set of attributes and attribute values that need to be transformed (normalized). These are: *Total Volume* and *WAP/Feresula or Quintal**.

Normalizing "Volume in Ton (Total Volume)" Attribute Values

This attribute is used to represent the total traded amount/volume (in Tons) obtained from a commodity trade, which is the concern of this study, and its value is directly obtained from the values of "Ton" attribute.

Thus, the newly transformed/normalized "Total Volume" attribute value represent the Sum of "Ton" attribute values of each Commodity per a specific period of trade duration (in our case: per Month, for Commodities: Coffee, Pea Beans and Sesame).

Normalizing "WAP/Feresula or Quintal" or "WAP" Attribute Values

WAP (Weighted Average Price per Feresula or Quintal) is another important price related attribute for this study. In the context of EMD, WAP is measured/indexed per Feresula (for Coffee) and Quintal (for other ECX Commodities). Thus, the newly transformed/normalized "WAP" attribute values are obtained from the "Total Value" & "Feresula/Quintal" attributes values. In such a way that:

- ✓ **WAP** (Month X) = Total Value in ETB (Month X) / Total Feresula or Quintal (Month X)

  Where,

  - **Total Value** in ETB (Month X) is the SUM of [Total Value (Beginning Trade Date Of Month X to End Trade Date Of Month X)], and
  - **Total Feresula or Quintal** (Month X) is the SUM of [Feresula or Quintal (Beginning Trade Date Of Month X to End Trade Date Of Month X)]

4.3.3. Description of the Pre-Processed Dataset

The EMD dataset, after passing through all above discussed pre-processing steps (Data Cleaning & Transformation), the final summary of constructed dataset used for this particular project experiments are presented in table 4.6.

Table 4.6 Summary of Final EMD Data Set

EMD Data Set 1 (For Building WAP Predictive Model)			**EMD Data Set 2 (For Building Sales Volume Predictive Model)**	
Category	**Description**		**Category**	**Description**
Duration	Jan 01, 2011-April 17, 2015		Duration	Jan 01, 2011- April 17, 2015
Attributes	Year, Month, Commodity & WAP		Attributes	Year, Month, Commodity, WAP & Total Volume
No. Of Selected Attributes	4		No. Of Selected Attributes	5
Total No. of Instances	65,551		Total No. of Instances	65, 551
Size of Data	3.5 MB		Size of Data	3.7 MB

As it is shown on Table 4.6, there are two separate datasets (EMD Dataset 1 and EMD Dataset 2), used for the experimentation of (WAP Predictive Model 1 and Total Sales Volume Predictive Model 2) respectively. Once the datasets are pre-processed and cleaned then the EMD data is ready for input to WEKA 3.6.12 data mining tool which is handled and discussed in detail on the next chapter.

Moreover, the detail of these datasets and their usage on the experimentation and result analysis is illustrated subsequently in the next chapter of this document.

CHAPTER FIVE

EXPERIMENTATION AND RESULT ANALYSIS

In this chapter, the experimental design, the procedures followed to build the predictive model and detailed results analysis are discussed. Moreover, an overview of the predictive model prototype development and usage is presented.

In this study an attempt is made to explore the application of DM approaches on the available EMD dataset to predict future market status of the commodities traded at the ECX. Experiments have been carried out to identify classification model and to extract relevant actionable predictive model. The purpose of experiments in classification is to find the best performing predictive model that is able to predict the WAP and Total Volume (Sales Volume) values of the selected commodities (Coffee, Pea Beans and Sesame) trades taking selected attributes variables as inputs.

5.1. Experimental Design

The full experimentation is done using WEKA 3.6.12 DM tool. Waikato Environment for Knowledge Learning (WEKA) is a computer program that was developed by the student of the University of Waikato in New Zealand for the purpose of identifying information from raw data gathered from agricultural domains [4, 17]. Data preprocessing, classification, clustering, association, regression and feature selection these standard data mining tasks are supported by Weka. It is an open source application which is freely available. In Weka datasets should be formatted to the ARFF format. The Weka Explorer will use these automatically if it does not recognize a given file as an ARFF file. Classify tab in Weka Explorer is used for the classification purpose.

Steps to apply classification techniques on data set and get result in Weka [27]**:**

Step 1**:** Take the input dataset.

Step 2**:** Apply the classifier algorithm on the whole data set.

Step 3: Note the accuracy given by it and time required for execution.

Step 4**:** Repeat step 1 to 3 for different classification algorithms on different datasets.

Step 5: Compare the different accuracy provided by the dataset with different classification algorithms and identify the significant classification algorithm for particular dataset.

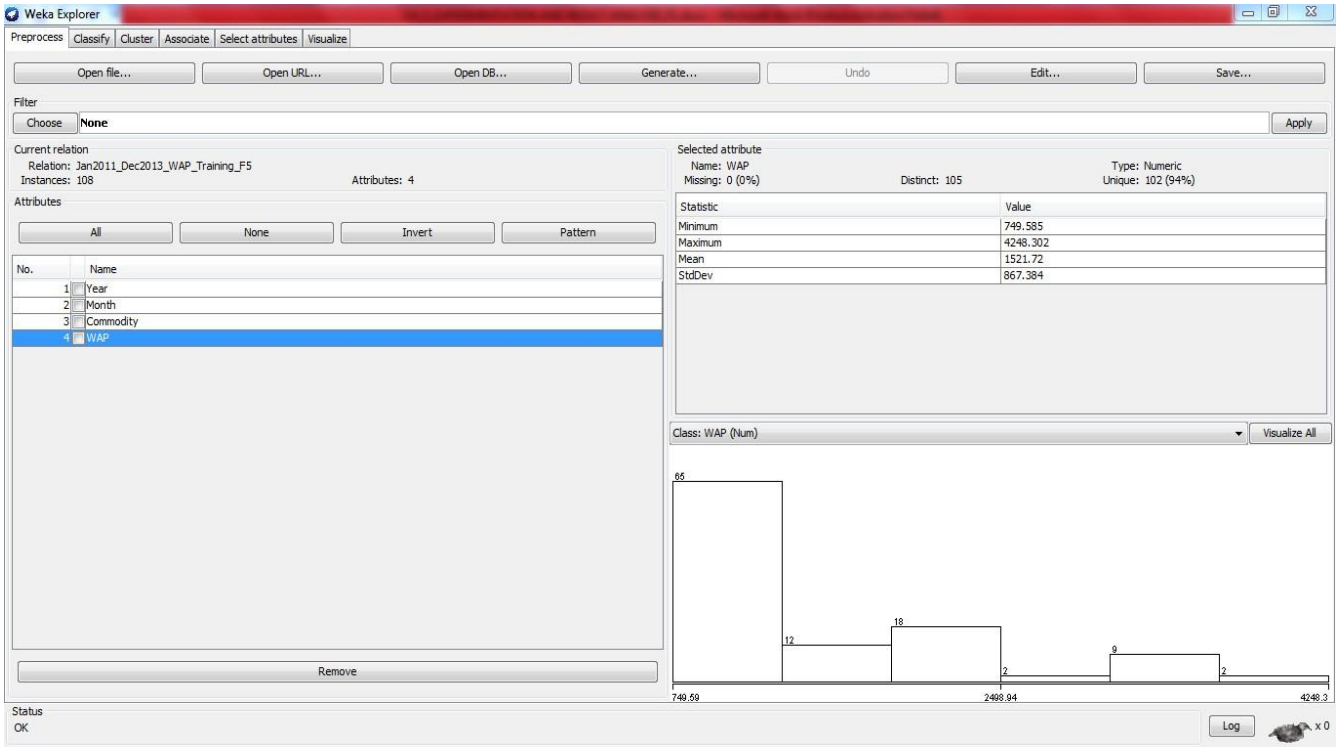

Figure 5.1 WEKA 3.6.12 Preprocess Explorer Windows Sample (the case of: WAP_DataSet_1)

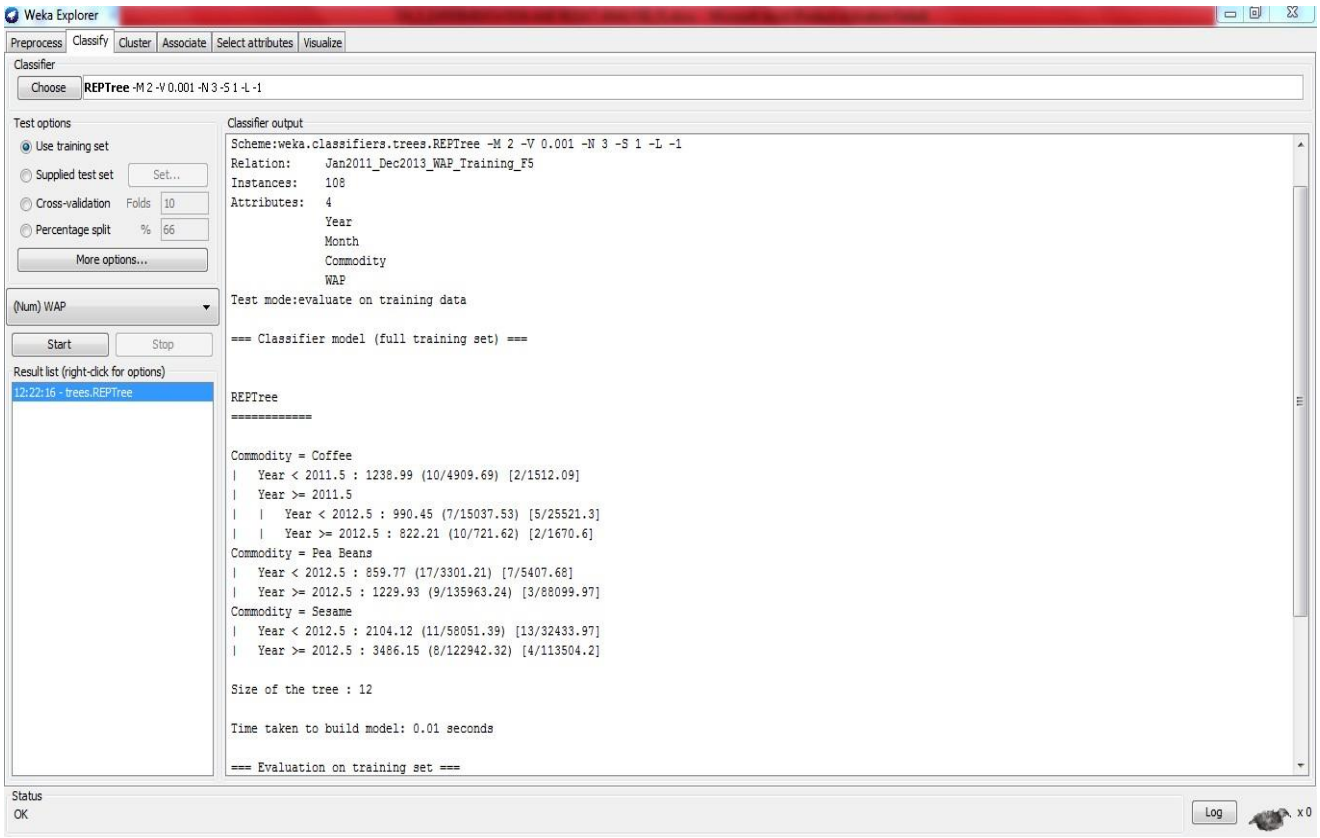

Figure 5.2 WEKA 3.6.12 Classify Explorer Windows Sample (Trees.REPTree Classifier on WAP_DataSet_1)

