



**Application of data mining for customs risk channel assignment: the
case of Ethiopian revenue and customs authority**

A Thesis Presented

By

Ephrem Bezabeh

The Faculty of Informatics

of

St. Marys' university

In Partial Fulfillment of the Requirements

for the Degree of Master of Science

In

Computer Science

July, 2019

ACCEPTANCE


Application of data mining for customs risk channel assignment: the case of Ethiopian revenue and customs authority

By

Ephrem Bezabeh

Accepted by the Faculty of Informatics, St. Mary's University, in partial fulfillment of the requirements for the degree of Master of Science in Computer Science

Thesis Examination Committee:

_____	_____
Internal Examiner	Signature &Date
Temtin Assefa (PhD)	 10/9/2020
External Examiner	Signature &Date
_____	_____
Dean, Faculty of Informatics	Signature &Date

Declaration

I, the undersigned, declare that this thesis work is my original work, has not been resented for a degree in this or any other universities, and all sources of materials used for the thesis work have been duly acknowledged.

Ephrem Bezabeh

Signature

Addis Ababa

Ethiopia

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

Dr. Getahun Semeon (PhD)

Signature

Addis Ababa

Ethiopia

July, 2019

Acknowledgement

I wish to express my profound gratitude to my advisors Dr. Getahun constructive suggestions and overall guidance during my work. I am also grateful to Ato Seweagen who is database administrator in customs for his professional inputs and support during my work I am also grateful to the Staff of Custom Risk management department who allowed access to this data and consulting to my work.

My special thanks are also due to all my friends for their all rounded contribution and support in the course of conducting this research work.

Abstract

Limiting intrusive customs examinations is recommended under the revised Kyoto Convention. It is also a proposal discussed in the context of World Trade Organization (WTO) trade facilitation negotiations. To limit these intrusive examinations, the more modern governments now intervene at all stages of the customs chain, using electronic data exchange and risk analysis, and focusing their resources on a posteriori inspection.

The Ethiopian Revenues and Customs Authority (ERCA) is one of the pioneers in implementing risk management in its customs processing. There is huge amount of data is being stored and processed daily activity for risk management. Just like other developing countries' customs office , it does not properly utilize its vast data records in a way that enables it to extract pattern and regularities important for forecasting problems related to risk in advance and take appropriate action. The Ethiopian revenue and customs authority is currently using Statistical calculation but mainly manual risk management scheme for selectivity for risk analysis as part of a process of putting analytics at the core of its business processes.

The problem is to be able to handle this huge amount of data and information in such a way that they can identify what is important and be able to extract it from the accumulated data. It is too complex and voluminous to be processed and analyzed by traditional methods. Now a day, data mining technology is being used as a tool that provides the techniques to transform these mounds of data into useful information which in turn enables to derive knowledge for decision making. A number of data mining techniques and tools are available to perform this task. The researcher considered selective techniques and tools which were used to explore the prevalence of Custom risk channel assignment and develop classification and prediction models.

Thus, the purpose of this study is to investigate the potential applicability of data mining techniques in exploring the prevalence of custom risk management using the data collected from Ethiopian Revenue and custom authority risk management database.

Three machine learning algorithms from WEKA software such as J48 Decision trees (DT), Naïve Bayes (NB) and K nearest neighbor classifiers are adopted to classify custom risk channel records on the basis of the values of attributes "Risk Level". Initially, a total dataset of 18814

records with 13 attributes were collected for the study. In this study CRISP-DM model was used as framework.

Results of the experiments have shown that K nearest neighbor (KNN) classifier has better classification and accuracy performance as compared to Naïve Bayes (NB) and Decision Tree classifier. The model selected in evaluation performance of these classifiers has an accuracy of 92.71 %.Overall, this study has proved that data mining techniques are valuable to support and scale up the efficacy of custom services provision process.