A large different number of classifiers are used in Weka. However, choosing an algorithm that best suits for building a predictive model is very challenging. Thus in this study, for creating WAP and Total Volume predictive model, the following eight WEKA 3.6.12 classifiers algorithms which are applicable to the datasets specific to this study are used :

- Weka.Classifiers.Functions.GaussianProcesses
- Weka.Classifiers.Functions.LinearRegression
- Weka.Classifiers.Functions.MultilayerPerceptron
- Weka.Classifiers.Functions.SMOreg
- Weka.Classifiers.Rules.DecisionTable
- Waka.Classifiers.Ruls.M5Rules
- Weka.Classifiers.Trees.M5P and
- Weka.Classifiers.Trees.REPTree.

In order to train, test and validate the WEKA classifiers a total of 65,551 data instances on six varieties of datasets are used.

The experiments are conducted in a configuration on Windows 7 Home Edition Operating System having Intel Core i5 Processor CPU @ 1.7 GHz, 6 GB Internal Memory and 500 GB HDD. Experiments are conducted 3 times each (i.e. a total of 144 experiments = 3 Times * 6 Datasets * 8 Classifiers) and an average accuracy and time is recorded.

Description of the Selected WEKA Classifiers

The eight Weka Classifier used for the experimentation of this particular study are categorized in to three main categories: *Functions*, *Rules* and *Trees*. A brief description of these classifiers is presented as follows. (*Moreover, See on Annex 6 for the details about each of these Weka Classifiers*).

➢ Functions (Weka.Classifiers.Functions)

- **Gaussian Processes:** Implements Gaussian processes for regression without hyper-parameter-tuning. To make choosing an appropriate noise level easier, this implementation applies normalization/standardization to the target attribute as well (if normalization/standardization is turned on). Missing values are replaced by the global mean/mode. Nominal attributes are converted to binary ones,

- **Linear Regression:** Class for using linear regression for prediction. Uses the Akaike criterion for model selection, and is able to deal with weighted instances.
- **Multilayer Perceptron:** A Classifier that uses back-propagation to classify instances. This network can be built by hand, created by an algorithm or both. The network can also be monitored and modified during training time. The nodes in this network are all sigmoid (except for when the class is numeric in which case the output nodes become unthresholded linear units).
- **SMOreg:** Implements the support vector machine for regression. The parameters can be learned using various algorithms. The algorithm is selected by setting the RegOptimizer. The most popular algorithm (RegSMOImproved) is due to Shevade, Keerthi et al and this is the default RegOptimizer.

➢ **Rules (Weka.Classifiers.Rules)**

- **Decision Table:** Class for building and using a simple decision table majority classifier.
- **M5Rules:** Generates a decision list for regression problems using separate-and-conquer. In each iteration, it builds a model tree using M5 and makes the "best" leaf into a rule.

➢ **Tree (Weka.Classiiers.Trees)**

- **M5P:** Implements base routines for generating M5 Model trees and rules. The original algorithm M5 was invented by R. Quinlan and Yong Wang made improvements.
- **REPTree:** Fast decision tree learner. Builds a decision/regression tree using information gain/variance and prunes it using reduced-error pruning (with backfitting). Only sorts values for numeric attributes once. Missing values are dealt with by splitting the corresponding instances into pieces (i.e. as in C4.5).

5.2. Model Building

One of the major tasks of data mining research is modeling. As mentioned in the literature part, data mining involves predictive and descriptive models. Prediction is very similar to classification. The difference is that in prediction, the class is not a qualitative discrete attribute but a continuous one. The goal of prediction is to find the numerical value of the target attribute for unseen objects; this problem is also known as regression, and if the prediction deals with time series data, then it is often called forecasting. Regression analysis, decision trees, and neural nets DM techniques are mostly applied for building predictive models.

In this study an attempt is made to build a model that predicts future market/trade status of agricultural commodities (Coffee, Pea Beans and Sesame) exchanged/traded at ECX, by comparing various types of classification algorithms. In doing so, WAP Predictive Model and Total Volume Predictive Model is constructed. The detail on building these models is discussed below.

5.2.1. WAP Predictive Model Building

The aim of this (WAP Predictive Model) is to predict the expected WAP/Weighted Average Price-per-Feresula or Quintal of the next/future year/12 months, based on the previous year market data. Thus, this WAP Predictive model predicts each 12 months (January to December) WAP of Year X +1, given (Year = X, Month = [January to December] and Commodity = [Coffee, Pea Beans, Sesame].

While building this WAP predictive model, three datasets were used for experimentation. The experimentation is made using different types of WEKA 3.6.12 classifiers (classification algorithms). Finally, by comparing and analyzing the results of each experiment, the best dataset and classifier is chosen for building this WAP Predictive Model.

Datasets

Monthly data for Coffee, Pea Beans and Sesame from year 2011 to 2014 are used this purpose. The dataset are further categorized in to three datasets all having four attributes (Year, Month, Commodity and WAP). The details of each datasets including the Training Dataset and Test Dataset that are used for building the WAP Predictive Model is presented below. *(See on Annex 2, for the full WAP Data Set ARFF file)*

WAP DataSet_1: Total Number of Instances=144

WAP DataSet_1_TrainingDataSet: Total Number of Instances=108

@relation Jan2011_Dec2013_WAP_Training

@attribute Year numeric

@attribute Month

{January,February,March,April,May,June,July,August,September,October,November,December}

@attribute Commodity {Coffee,'Pea Beans',Sesame}

@attribute WAP numeric

WAP DataSet_1_TestDataSet_1: Total Number of Instances=36

@relation Jan2014_Dec2014_WAP_Test

@attribute Year numeric

@attribute Month

{January,February,March,April,May,June,July,August,September,October,November,December}

@attribute Commodity {Coffee,'Pea Beans',Sesame}

@attribute WAP numeric

WAP DataSet_2: Total Number of Instances=108

WAP DataSet_2_TrainingDataSet Total Number of Instances=72

@relation Jan2012_Dec2013_WAP_Training

@attribute Year numeric

@attribute Month

January,February,March,April,May,June,July,August,September,October,November,December}

@attribute Commodity {Coffee,'Pea Beans',Sesame}

@attribute WAP numeric

WAP DataSet_2_TestDataSet: Total Number of Instances=36

@relation Jan2014_Dec2014_WAP_Test

@attribute Year numeric

@attribute Month

{January,February,March,April,May,June,July,August,September,October,November,December}

@attribute Commodity {Coffee,'Pea Beans',Sesame}

@attribute WAP numeric

WAP DataSet_3: Total Number of Instances=72

WAP DataSet_3_TrainingDataSet: Total Number of Instances=36

@relation Jan2013_Dec2013_WAP_Training

@attribute Year numeric
@attribute Month
January,February,March,April,May,June,July,August,September,October,November,December}
@attribute Commodity {Coffee,'Pea Beans',Sesame}
@attribute WAP numeric

WAP DataSet_3_TestDataSet: Total Number of Instances=36

@relation Jan2014_Dec2014_WAP_Test

@attribute Year numeric

@attribute Month

{January,February,March,April,May,June,July,August,September,October,November,December}

@attribute Commodity {Coffee,'Pea Beans',Sesame}

@attribute WAP numeric

Procedures

Experimentation is made using the above datasets on selected WEKA 3.6.12 classifiers. The WEKA classifiers used for this WAP Predictive Model building experimentation purposes are: functions.GaussianProcesses, functions.LinearRegression, functions.MultilayerPerceptron and functions.SMOreg, rules.DecisionTable, rules.M5Rules, trees.M5P and trees.REPTree. All classifiers use the values of [Year, Month and Commodity] as *Predictor* values and [WAP] as the *Predicted* value.

Performance Evaluation

As most of the predictive models use, the statistical parameters used for the evaluation of the model are: the time taken to build model, the correlation coefficient and the root mean square percentage error (RMSPE) values. Thus, the selected model is the one with the maximum correlation coefficient and minimum RMSPE of the average outputs corresponding to the testing phase that takes an optimal time to build the model. The statistical summary is discussed in detail on “Results and Discussions” subsequent section of this document.

5.2.2. Total Sales Volume Predictive Model Building

The aim of this (Total Volume Predictive Model) is to predict the expected Total Volume/Total Volume trade amount in tons of the next/future year/12 months, based on the previous year market data. Thus, this Total Volume Predictive model predicts each 12 months (January to December) Total Volume of Year X +1, given (Year = X, Month = [January to December], Commodity = [Coffee, Pea Beans, Sesame] and monthly WAP of Year X. This Total Volume Predictive model depends on the previous WAP Predictive model, as it uses results of monthly WAP_Pridicted value of Year X + 1 for executing Total Volume Predictive model.

While building this total volume predictive model, three datasets were used for experimentation. The experimentation is made using different types of WEKA 3.6.12 classifiers (classification algorithms). Finally, by comparing and analyzing the results of each experiment, the best dataset and classifier is chosen for building this Total Volume Predictive Model.

Datasets

Monthly data for Coffee, Pea Beans and Sesame from year 2011 to 2014 are also used this purpose. However, here the dataset are further categorized in to three datasets all having five attributes (Year, Month, Commodity, WAP and Total Volume). The details of each datasets including the Training Dataset and Test Dataset that are used for building the Total Volume Predictive Model is presented below. *(See on Annex 3, for the full VOLUME Data Set ARRF file)*

Volume DataSet_1: Total Number of Instances=144

Volume DataSet_1_TrainingDataSet: Total Number of Instances=108

@relation Jan2011_Dec2013_VOLUME_Training

@attribute Year numeric
@attribute Month {January,February,March,April,May,June,July,August,September,October,November,December}
@attribute Commodity {Coffee,'Pea Beans',Sesame}
@attribute WAP numeric
@attribute 'Total Volume' numeric

Volume DataSet_1_TestDataSet_1: Total Number of Instances=36

@relation Jan2014_Dec2014_VOLUME_Test

@attribute Year numeric

@attribute Month
{January,February,March,April,May,June,July,August,September,October,November,December}
@attribute Commodity {Coffee,'Pea Beans',Sesame}
@attribute WAP numeric
@attribute 'Total Volume' numeric

Volume DataSet_2: Total Number of Instances=108

Volume DataSet_2_TrainingDataSet Total Number of Instances=72

@relation Jan2012_Dec2013_VOLUME_Training

@attribute Year numeric
@attribute Month
{January,February,March,April,May,June,July,August,September,October,November,December}
@attribute Commodity {Coffee,'Pea Beans',Sesame}
@attribute WAP numeric
@attribute 'Total Volume' numeric

Volume DataSet_2_TestDataSet: Total Number of Instances=36

@relation Jan2014_Dec2014_VOLUME_Test

@attribute Year numeric
@attribute Month
{January,February,March,April,May,June,July,August,September,October,November,December}
@attribute Commodity {Coffee,'Pea Beans',Sesame}
@attribute WAP numeric
@attribute 'Total Volume' numeric

Volume DataSet_3: Total Number of Instances=72

Volume DataSet_3_TrainingDataSet: Total Number of Instances=36

@relation Jan2013_Dec2013_VOLUME_Training

@attribute Year numeric
@attribute Month
{January,February,March,April,May,June,July,August,September,October,November,December}
@attribute Commodity {Coffee,'Pea Beans',Sesame}
@attribute WAP numeric
@attribute 'Total Volume' numeric

Volume DataSet_3_TestDataSet: Total Number of Instances=36

@relation Jan2014_Dec2014_VOLUME_Test

@attribute Year numeric
@attribute Month
{January,February,March,April,May,June,July,August,September,October,November,December}

@attribute Commodity {Coffee,'Pea Beans',Sesame}
@attribute WAP numeric
@attribute 'Total Volume' numeric

Procedures

Experimentation is made using the above three datasets on selected WEKA 3.6.12 classifiers. The WEKA classifiers used for this WAP Predictive Model building experimentation purposes are: functions.GaussianProcesses, functions.LinearRegression, functions.MultilayerPerceptron and functions.SMOreg, rules.DecisionTable, rules.M5Rules, trees.M5P and trees.REPTree. All classifiers use the values of [Year, Month, Commodity and WAP] as *Predictor* values and [Total Volume] as the *Predicted* value.

Performance Evaluation

As most of the predictive models use, the statistical parameters used for the evaluation of the model are: the time taken to build model, the correlation coefficient and the root mean square percentage error (RMSPE) values. Thus, the selected model is the one with the maximum correlation coefficient and minimum RMSPE of the average outputs corresponding to the testing phase that takes an optimal time to build the model. The statistical summary is discussed in detail on "Results and Discussions" subsequent section of this document.

5.3. Experimentation Results and Discussions

5.3.1. Model Evaluations and Comparisons

Most predictive model evaluation methods dealt with factors such as the accuracy and generality of the model [25, 26, 27]. The model evaluation process assesses the degree to which the model meets the business objectives and seeks to determine if there is some business reason why the model is deficient. It compares results with the evaluation criteria defined at the start of the project. Another option of evaluation is to test the model(s) on test applications in the real application if time and budget permits.

In this study, the statistical parameters used for the evaluation of the model are: the time taken to build model (Time), the correlation coefficient percentage (CCP) and the root mean square percentage error (RMSPE) values.

Time: is the time taken to build the model using WEKA 3.6.12 classifiers and is measured in seconds.

CCP: is the Correlation Coefficient Percentage value and is given by: **CCP =** ***r*** * 100%

The quantity ***r***, called *the linear correlation coefficient*, measures the strength and the direction of a linear relationship between two variables (***x*** and ***y***). The mathematical formula for computing ***r*** is:

$$r = \frac{n\sum xy - \left(\sum x\right)\left(\sum y\right)}{\sqrt{n\left(\sum x^2\right) - \left(\sum x\right)^2}\sqrt{n\left(\sum y^2\right) - \left(\sum y\right)^2}}$$

where, ***n*** is the number of pairs of data.

The value of r is such that $-1 \leq r \leq +1$. The + and – signs are used for positive linear correlations and negative linear correlations, respectively. A correlation greater than 0.8 is generally described as *strong*, whereas a correlation less than 0.5 is generally described as *weak*. These values can vary based upon the "type" of data being examined [54].

RMSPE: is the Root Mean Square Percentage Error

To compare the closeness of modeled/predicted outputs, to the observed outputs, the root mean square percentage error (RMSPE). The mathematical formula for computing RMSPE is given by [54]:

$$RMSPE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\frac{t - t^*}{t}\right)^2}.100\%$$

where **t** is the observed output; **t*** is the predicted output and **N** is the number of observations.

Since the above statistics is measures of error in estimation, the smaller the values of this measures, the closer is the modeled/predicted output to the target/observed output. Ideally, values of zero for each of these statistics would mean that the estimated output is the same as the target output. Normally, the desirability of having low or zero values of these statistics depends on how close the target is to the (unknown) true output.

Thus, the selected/best model is the one with the maximum *CCP* and minimum *RMSPE* of the average outputs corresponding to the testing phase that consumes an optimal *Time* to build the model.

5.3.1.1. Experimentation Results & Analysis: On WAP Predictive Model

Table 5.1 Summary of WEKA Classifiers Performance Evaluation on WAP_Dataset_1

	Classifier Functions				Classifier Rules		Classifier Trees	
	Gaussian Processes	**Linear Regression**	**Multilayer Perceptron**	**SMOreg**	**Decision Table**	**M5Rules**	**M5P**	**REPTree**
Time	0	0.02	0.31	0.1	0.04	0.32	0.06	0
CC	0.9339	0.9551	0.8982	0.9547	0.9775	0.9369	0.9541	0.9775
RMSPE	7.6571	5.6209	9.9361	9.0187	3.8914	6.0552	4.2111	3.8914

As it is shown in Table 5.1, the REPTree Classifier has the highest accuracy score on WAP_Dataset_1 with maximum *CCP* (97.75 %), minimum *RMSPE* (3.8914 %) and an optimal *Time* (almost 0 seconds) to build the WAP Predicive model.

Table 5.2 Summary of WEKA Classifiers Performance Evaluation on WAP_Dataset_2

	Classifier Functions				Classifier Rules		Classifier Trees	
	Gaussian Processes	**Linear Regression**	**Multilayer Perceptron**	**SMOreg**	**Decision Table**	**M5Rules**	**M5P**	**REPTree**
Time	0.72	0	0.14	0.01	0.02	0.05	0.04	0
CC	0.9326	0.9461	0.9173	0.9647	0.9775	0.9356	0.9526	0.9715
RMSPE	10.4936	5.0179	5.2886	5.2989	3.8914	5.7591	4.9456	4.3787

As it is shown in Table 5.2, the Decision Table Classifier has the highest accuracy score on WAP_Dataset_2 with maximum *CCP* (97.75 %), minimum *RMSPE* (3.8914 %) and an optimal *Time* (0.02 seconds) to build the WAP Predictive model.

Table 5.3 Summary of WEKA Classifiers Performance Evaluation on WAP_Dataset_3

	Classifier Functions				Classifier Rules		Classifier Trees	
	Gaussian Processes	**Linear Regression**	**Multilayer Perceptron**	**SMOreg**	**Decision Table**	**M5Rules**	**M5P**	**REPTree**
Time	0.09	0.01	0.07	0.01	0.01	0.01	0.02	0
CC	0.9354	0.9405	0.911	0.9227	0.9775	0.9397	0.9397	0.9775
RMSPE	6.0704	5.0778	5.9487	5.4472	3.8914	5.1002	5.1002	3.8914

As it is shown in Table 5.3, again the REPTree Classifier has the highest accuracy score on WAP_Dataset_3 with maximum *CCP* (97.75 %), minimum *RMSPE* (3.8914 %) and an optimal *Time* (almost 0 seconds) to build the WAP Predictive model.

5.3.1.2. Experimentation Results & Analysis: On Sales Volume Predictive Model

Table 5.4 Summary of WEKA Classifiers Performance Evaluation on VOLUME_Dataset_1

	Classifier Functions				Classifier Rules		Classifier Trees	
	Gaussian Processes	**Linear Regression**	**Multilayer Perceptron**	**SMOreg**	**Decision Table**	**M5Rules**	**M5P**	**REPTree**
Time	0.11	0.04	0.34	0.13	0.05	0.25	0.08	0.01
CC	0.8737	0.796	0.6021	0.647	0.9441	0.8687	0.8677	0.8887
RMSPE	11.2044	9.6389	12.5825	12.0804	5.0085	8.2345	8.3767	6.8494

As it is shown in Table 5.4, the Decision Table Classifier has the highest accuracy score on VOLUME_Dataset_1 with maximum *CCP* (94.41 %), minimum *RMSPE* (5.0085%) and an optimal *Time* (0.05 seconds) to build the Total Volume Predictive model.

Table 5.5 Summary of WEKA Classifiers Performance Evaluation on VOLUME_Dataset_2

	Classifier Functions				Classifier Rules		Classifier Trees	
	Gaussian Processes	**Linear Regression**	**Multilayer Perceptron**	**SMOreg**	**Decision Table**	**M5Rules**	**M5P**	**REPTree**
Time	0.09	0.03	0.18	0.02	0.05	0.06	0.08	0
CC	0.8705	0.7587	0.673	0.6514	0.6249	0.7587	0.7587	0
RMSPE	13.26078	10.1628	22.9302	13.2666	11.7946	10.1628	10.1628	14.7314

As it is shown in Table 5.5, the Linear Regression Classifier has the highest accuracy score on VOLUME_Dataset_2 with maximum *CCP* (75.87 %), minimum *RMSPE* (10.1628%) and an optimal *Time* (0.03 seconds) to build the Total Volume Predictive model.