Keywords— Customs Risk Channel Assignment, Data mining, J48 Decision Tree, Naïve Bayes, K-NN

Table of Contents

Declaration	iii
Acknowledgement.....	iv
Abstract	v
List of Figures	xii
List of Tables.....	xiii
Acronyms and Abbreviations.....	xiv
CHAPTER ONE	1
INTRODUCTION	1
1.1. Background	1
1.2. Statement of the Problem.....	4
1.3. Objectives.....	7
1.3.1. General Objective	7
1.3.2. Specific Objectives	7
1.4. Methodology	8
1.5. Scope of the Study	9
1.6. Significance of the Study	9
1.7 Thesis Organization	10
CHAPTER TWO.....	11
LITRATURE REVIEW	11

2.1. DATA MINING.....	11
2.1.1. Introduction	11
2.2 The Data Mining Process.....	12
2.2.1 Data acquisition	13
2.2.2 Data preprocessing	13
2.2.3 Building model	13
2.2.4 Interpretation and model evaluation	13
2.3 Data Mining Tasks.....	14
2.3.1 Predictive modeling	14
2.3.2 Classification	14
2.3.2.1.1. Constructing Decision Tree	17
2.3.2.1.2. Rule induction.....	19
2.3.3 Naive Bayes (NB) Classifier	19
3.3.3.1. Naive Bayesian Classification Algorithm	20
2.3.4. The K-nearest Neighbor (K-NN) Algorithm	22
2.4. Descriptive modelling.....	25
2.4.1. Clustering	26
2.4.2 Association rule discovery	28
2.5. Types of Data Mining Systems.....	29
2.6. The Data Mining Models.....	30

2.6.1. The six step Cios model	30
2.6.2 The KDD process model	32
2.6.3. The CRISP-DM process	34
2.6.4. The SEMMA Process	35
2.6.5. Comparison of SEMMA, KDD and CRISP-DM	36
2.7. Data mining in customs	37
2.7.1. Overview of risk management	37
2.7.2. Risk Management at Customs	38
2.7.3. The Concept of Data Mining and Its Role in Identifying Risks at Customs	42
2.8. Risk Preparation/Profiling	43
2.9. Experience from other countries	46
2.9.1. Data Mining Project at Turkish Custom Authority (TCA)	47
2.10. Related works	49
CHAPTER THREE	52
RESEARCH METHODOLOGY	52
3.1 Introduction	52
3.2 Cross-Industry Standard Process for Data Mining (CRISP-DM) Process Model	53
3.2.1. Understanding the problem domain	53
3.2.1.1. International Trade and Customs	53
3.2.2. Data understanding	58

3.2.3 Data Preparation	60
3.2.3.1 Data preprocessing	60
3.2.3.1.2 Data Protection and Privacy Issues	64
3.2.4 Training and Building Models	65
3.2.4.1. Data Formatting	66
3.2.4.2. Algorithms Deployed	67
3.2.5 Evaluation	67
3.2.6 Deployment	68
CHAPTER FOUR	70
EXPERIMENTATION	70
4.1 Experimental setup.....	70
4.2 Model Building and Result Analysis	71
4.2.1 Decision Tree Model Building.....	71
4.3.2 Naive Bayes (NB) Model Building	76
4.3.2 K Nearest Neighbor (KNN) Model Building	79
4.4 Performance Evaluation of the models	82
CHAPTER FIVE	84
CONCLUSION AND RECOMMENDATION	84
5.1 Conclusion	84
5.2 Recommendation	86

Reference	87
Appendices.....	92
Appendix A:	92
Appendix B:.....	93

List of Figures

Figure 1.1: Risk Assignment Process.	3
Figure 2.1: KNN proximate algorithm map.....	23
Figure 2.2: The Six Step Cios et al. (2000) process model	32
Figure 2.3: The KDD Process	33
Figure 2.4: The CRISP-DM Process.....	34
Figure 2.5: Data mining process in customs	41
Figure 2.6: Development and Characteristics of a Risk Profile.....	44
Figure 3. 1: Statistic about class labels distribution in a data set based on ‘Risk Level’ as a target class before ‘Class Balancer’ was used.	62
Figure 3.2: Statistic about class labels distribution in a data set based on ‘Risk Level’ as a target class after ‘Class Balancer’ was used.	63
Figure 3.3: Rank of attributes	64
Figure 3.4: Sample data set to convert ARFF format	67
Figure 4. 1: Default Run option on Weka.....	73