Table 5.6 Summary of WEKA Classifiers Performance Evaluation on VOLUME_Dataset_3

	Classifier Functions				Classifier Rules		Classifier Trees	
	Gaussian Processes	**Linear Regression**	**Multilayer Perceptron**	**SMOreg**	**Decision Table**	**M5Rules**	**M5P**	**REPTree**
Time	0.2	0.001	0.1	0.01	0.04	0.03	0.02	0
CC	0.91	0.7881	0.9342	0.676	0.6064	0.7881	0.7881	0
RMSPE	10.1708	9.8421	6.1496	11.8802	12.0469	9.8421	9.8421	14.9216

As it is shown in Table 5.6, the Multilayer Perceptron Classifier has the highest accuracy score on VOLUME_Dataset_3 with maximum *CCP* (93.42 %), minimum *RMSPE* (6.1496 %) and an optimal *Time* (0.1 seconds) to build the Total Volume Predictive model.

5.3.2. Results and Discussions

As it was shown in the above experimentation result, while comparing the performance of various classifiers, it is found that the performance of classification techniques varies with different data sets. As noted by [27], factors that affect the classifier's performance are: Data set, Number of tuples and attributes, Type of Attributes, and System configuration. Thus further comparison should be done to choose the best classifier that performs well on all the datasets while building both (the WAP & Sales Volume) predictive models respectively. To do so, the average CCP, RMSPE and Time values is considered and the statistical summary is depicted on the following tables.

Table 5.7 Summary of WEKA Classifiers Performance Evaluation on WAP_Dataset_1, 2 & 3

	Classifier Functions				Classifier Rules		Classifier Trees	
	Gaussian Processes	**Linear Regression**	**Multilayer Perceptron**	**SMOreg**	**Decision Table**	**M5Rules**	**M5P**	**REPTree**
Average Time	0.27	0.01	0.17	0.04	0.02	0.17	0.04	0
Average CC	0.9340	0.9472	0.9088	0.9474	0.9775	0.941	0.9488	0.9755
Average RMSPE	8.0737	5.2389	7.0578	6.5883	3.8914	5.2352	4.7523	4.0538

As it is depicted in Table 5.7, the Decision Table Classifier has the highest optimal accuracy score on WAP_Dataset_1, WAP_Dataset_2 and WAP_Dataset_3 with maximum optimal *CCP* (97.75 %), minimum optimal *RMSPE* (3.8914 %) and an optimal *Time* (0.02 seconds) to build the WAP Predictive model. (*See on Annex 4, for the Weka Result Buffer full output of this WAP predictive model*)

Table 5.8 Summary of WEKA Classifiers Performance Evaluation on VOLUME_Dataset_1, 2 & 3

	Classifier Functions				**Classifier Rules**		**Classifier Trees**	
	Gaussian Processes	**Linear Regression**	**Multilayer Perceptron**	**SMOreg**	**Decision Table**	**M5Rules**	**M5P**	**REPTree**
Average Time	0.13	0.02	0.21	0.05	0.05	0.11	0.06	0.003
Average CC	0.8847	0.7809	0.7364	0.6581	0.8052	0.8052	0.8048	0.2962
Average RMSPE	11.5453	9.8813	13.8874	12.4091	9.6167	9.4131	9.4605	12.1675

As it is depicted in Table 5.8, the M5Rules Classifier has the highest optimal accuracy score on VOLUME_Dataset_1, VOLUME_Dataset_2 and VOLUME_Dataset_3 with maximum optimal *CCP* (80.52 %), minimum optimal *RMSPE* (9.4131 %) and an optimal *Time* (0.11 seconds) to build the Total Volume Predictive model. (*See on Annex 5, for the Weka Result Buffer full output of this Total Volume predictive model*)

Thus, the Weka.Classifier.Rules.DecisionTable and Weka.Classifier.Rules.M5Rules are implemented for building the Price & Sales Volume Predictive model respectively. Moreover, farther works would be done on the improvement of the selected classification technique thereby improving the efficiency of classifiers much more.

5.3.3. Model Prototyping and Usage

Creation of the model is generally not the end of the project. Even if the purpose of the model is to increase knowledge of the data, the knowledge gained will need to be organized and presented in a way that the client can use. Depending on the requirements, the deployment phase can be as simple as generating a report or as complex as implementing a repeatable data mining process. In many cases it will be the client, not the data analyst, who will carry out the deployment steps. However, even if the analyst will not carry out the deployment effort it is important for the client to understand up front what actions will need to be carried out to make use of the models created [1, 21, 26].

Thus, this (Deployment or in our case: Model Prototyping & Usage) is the final and most important phase of this study. To this end, series of experiments were run using WEKA frame work to achieve model training and prediction. In this phase, the implementation approach we have followed is: Customizing and Extending WEKA (Weka-Src V3.6.12 Framework).

The Development Environment is, for the front end (NetBeans IDE 7.1.2, JDK 1.6.0_26, JRE 6), and Microsoft SQL Server 2008 R2 for the back end (data base & data warehouse management). The system architecture is a 2-tier system architecture i.e. the application is deployed in end users machine while the database is in another central database server. Thus, application is a multi-user desktop application that can be accessed in local area network environment with "Client-Server" Architecture.

The overall system is then termed as "***ECX Market Price and Sales Volume Prediction System***". The two major components (sub-systems) in this system are *WAP Predictive Model Prototype* and *Sales Volume Predictive Model Prototype*. Where,

- *WAP Predictive Model Prototype*: The sub-system takes the Year, Month, Commodity and WAP (Class) for which the WAP is to be predicted and outputs the Predicted WAP.
- *Sales Volume Predictive Model Prototype*: The sub-system takes the Year, Month, Commodity, WAP_Predicted and Total Volume (Class) for which the Total Volume is to be predicted and outputs the Predicted Total Sales Volume.

The application has three basic menus/panels: *Preprocess, SQL* and *Prediction Panels* and one supportive menu/panel i.e. *Experiment Panel*. A screen shot of these main panels are shown in the following figures

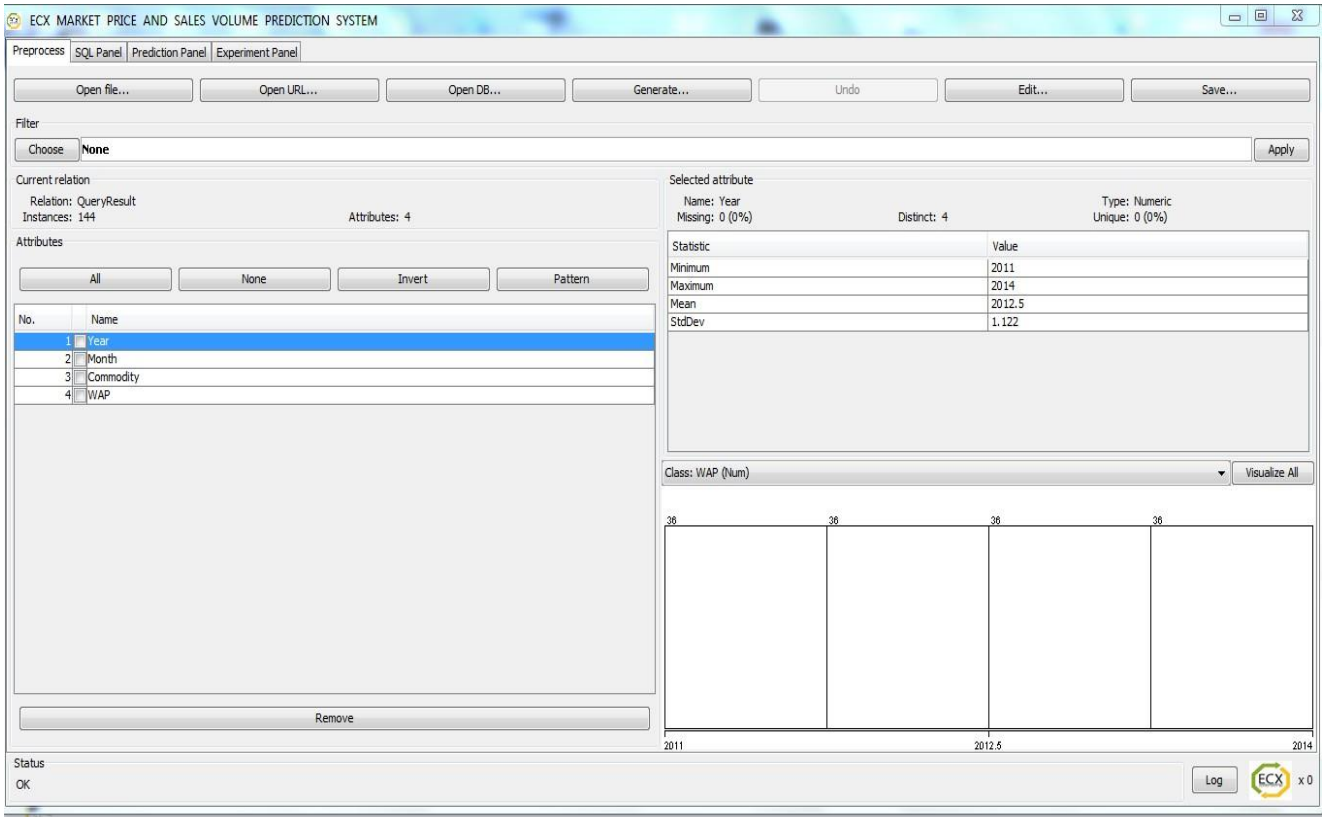

Figure 5.3 ECX Price & Sales Volume Prediction System: Preprocess Panel Main Window

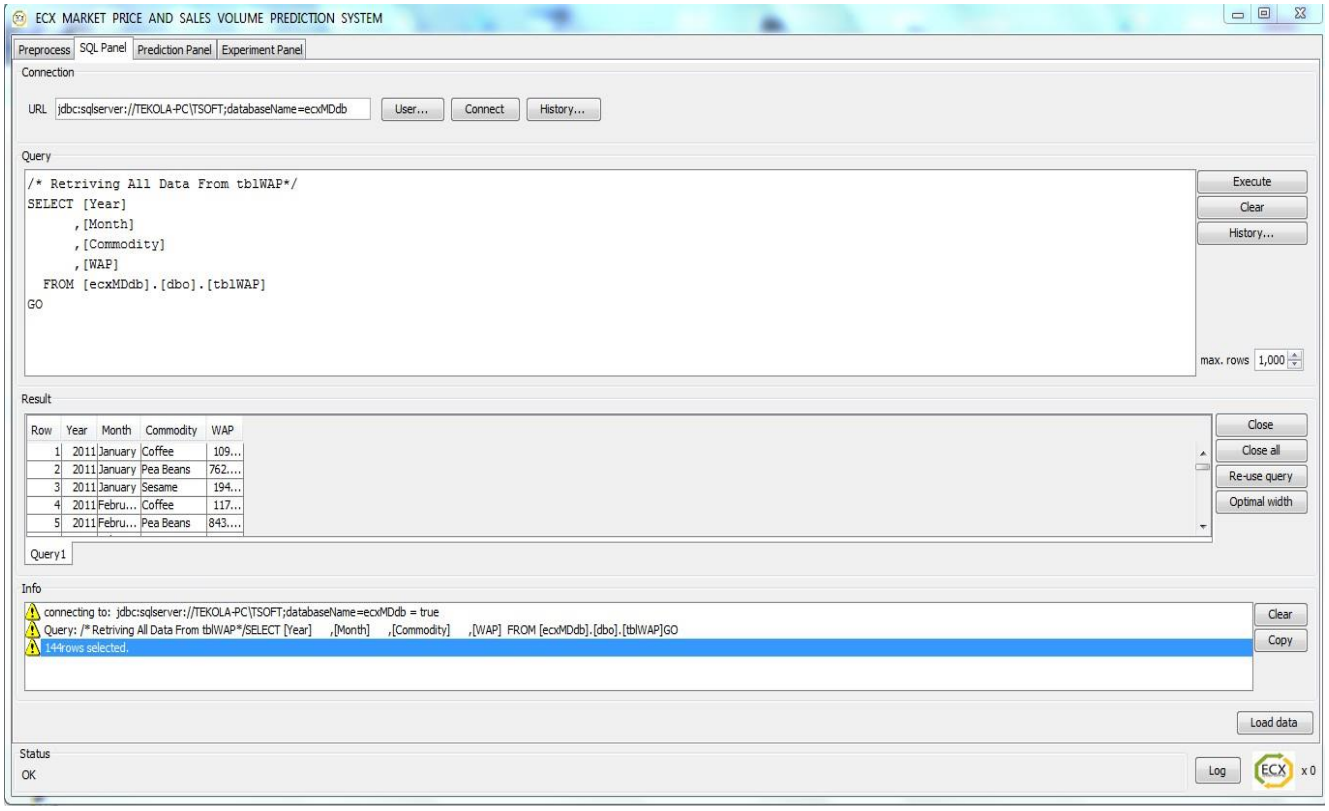

Figure 5.4 ECX Price & Sales Volume Prediction System: SQL Panel Main Window

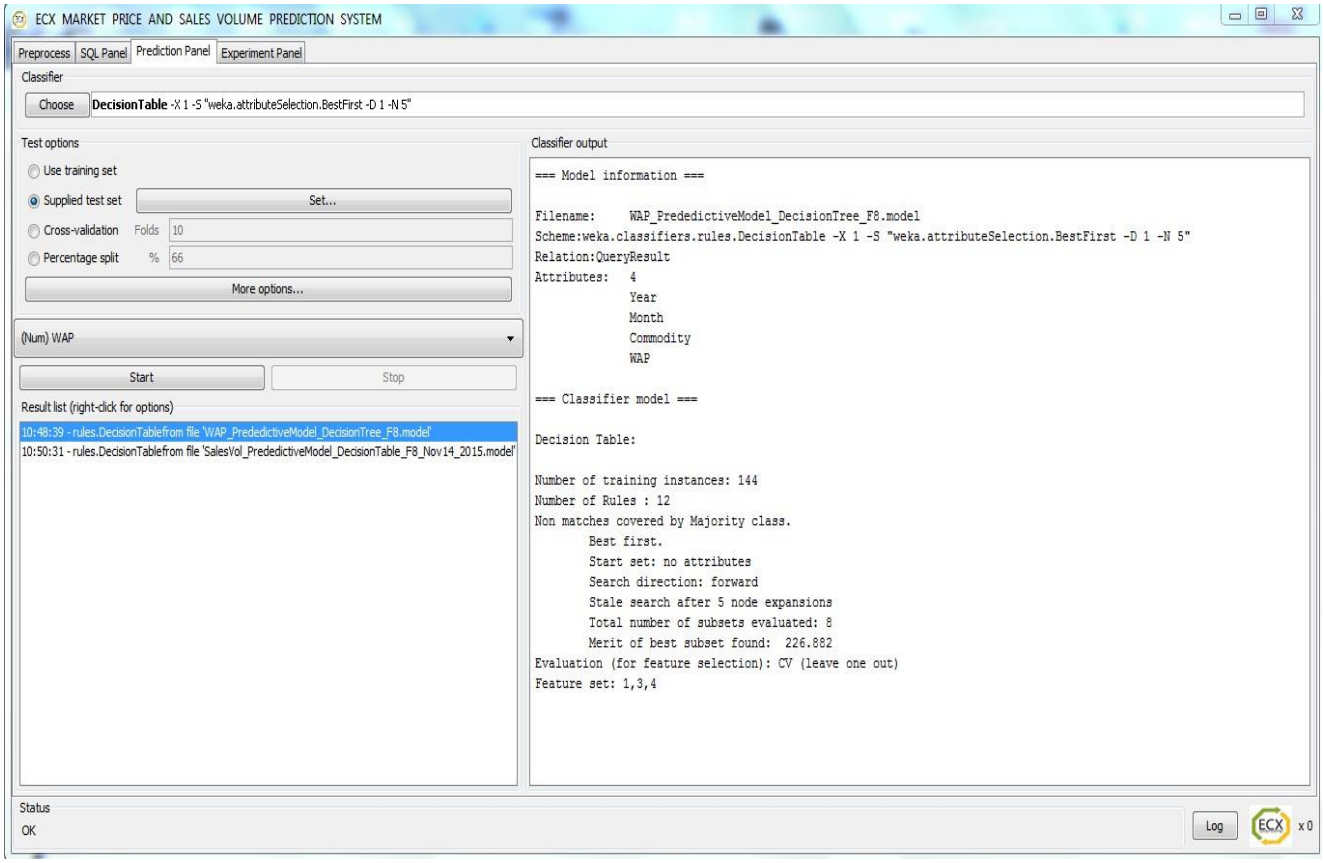

Figure 5.5 ECX Price & Sales Volume Prediction System: Prediction Panel Main Window

Finally, the system generates the predicted Price (WAP) and Total (Sales) Volume values in a prediction result buffer tables prospectively. The prediction result buffer tables provide a good and easy interface to the forecasters (domain experts), so that the predicted values will be compared with the actual values when the time comes.

CHAPTER SIX

CONCLUSION AND RECOMMENDATION

6.1. Conclusion

An essential for the design and implementation of predicting price and sales volume of agricultural commodities traded at commodity exchange organizations such as the Ethiopia Commodity Exchange (ECX), data mining researches are able to reveal such facts through analysis of the available historical market data. Thus, the purpose of this study is to analyze the historical market data available on the Ethiopia Commodity Exchange organization using the best applicable data mining methods and build a predictive data mining model/prototype which will be termed as "*ECX Price & Sales Volume Predictive Model*" for the selected commodities (Coffee, Pea Beans and Sesame) traded at ECX.

In this particular research, ECX Market Data (EMD) has been utilized for the purpose of predicting future market status of commodities at ECX, as it is believed that EMD datasets contain relevant information about past market data that can show what will happen in the future. Raw data which were simply accumulated as result of day to day trade activities and used for the purpose of template based repetitive statistical report have been utilized for data mining research to show important predictor variables and prediction class in the data.

Before these data has been used for the purpose of classification mining some pre-processing steps such as data reduction, integration and transformation have been effectively performed which helped in achieving the objectives finally. Here, trivial data cleansing such as cleaning of error data and missing values were not a big issue as the EMD data were obtained from the ECX data warehouse.

After the EMD data is prepared and the datasets are ready for exploration, series of experiments have been carried to identify classification model and to extract relevant actionable predictive model. The purpose of experiments in various classification data mining techniques is to find the best performing predictive model that is able to predict the Price and Sales Volume (WAP and Total Volume) values of commodity trades taking selected variables as inputs. Thus in this particular study, for building Price and Sales Volume predictive model, the eight WEKA 3.6.12 classifiers algorithms (Gaussian Processes, Linear Regression, Multilayer Perceptron, SMOreg, Decision Table, M5Rule, M5P and REPTree) which are applicable to the EMD datasets specific to this study are used in the

experimentation. Moreover, different data sets are used on the experimentations, and it is found that the performance of classification techniques varies with different datasets.

The statistical parameters are: the time taken to build model (Time), the correlation coefficient percentage (CCP) and the root mean square percentage error (RMSPE) values are used for the evaluation of the models. And the selected (best) model is the one with the maximum *CCP* and minimum *RMSPE* of the average outputs corresponding to the testing phase that consumes an optimal *Time* to build the model. Accordingly, the experimental results obtained from this study shows; the Decision Table Classifier has the highest optimal accuracy score on the three EMD WAP Datasets with maximum optimal *CCP* (97.8 %), minimum optimal *RMSPE* (3.9 %) and an optimal *Time* (0.02 seconds) to build the Price Predictive model. While, the M5Rules Classifier has the highest optimal accuracy score on the three EMD TOTAL VOLUME Datasets with maximum optimal *CCP* (80.5 %), minimum optimal *RMSPE* (9.4 %) and an optimal *Time* (0.11 seconds) to build the Sales Volume Predictive model.

In general, in spite of the challenges and limitations of obtaining more predictor variables from external/international reference market data sources (as predictor variables are from internal/ECX market data only) and limited utilization of classification algorithms (as the WEKA 3.6.12 default options/parameters of each classifies are used) ; the results are encouraging and answers the research questions of this study.

Therefore, the results obtained from this research indicate that data mining predictive models are useful in predicting future market status to commodity exchange organizations like ECX for the effective and efficient utilization of massive amount of market data to support experts and stockholders (mainly traders) in making strategic planning as well as proactive and knowledge-driven decisions.