List of Tables

Table 2.1: Comparison of Data mining process model.....	36
Table 2.2: Origin of Risks for Different Customs Objectives. (COMCEC, 2018).....	40
Table 2.3:Risk Analysis System Powered by Data Mining in Turkish.	49
Table 3.1: Attributes selected for experiments	59
Table 4.1: Output from J48 Decision Tree classifier based on ‘Risk Level’ as a target class (using 10 folds cross validation)	74
Table 4.2: Output from J48 Decision Tree classifier based on ‘Risk Level’ as a target class (using 66% split option)	75
Table 4.3: Output of Naive Bayes classifier based on” Risk Level” as a target class (using 10 folds cross validation)	77
Table 4.4: Output of Naive Bayes classifier based on “Risk Level” as a target class (using 66 percent split option)	78
Table 4.5: Output of Nearest Neighbor classifier based on “Risk Level’ as a target class (using 10 folds cross validation)	80
Table 4.6: Output of Nearest Neighbor classifier based on “Risk Level’ as a target class (using 66 percent splitting option)	81
Table 4.7: comparison of best accuracy between J48, Naïve Bayes and K Nearest Neighbor	83

Acronyms and Abbreviations

WCO	World Custom Organization
KDD	Knowledge Data Discovery
KDP	Knowledge Discovery Process
ERCA	Ethiopian Revenue and Customs Authority
WEKA	Waikato Environment for Knowledge Analysis
CEN	Custom Enforcement Network
DM	Data Mining
DT	Decision Trees
NB	Naive Bayes
KNN	K Nearest Neighbor
AEO	Authorized Economic Factor
WTO	World Trade Organization
ASYCUDA	Automated System for Customs Data
eCMEI	Electronic Custom Management System
CPC	Custom Procedure Code
Hs Code	Harmonized code
CRISP	Cross Industry Standard Process
SEMMA	Sample Explore Modify Model Assess
ANN	Artificial neural network
RN	Recurrent Neural
DT	Decision Tree
NB	Naïve Bayes

KNN	K Nearest Neighbor
KDD	Knowledge Discovery in Databases
AEO	Authorized economic operator
NN	Neural Network
UNCTAD	United Nation Conference on Trade and Development
AFRITAC	Africa Region Technical Assistance Center
IMF	International Monetary Fund's
IG	Information Gain
KDP	Knowledge proceeding
SMOTE	Synthetic Minority Oversampling Technique

CHAPTER ONE

INTRODUCTION

1.1. Background

A common characteristic of Customs work is the high volume of transactions and the impossibility of checking all of them. Customs administrations therefore face the challenge of facilitating the movement of legitimate passengers and cargo while applying controls to detect Customs fraud and other offences. Customs service's find themselves increasingly under pressure from national governments and international organizations to facilitate the clearance of legitimate passengers and cargo while also responding to increase in transactional crime and terrorism. These competing interests mean that it is necessary to find a balance between facilitation and control. Customs controls should ensure that the movement of vessels, vehicles, aircraft, goods and persons across international borders occurs within the framework of laws, regulations and procedures that comprise the Customs clearance process. Given the high number of export, import and transit transactions many Customs administrations use risk analysis to determine which persons, goods, and means of transport should be examined and to what extent [2]. Risk analysis and risk assessment are analytical processes that are used to determine which risks are the most serious and should have priority for being treated or having corrective action taken. Inspection selectivity programs make use of risk profiles, which have been established in a process of risk analysis and assessment. Risk profiles encompass various indicators, such as: type of good, know trader and compliance records of traders, value of goods and applicable duties, destination and origin countries, mode of transport and routes and are built based on characteristics displayed by unlawful consignments (or offending passengers). The development of profiles relies heavily on the gathering, charting and analysis of intelligence and the WCO has developed various tools to assist its member countries in the establishment of profiles and the management of intelligence collection. The WCO Customs Enforcement Network (CEN) database can, for example, provide useful intelligence for the establishment of risk profiles. These profiles then drive inspection selectivity programs, through which data declared will be analyzed on the basis of the identified risk parameters and consignments, and depending on the

selected risk level, goods and persons are routed through different channels of Customs control. This process is obviously labor intensive and inconsistent as the analysis and data collection is done manually it is highly dependent on the effectiveness, skill and knowledge of the officers.

The Ethiopian revenue and customs authority (ERCA) came into existence on 2008 by the merger of the ministry of revenue, customs authority and the federal Inland Revenue authority. According to proclamation No. 587/2008 of the Ethiopia federal Negarit Gazeta,[3] ERCA has the power and duties are establish and implement modern revenue assessment and collection system; provide efficient, equitable and quality service; properly enforce incentive of tax exemption given to investors and ensure such incentives are used as per intended purposes. It has also the responsibility to collect and analyze information necessary for the control of import and export goods and the assessment of duty and taxes; compile statistical data on criminal offences relating to the sector, and disseminate the information to stakeholders.[4]

ERCA implemented risk analysis and uses selectivity where the analysis methodology is based on 15 criteria. From which 8 criteria are based on the core particulars of the customs declaration: customs value, customs classification (tariff), country of origin, consignment, the CPC (customs procedure code), special certificate, the company and those of the customs clearing agents. The remaining 7 criteria are Car list, Diplomatic list, Government list, 5% random selection (from Yellow and Green declaration), Yellow (Raw material and chemical) list, Authorized economic operator (AEO) list, and Manufacturing (Especially privileged company) list.

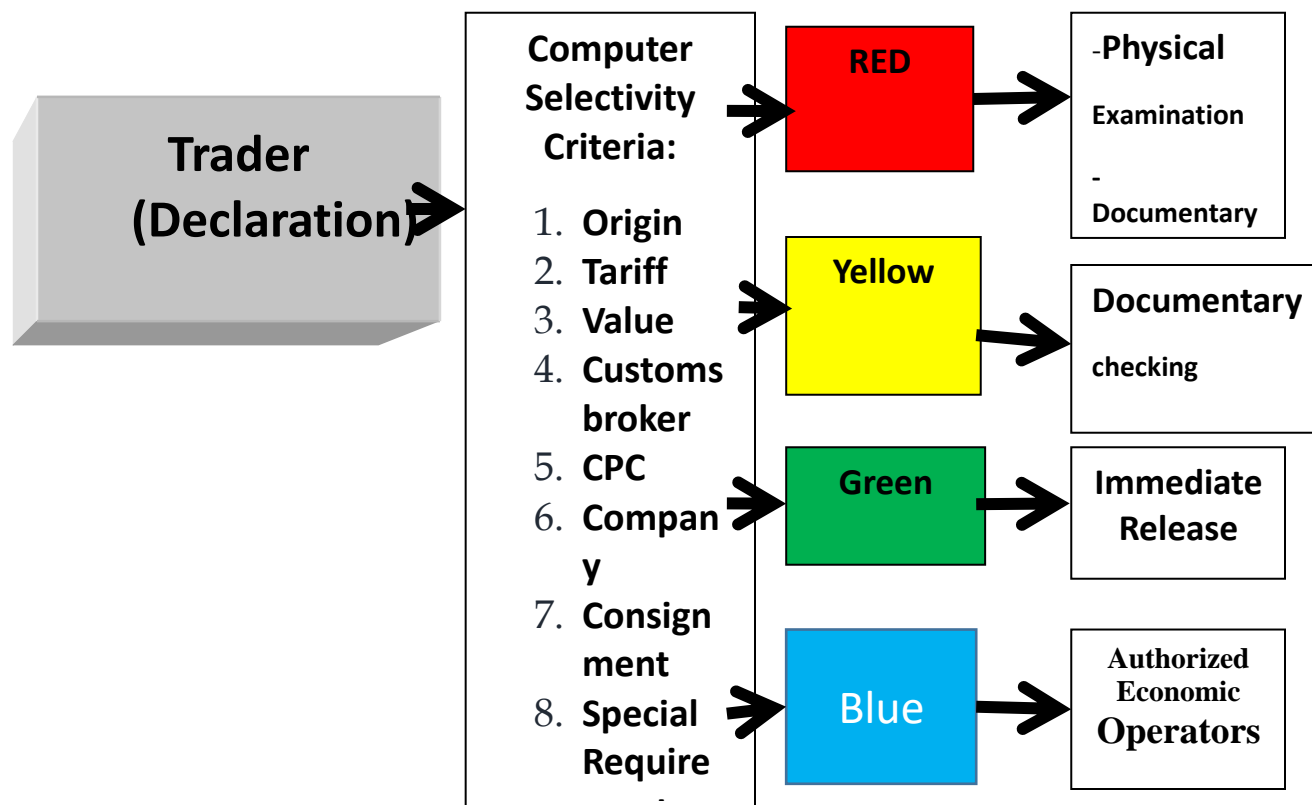


Figure 1.1: Risk Assignment Process.

In an attempt to address the issue of fraud investigation and protection in the case of ERCA, Mamo [5] proposed computer-based tools like data mining to be experimented in the area of risk management. Data Mining is a technology that use various technique to discover hidden knowledge from heterogeneous and distributed data stored in large databases, warehouses and other massive information repositories so as to find patterns that are valid, novel, useful, and understandable [6].

Risk Management is the systematic application of management procedures and practices providing Customs with the necessary information to address movements or consignments which present a risk. This focus is necessary since the fundamental task of the Customs is to control the movements or consignments across national frontiers and ensure compliance with national laws. When adopted as a management philosophy it enables the Customs not only carry out its key

responsibilities effectively but also organize its resources and deploy them in a manner so as to improve its overall performance.

The main purpose of risk management is to decide whether the consignment requires physical examination, documentary checks and direct release. ERCA apply four levels of risks. Red channel - when the consignment is subjected to Physical Examination and documentary checking, Yellow channel - when the consignment is subjected to documentary checks, Green channel - Direct release of goods or the consignment is subjected to any checks before release and Blue- a type of risk channel which will be used like Authorized Economic Operators, Privileged companies, etc. and the consignment is not subjected for both physical examination or documentary checks.