6.2. Recommendation

To the knowledge of the researcher, no attempt has been made using data mining classification technique to predict the price and sales volume of agricultural commodities (Coffee, Pea Beans and Sesame) at the ECX. Moreover, this work has been the first attempt to evaluate the applicability of DM for predicting both the price and sales volume of agricultural commodities (Coffee, Pea Beans and Sesame) at the ECX. Last not least, by extending WEKA software source code, an application (predictive-model-prototype) which is termed as “ECX Price and Sales Volume Predictive System” with a user-friendly GUI have been developed and deployed for the usage of domain experts (end

users); which was not used in most of the previous studies.

However, in this research, we have not analyzed the sparse property (low complexity) of the selected classifiers, since we have worked with relatively medium data set (144 monthly observations) to train the models. Future research would be performed by analyzing weekly market data with much more observations and fully exploit the sparse characteristics of the classifiers approach when dealing with large dataset. It is also recommended to incorporate more predictor variables (obtained from International/Reference companies' market data) in the predictive analytics. Moreover, only the three commodities (Coffee, Pea Beans and Sesame) are considered on this study, thus further works could be done to get a more generic model that applies for all the commodities traded at the exchange.

Finally, future works could also focus on the improvement of selected classification techniques thereby improving the efficiency of predictive model by fully utilizing all the possible options (parameters) of each classifier. Also, a combination of other data mining techniques (such as: association and clustering) with classification technique could be used to excel the performance and accuracy of the predictive model. Moreover the developed predictive model prototype/ application could be integrated with Knowledge Based Systems and Case Based Systems for a higher level of data & knowledge extractions.

REFERENCES

[1] Joyce Jackson (2002). Data Mining: A Conceptual Overview. In: *Communications of the Association for Information Systems, Volume 8: 267-296*

[2] Mykola Pechenizkiy, Seppo Puuronen, and Alexey Tsymbal (2006). Does Relevance Matter to Data Mining Research? In: *Communications of the Association for Information Systems, Vol. 9: 1-25*

[3] Data Mining: What is Data Mining?

http://www.anderson.ucla.edu/faculty/jason.frand/teacher/technologies/palace/datamining.htm *Accessed: 8 April, 2015.*

[4] Data Mining Software, Tools and Applications.

http://www.the-data-mine.com/Software/DataMiningSoftware *Accessed: 12 May, 2015.*

[5] S.D.Gheware, A.S.Kejkar, and S.M.Tondare (2014). Data Mining: Task, Tools, Techniques and Applications. In: *International Journal of Advanced Research in Computer and Communication Engineering. Vol. 3, Issue 10.*

[6] Yongjian FU (2000). Data Mining: Tasks, Techniques and Applications. *Department of Computer Science; University of Missouri, Rolla.*

[7] Cha and Lweis (2002). Data Mining Tasks. In: *Communications of the Association for Information Systems, Ch 8.*

[8] Blaz Zupan and Janez Demsar (2008). Open-Source Tools for Data Mining. In: *Clin Lab Med 28,37-54*

[9] Berry, M. and Linoff , G. (2004). Data mining techniques for marketing, sales, and customer relationship management (2nd ed).

[10] Kantardzic, M.(2003). Data Mining: concepts, models, methods, and algorithms. In: *Database Mining Journal* vol. I

[11] Han, J & Kamber, M. (2006).*Data mining: concepts and techniques*. (2nd edition).

[12] Fayyad U., Piatetsky-Shapiro G, and Smyth P. (1996). Data mining to knowledge discovery in databases. In: *American Association for Artificial Intelligence: 36-51.*

[13] Samuel Odei Danso (2006). An Exploration of Classification Prediction Techniques in Data Mining: The Insurance Domain. In: *Predictive Data Mining Journal* vol. 2

[14] Mihai ANDRONIE and Daniel CRISAN (2010).Commercially Available Data Mining Tools used in the Economic Environment .In: *Database Systems Journal* vol. I, no. 2/2010

[15] Alex A. Freitas (2001). A Survey of Evolutionary Algorithms for Data Mining and Knowledge Discovery.

[16] Xindong Wu and Vipin Kumar (2008). The Top Ten Algorithms in Data Mining. From: *Taylor & Francis Group, LLC, 2009.*

[17] Michael Abernethy. Data mining with WEKA, Part 2: Classification and Clustering.pdf. From: www.ibm.com/developerWorks/ *Accessed 10 May 2015*.

[18] Neil Patel. 10 Ways Data Mining Can Help You Get a Competitive Edge. From: https://blog.kissmetrics.com/**data**-**mining**/ *Accessed 27 March 2015.*

[19] Efstathios Kirkos1and Yannis Manolopoulos (2006). Data Mining in Finance and Accounting: A Review of Current Research Trends. In: *Data Mining in Finance & Accounting Journal* vol. 1

[20] Boris Kovalerchuk and Evgenii Vityaev (2005). Data Mining for Financial Applications. In: *Data Mining for Financial Applications Journal* vol. 10

[21] Timothy D. Rey, Chip Wells, and Justin Kauhl (2013). Using Data Mining in Forecasting Problems. In: *SAS Global Forum; Paper*: *085-2013*

[22] Stephen Langdell. Examples of the use of data mining in financial applications. In*: The Data Analysis and Visualization Group at NAG Ltd.*

[23] Ahmedabad. (2011). Advanced Data Analysis, Business Analysis and Intelligence. In:2nd *IIMA International Conference.*

[24]) Cios K. and Kurgan L. (2005). Trends in Data Mining and Knowledge discovery. In: *Advanced Techniques in Knowledge Discovery and Data Mining, pp. 1–26.*

[25] M.K.Saranya, R.Rathnavathy and Dr.G.N.K. Suresh Babu. A Study of Predictive Data Mining Techniques. (2014). In: *International Journal of Computer Science and Information Technology Research ISSN 2348-120X (online), Vol. 2, Issue 2, pp: (258-264),* Available at: *www.researchpublish.com*

[26] Martin Butler (2013). Predictive Analytics in Business: Strategy, methods, technology. In: *Predictive Analytics Journal* vol. 2

[27] V. Vaithiyanathan, K. Rajeswari, Kapil Tajane, and Rahul Pitale (2013). Comparison of Different Classification Techniques Using Different Datasets. In: *International Journal of Advances in Engineering & Technology. ©IJAET ISSN: 2231-1963*

[28] Reinartz, T. (1999) Focusing Solutions for Data Mining: Analytical Studies and Experimental Results in Real-World Domains. In: *Data Mining Analytics Journal* vol. 1

[29] Portia A. Cerny and Aim Proximity (2010). Data mining and Neural Networks from a Commercial Perspective. In: *Data Mining & Neural Networks Journal* vol. 3

[30] Vincent Meens (2012). Data analytics at work: Introducing a data mining process framework to enable consultants to determine effective data analytics tasks. In: *Data Mining Process Journal* vol. 3

[31] D. Venugopal Setty, T.M.Rangaswamy, and K.N.Subramanya. (2010). A Review on Data Mining Applications to the Performance of Stock Marketing. In: *International Journal of Computer Applications (0975 – 8887. Volume 1 – No. 3.*

[32] Mrs. Keerti. S. Mahajan & R. V. Kulkarni. (2013). A Review :Application of Data Mining Tools for Stock Market. In: *Keerti S Mahajan et al ,Int.J. Computer Technology & Applications,Vol 4 (1), 19-27.*

[33] Dr. M.Hanumanthappa1 & Sarakutty.T.K (2011).Predicting the Future of Car Manufacturing Industry using Data Mining Techniques.In: *ACEEE Int. J. on Information Technology, Vol. 01, No. 02.*

[34] QASEM A. AL-RADAIDEH, ADEL ABU ASSAF & EMAN ALNAGI. Predicting Stock Prices Using Data Mining Techniques (2013). In: *The International Arab Conference on Information Technology.*

[35] Magdalena Daniela NEMEŞ & Alexandru BUTOI. (2013). Data Mining on Romanian Stock Market Using Neural Networks for Price Prediction. In: *Informatics Economică; vol. 17, No. 3/2013*

[36] Valtteri Ahti.(2009).Forecasting Commodity Prices with Nonlinear Models. In: *HECER Discussion Paper No. 268.*

[37] Dr. G.Manoj Someswar, B. Satheesh, and G.Vivekanand. (2009).Finance Mining – Analysis of Stock Market Exchange for Foreign Using Classification Techniques. In: *International Journal of Engineering Research and Applications (IJERA) ISSN: 2248-9622 www.ijera.com Vol. 2, Issue 4, pp.717-723.*

[38] Ramesh A. Medar and Vijay. S. Rajpurohit. (2014). Data mining in Agriculture on Crop Price Prediction Techniques and Applications. From: *International Journal of Computer Applications (0975 – 8887) Vol 99, No.12.*

[39] Andres M. Ticlavilca, Dillon M. Feuz, and Mac McKee. (2010).Forecasting Agricultural Commodity Prices Using Multivariate Bayesian Machine Learning Regression. In: *Paper presented at the NCCC-134 Conference on Applied Commodity Price Analysis, Forecasting, and Market Risk Management.*

[40] G.M.Nasira and N.Hemageetha. (2012).Forecasting Model for Vegetable Price Using Back Propagation Neural Network. In: *International Journal of Computational Intelligence and Informatics, Vol. 2: No. 2.*

[41] Abubakar S. Magaji and Adeboye K.R. (2014). An Intense Nigerian Stock Exchange Market Prediction: Using Logistic with Back-Propagation ANN Model. In: *Science World Journal Vol 9 (No 2).*

[42] Wilma Latuny. (2012). Predicting the Selling Price of Dried EUCHEUMACOTTONII in Indonesia with Four Classifier of Data Mining Techniques. In: *ARIKA, Vol. 06, No. 2, ISSN: 1978-1105.*

[43] Mohammad Rafiuzaman. (014). Forecasting Chaotic Stock Market Data using Time Series Data Mining. In: *International Journal of Computer Applications (0975 – 8887) Volume 101– No.10.*

[44] Khaled Assaleh, Hazim El-Baz, and Saeed Al-Salkhadi.(2012).Predicting Stock Prices Using Polynomial Classifiers-The Case of Dubai Financial Market. In: *Journal of Intelligent Learning Systems and Applications, Vol 3, PP 82-89.*

[45] Mahdi Pakdaman Naeini, Hamidreza Taremian and Homa Baradaran Hashemi.(2010).Stock Market Value Prediction Using Neural Networks. From: *2010 International Conference on Computer Information Systems and Industrial Management Applications (CISIM).*

[46] Halim Budi Santoso , Albertus Joko Santoso and Eduard Rusdianto.(2013). Hybrid Clustering Method for Stock Price and Commodity Price. In: *International Journal of Science and Advanced Technology. Vol. 3 No 7.*

[47] Heping Pan, Chandima Tilakaratne, John Yearwood .(2013). Predicting Australian Stock Market Index Using Neural Networks Exploiting Dynamical Swings and Inter market Influences. In: *Journal of Research and Practice in Information Technology, Vol. 37, No.*

[48] Asif Ullah Khan, T. K. Bandopadhyaya and Sudhir Sharma. (2011). Stocks selected using SOM and Genetic Algorithm based backpropagation Neural Network gives better returns. In: *IJCSI International Journal of Computer Science Issues, Vol. 8, Issue 2.*

[49] Mayankkumar B Patel and Sunil R Yalamalle.(2014).Stock Price Prediction Using Artificial NN. From: *International Journal of Innovative Research in Science, Engineering and Technology, Vol. 3, Issue 6.*

[50] S Abdulsalam Sulaiman Olaniyi, Adewole, Kayode S. and Jimoh, R. G.(2011).Stock Trend Prediction Using Regression Analysis – A Data Mining Approach. In: *ARPN Journal of Systems and Software, Volume 1 No. 4.*

[51] Ramesh A. Medar and Vijay. S. Rajpurohit. (2014).A survey on Data Mining Techniques for Crop Yield Prediction. From: *International Journal of Advance Research in Computer Science and Management Studies Volume 2, Issue 9***.**

[52] Niket Kumar Choudhary, Yogita Shinde, Rajeswari Kannan and Vaithiyanathan Venkatraman.(2014).Impact of attribute selection on the accuracy of Multilayer Perceptron.From: *IJITKMI Volume 7, Number 2, pp. 32-36.*

[53] Kate A. Smith and Jatinder N.D. Gupta.(2000).Neural networks in business: techniques and applications for the operation researchers. From: *Computers & Operations Research 27.*

[54] Priyanka Pitale, Asha Ambhaikar. (2012). Sensitive Region Prediction using Data Mining Technique. In: *International Journal of Engineering and Advanced Technology (IJEAT) ISSN: 2249 – 8958, Volume-1, Issue-5.*

[55] Marin Cerjan (2013). A hybrid model of dynamic electricity price forecasting with emphasis on price volatility. In: *Data Mining Hybrid Model* vol. 4

[56] Deena Younis Abo Khashan. (2014). Using Data Mining and Text Mining Techniques in Predicting the Price of Real Estate Properties in Dubai. In: *Thesis Paper*.

[57] Andre Christoffer. (2012) .A Novel Algorithmic Trading Framework Applying Evolution and Machine Learning for Portfolio optimization. In: *Data Mining & Machine Learning Journal* vol. 103

[58] Ming Zhu and Lipo Wang. Intelligent (2012) Trading Using Support Vector Regression and Multilayer Perceptrons Optimized with Genetic Algorithms. In: *Data Mining & Multilayer Perceptrons Journal* vol. 10

[59] Gray P., Watson, H.J. (1998a), "Professional Briefings...Present and Future Directions in Data Warehousing", *Database for Advances in Information Systems*, *Summer*, *(29:3), 83-90.*

[60] Debashish Das and Mohammad Shorif Uddin. (2013). DATA MINING AND NEURAL NETWORK TECHNIQUES IN STOCK MARKET PREDICTION: AMETHODOLOGICAL REVIEW. From: *International Journal of Artificial Intelligence & Applications (IJAIA), Vol.4, No.1.*

[61] JAESOO KIM AND HEEJUNE AHN .(2009).A New Perspective for Neural Networks: Application to a Marketing Management Problem. In: *JOURNAL OF INFORMATION SCIENCE AND ENGINEERING 25, 05-16.*

[62] Kyoung-jae Kim. (2006). Artificial neural networks with evolutionary instance selection for financial forecasting. In: *Expert Systems with Applications 3,(519–526.*

[63] Lukáš Civín (2010). Back Propagation Neural Networks User's Manual. (2nd edition).

[64] ELIAS LEMUYE (2011). HIV STATUS PREDICTIVE MODELING USING DATA MINING TECHNOLOGY, *AAU - Thesis Paper*.

[65] LEUL WOLDU ASEGEHGN.(2003). THE APPLICATION OF DATA MINING IN CRIME REVENTION: THE CASE OF OROMIA POLICE COMMISSION, *AAU - Thesis Paper*.

[66] DENEKEW ABERA JEMBERE.(2003). THE APPLICATION OF DATA MINING TO SUPPORT CUSTEMER RELATIONSHIP MANAGEMENT AT ETHIOPIAN AIRLINES, *AAU - Thesis Paper*.

[67] Tariku Debela.(2013). DEVELOPING A PREDICTIVE MODEL FOR FERTILITY PREFERENCE OF WOMEN OF REPRODUCTIVE AGE USING DATA MINING TECHNIQUES, *AAU - Thesis Paper*.

[68] Ethiopia Commodity Exchange-Official Website. www.ecx.com.et (*Accessed on: March 2015*)

[69] ECX Market Data Presentation (By: ECX Market Data Unit). From: ECX Portal, *Accessed on June 2015.*

[70] ECX market Bulletin (From March 31 to April 04, 2014). From: www.ecx.com.et , *Accessed on March 2015.*

ANNEXES

ANNEX 1: Description of the Main EMD Attributes

Attribute Name	Description	Type	Valid Values
Year	Commodity trading year	Numeric	4 Digit Numeric (Starting from 2010)
Month	Commodity trading month	Character String	{January, February, March, April, May, June, July, August, September, October, November, December}
Commodity	The commodity type	Character String	{Coffee, Pea beans, Sesame}
Ton	The total amount of commodity traded in tons	Numeric	Numeric value > 0
Kg	The total amount of commodity traded in Kilogram	Numeric	Numeric value > 0
Feresula/Quintal	The total amount of commodity traded (in Feresula for Coffee and in Quintal for other commodities)	Numeric	Numeric value > 0
Price	The price (in Et. Birr per LOT) of a commodity traded	Numeric	Numeric value > 0
Total Value	Total amount (in Ethiopian Birr) obtained from a commodity trade	Numeric	Numeric value > 0
Total Volume	Total traded amount/volume (in Tons) obtained from a commodity trade.	Numeric	Numeric value > 0
WAP	Weighted Average Price (WAP)= Total Value divided by Total Feresula/Quintal	Numeric	Numeric value > 0

ANNEX 2: WAP Data Set ARRF file

Note*:* Following is a sample WAP Data Set ARRF File (of Year 2011 and Months January to April only). ***The whole WAP Data set is available separately!***

@relation Jan2011_Dec2013_WAP_Training

@attribute Year numeric

@attribute Month {January ,February, March, April, May, June, July, August, September, October, November, December}

@attribute Commodity {Coffee, 'Pea Beans', Sesame}

@attribute WAP numeric

@data

2011,January,Coffee,1091.82575

2011,January,'Pea Beans',762.453731

2011,January,Sesame,1949.037209

2011,February,Coffee,1178.919345

2011,February,'Pea Beans',843.420964

2011,February,Sesame,1950.109727

2011,March,Coffee,1205.58947

2011,March,'Pea Beans',920.662092

2011,March,Sesame,1911.541685

2011,April,Coffee,1264.152996

2011,April,'Pea Beans',948.431576

2011,April,Sesame,1955.281235

ANNEX 3: VOLUME Data Set ARRF file

Note*:* Following is a sample VOLUME Data Set ARRF File (of Year 2011 and Months January to April only). ***The whole VOLUME Data set is available separately!***

@relation Jan2011_Dec2013_VOLUME_Training_F5

@attribute Year numeric

@attribute Month {January, February, March, April, May, June, July, August, September, October, November, December}

@attribute Commodity {Coffee, 'Pea Beans', Sesame}

@attribute WAP numeric

@attribute 'Total Volume' numeric

@data

2011,January,Coffee,1091.82575,19989.994

2011,January,'Pea Beans',762.453731,6633

2011,January,Sesame,1949.037209,40065.4

2011,February,Coffee,1178.919345,20397.841

2011,February,'Pea Beans',843.420964,4733.9

2011,February,Sesame,1950.109727,37053

2011,March,Coffee,1205.58947,20105.237

2011,March,'Pea Beans',920.662092,4211.2

2011,March,Sesame,1911.541685,30185.8

2011,April,Coffee,1264.152996,20801.404

2011,April,'Pea Beans',948.431576,1735.5

2011,April,Sesame,1955.281235,17449.8

ANNEX 4: Output of the Selected WAP Predictive Model

=== Run information ===

Scheme:weka.classifiers.rules.DecisionTable -X 1 -S "weka.attributeSelection.BestFirst -D 1 -N 5"

Relation: Jan2011_Dec2013_WAP_Training_F5

Instances: 72

Attributes: 4

Year

Month

Commodity

WAP

Test mode:evaluate on training data

=== Classifier model (full training set) ===

Decision Table:

Number of training instances: 72

Number of Rules : 6

Non matches covered by Majority class.

Best first.

Start set: no attributes

Search direction: forward

Stale search after 5 node expansions

Total number of subsets evaluated: 8

Merit of best subset found: 248.328

Evaluation (for feature selection): CV (leave one out)

Feature set: 1,3,4

Time taken to build model: 0.02 seconds

=== Evaluation on training set ===

=== Summary ===

Correlation coefficient 0.9735

Mean absolute error 150.785

Root mean squared error 227.6343

Relative absolute error 17.7034 %

Root relative squared error 22.8521 %

Total Number of Instances 72

=== Re-evaluation on test set ===

User supplied test set

Relation: Jan2014_Dec2014_WAP_Test1_F5

Instances: unknown (yet). Reading incrementally

Attributes: 4

=== Summary ===

Correlation coefficient 0.9775

Mean absolute error 338.0339

Root mean squared error 389.1418

Total Number of Instances 36

=== Re-evaluation on test set ===

User supplied test set

Relation: Jan2015_Mar2015_WAP_Test2_F5

Instances: unknown (yet). Reading incrementally

Attributes: 4

=== Summary ===

Correlation coefficient 0.9617

Mean absolute error	494.555
Root mean squared error	564.1907
Total Number of Instances	9

ANNEX 5: Output of the Selected Total VOLUME Predictive Model

=== Run information ===

Scheme:weka.classifiers.rules.M5Rules -M 4.0

Relation: Jan2011_Dec2013_VOLUME_Training_F5

Instances: 108

Attributes: 5

Year

Month

Commodity

WAP

Total Volume

Test mode:evaluate on training data

=== Classifier model (full training set) ===

M5 pruned model rules

(using smoothed linear models) :

Number of Rules : 2

Rule: 1

IF

Commodity=Coffee,Sesame > 0.5

Month=May,April,November,March,February,January,December > 0.5

THEN

Total Volume =

3867.6014 * Month=May,April,November,March,February,January,December

+ 856.1326 * Month=November,March,February,January,December

+ 9660.7439 * Month=March,February,January,December

+ 1105.9287 * Month=February,January,December

+ 2112.0341 * Commodity=Coffee,Sesame

+ 17548.6938 * Commodity=Sesame

- 5.0347 * WAP

+ 14983.2456 [42/85.946%]

Rule: 2

Total Volume =

1494.8288 * Year

+ 2075.3789 * Month=July,June,May,April,November,March,February,January,December

+ 1505.9761 * Month=June,May,April,November,March,February,January,December

+ 2535.4486 * Month=November,March,February,January,December

+ 3132.0226 * Month=January,December

+ 10071.7331 * Commodity=Coffee,Sesame

- 4.4171 * WAP

- 3001698.3092 [66/46.927%]

Time taken to build model: 0.33 seconds

=== Evaluation on training set ===

=== Summary ===

Correlation coefficient	0.8346
Mean absolute error	5102.7077
Root mean squared error	7936.444
Relative absolute error	49.2113 %
Root relative squared error	55.2178 %

Total Number of Instances 108

=== Re-evaluation on test set ===

User supplied test set

Relation: Jan2014_Dec2014_VOLUME_Test1_F5

Instances: unknown (yet). Reading incrementally

Attributes: 5

ANNEX 6: Overview of the Selected WEKA Classifiers

The eight Weka Classifier used for the experimentation of this particular study are categorized in to three main categories: *Functions*, *Rules* and *Trees*. The detail of these classifiers is presented as follows. (*Source: Weka 3.6.12/Documentation*).

➢ Weka.Classifiers.Functions

- **Gaussian Processes (**An overview of "Weka.Classifiers.Functions.GaussianProcesses")

NAME	weka.classifiers.functions.GaussianProcesses
SYNOPSIS	Implements Gaussian processes for regression without hyper-parameter-tuning. To make choosing an appropriate noise level easier, this implementation applies normalization/standardization to the target attribute as well (if normalization/standardization is turned on). Missing values are replaced by the global mean/mode. Nominal attributes are converted to binary ones
CAPABILITIES	Class -- Numeric class, Date class, Missing class values Attributes -- Binary attributes, Empty nominal attributes, Numeric attributes, Unary attributes, Missing values, Nominal attributes Additional -- min # of instances: 1
OPTIONS	debug -- If set to true, classifier may output additional info to the console. filterType -- Determines how/if the data will be transformed. kernel -- The kernel to use. noise -- The level of Gaussian Noise (added to the diagonal of the Covariance Matrix).

- **Linear Regression (**An overview of "Weka.Classifiers.Functions.LinearRegression")

NAME	weka.classifiers.functions.LinearRegression
SYNOPSIS	Class for using linear regression for prediction. Uses the Akaike criterion for model selection, and is able to deal with weighted instances.
CAPABILITIES	Class -- Numeric class, Date class, Missing class values Attributes -- Binary attributes, Date attributes, Empty nominal attributes, Numeric attributes, Unary attributes, Missing values, Nominal attributes Additional -- min # of instances: 1
OPTIONS	attributeSelectionMethod -- Set the method used to select attributes for use in the linear regression. Available methods are: no attribute selection, attribute selection using M5's method (step through the attributes removing the one with the smallest standardised coefficient until no improvement is observed in the estimate of the error given by the Akaike information criterion), and a greedy selection using the Akaike information metric.

	debug -- Outputs debug information to the console. eliminateColinearAttributes -- Eliminate colinear attributes. ridge -- The value of the Ridge parameter.

- **Multilayer Perceptron (**An overview of "weka.Classifiers.Functions.MultilayerPercepton")

NAME	weka.classifiers.functions.MultilayerPercepton
SYNOPSIS	A Classifier that uses back-propagation to classify instances. This network can be built by hand, created by an algorithm or both. The network can also be monitored and modified during training time. The nodes in this network are all sigmoid (except for when the class is numeric in which case the output nodes become unthresholded linear units).
CAPABILITIES	Class -- Binary class, Numeric class, Date class, Missing class values, Nominal class Attributes -- Binary attributes, Date attributes, Empty nominal attributes, Numeric attributes, Unary attributes, Missing values, Nominal attributes Additional -- min # of instances: 1
OPTIONS	autoBuild -- Adds and connects up hidden layers in the network. debug -- If set to true, classifier may output additional info to the console. decay -- This will cause the learning rate to decrease. hiddenLayers -- This defines the hidden layers of the neural network. This is a list of positive whole numbers. learningRate -- The amount the weights are updated. momentum -- Momentum applied to the weights during updating. nominalToBinaryFilter -- This will preprocess the instances with the filter. normalizeAttributes -- This will normalize the attributes. normalizeNumericClass -- This will normalize the class if it's numeric. reset -- This will allow the network to reset with a lower learning rate. seed -- Seed used to initialise the random number generator. trainingTime -- The number of epochs to train through. validationSetSize -- The percentage size of the validation set validationThreshold -- Used to terminate validation testing.

- **SMOreg (**An overview of "Weka.Classifiers.Functions.SMOreg"**)**

NAME	weka.classifiers.functions.SMOreg
SYNOPSIS	SMOreg (Support Vector Matchine for Regression) implements the support vector machine for regression. The parameters can be learned using various algorithms. The algorithm is selected by setting the RegOptimizer. The most popular algorithm (RegSMOImproved) is due to Shevade, Keerthi et al and this is the default RegOptimizer.
CAPABILITIES	Class -- Numeric class, Date class, Missing class values Attributes -- Binary attributes, Empty nominal attributes, Numeric attributes, Unary attributes, Missing values, Nominal attributes Additional -- min # of instances: 1
OPTIONS	c -- The complexity parameter C. debug -- If set to true, classifier may output additional info to the console. filterType -- Determines how/if the data will be transformed. kernel -- The kernel to use. regOptimizer -- The learning algorithm.

➢ Weka.Classifiers.Rules

- **Decision Table (**An overview of "Weka.Classifiers.rules.DecisionTable"

NAME	weka.classifiers.rules.DecisionTable
SYNOPSIS	Class for building and using a simple decision table majority classifier.
CAPABILITIES	Class -- Binary class, Numeric class, Date class, Missing class values, Nominal class Attributes -- Binary attributes, Date attributes, Empty nominal attributes, Numeric attributes, Unary attributes, Missing values, Nominal attributes Additional -- min # of instances: 1
OPTIONS	crossVal -- Sets the number of folds for cross validation (1 = leave one out). debug -- If set to true, classifier may output additional info to the console. displayRules -- Sets whether rules are to be printed. evaluationMeasure -- The measure used to evaluate the performance of attribute combinations used in the decision table. search -- The search method used to find good attribute combinations for the decision table. useIBk -- Sets whether IBk should be used instead of the majority class.

- **M5Rules (**An overview of "Weka.Classifiers.Rules.M5Rules")

NAME	weka.classifiers.rules.M5Rules
SYNOPSIS	Generates a decision list for regression problems using separate-and-conquer. In each iteration it builds a model tree using M5 and makes the "best" leaf into a rule.
CAPABILITIES	Class -- Numeric class, Date class, Missing class values Attributes -- Binary attributes, Date attributes, Empty nominal attributes, Numeric attributes, Unary attributes, Missing values, Nominal attributes Additional -- min # of instances: 1
OPTIONS	buildRegressionTree -- Whether to generate a regression tree/rule instead of a model tree/rule. debug -- If set to true, classifier may output additional info to the console. minNumInstances -- The minimum number of instances to allow at a leaf node. unpruned -- Whether unpruned tree/rules are to be generated. useUnsmoothed -- Whether to use unsmoothed predictions.

➢ Weka.Classiiers.Trees

- **M5P (**An overview of "Weka.Classifiers.Trees.M5P")

NAME	weka.classifiers.trees.M5P
SYNOPSIS	M5Base. Implements base routines for generating M5 Model trees and rules The original algorithm M5 was invented by R. Quinlan and Yong Wang made improvements.
CAPABILITIES	Class -- Numeric class, Date class, Missing class values Attributes -- Binary attributes, Date attributes, Empty nominal attributes, Numeric attributes, Unary attributes, Missing values, Nominal attributes Additional -- min # of instances: 1
OPTIONS	buildRegressionTree -- Whether to generate a regression tree/rule instead of a model tree/rule. debug -- If set to true, classifier may output additional info to the console. minNumInstances -- The minimum number of instances to allow at a leaf node. saveInstances -- Whether to save instance data at each node in the tree for visualization purposes. unpruned -- Whether unpruned tree/rules are to be generated. useUnsmoothed -- Whether to use unsmoothed predictions.

\

- **REPTree (**An overview of "Weka.Classifiers.Trees.REPTree")

NAME	weka.classifiers.trees.REPTree
SYNOPSIS	Fast decision tree learner. Builds a decision/regression tree using information gain/variance and prunes it using reduced-error pruning (with backfitting). Only sorts values for numeric attributes once. Missing values are dealt with by splitting the corresponding instances into pieces (i.e. as in C4.5).
CAPABILITIES	Class -- Binary class, Numeric class, Date class, Missing class values, Nominal class Attributes -- Binary attributes, Date attributes, Empty nominal attributes, Numeric attributes, Unary attributes, Missing values, Nominal attributes Additional -- min # of instances: 1
OPTIONS	debug -- If set to true, classifier may output additional info to the console. maxDepth -- The maximum tree depth (-1 for no restriction). minNum -- The minimum total weight of the instances in a leaf. minVarianceProp -- The minimum proportion of the variance on all the data that needs to be present at a node in order for splitting to be performed in regression trees. noPruning -- Whether pruning is performed. numFolds -- Determines the amount of data used for pruning. One fold is used for pruning, the rest for growing the rules. seed -- The seed used for randomizing the data.

********** ***END*** **********