When companies or individual seek to import goods into Ethiopia they typically have to pay a government imposed charged called a duty. Companies that engage in conduct designed to avoid paying custom duties and charges, is coming a fraud against the government. On the other hand there are goods that do not allowed to enter the country to protect the security of the people. The false Claims Act now presents a new approach to combating customs violations that are difficult for customs officers to catch.

The following are common types of fraud in ERCA.

- Undervaluation of goods upon entry into a customs territory
- Inaccurate country of origin marking
- Misclassification of goods
- Failure to pay anti-dumping or countervailing duties
- Claim by Reduce quantity of goods

Therefore, the basic question here is how could data mining help for risk management, risk channel assignment and related need in the case of Ethiopian revenue and customs authority (ERCA).

1.2. Statement of the Problem

Recently the increased complexity and volume of international trade, fueled by technological advances that have revolutionized global trading practices, have significantly affected the way Customs administrations carry out their responsibilities and organize their business operations.

Today Customs is required to provide extensive facilitation of trade while maintaining control over the international movement of goods, persons and means of transport. In seeking to achieve a balance between these goals, Customs has been moving away from traditional control methods and adopting new thinking and approaches to its tasks. Custom clearance efficiency and service level affects the overall trade efficiency, investment flow, employment level and even regional economic development. Laporte [46] stated that in many of developing countries, particularly in Sub-Sahara African, custom administration continue to carry out intrusive inspection on large number of containers even the detected incidence of fraud generally being less than 3% (as is the case in Benin, Côte d'Ivoire, Mali, and Senegal). As he explained, the existing selectivity methods used in risk management in these countries remain very much dependent on human judgment which represents a major shortcoming given moral dilemma. Therefore for optimal level of facilitation and control, customs administrations should aim for a reasonable and equitable balance between ensuring compliance and minimizing disruption and cost to legitimate trade through the adoption of a holistic risk-based compliance management approach[8].So in order to relieve the contradiction between the shortage of custom enforcement power and the growth of business volume, custom in different nations implement modern cargo clearance system centered in overall risk management [2]. It adopts scientific method like data mining and evaluates the risk dynamically to allocate limited manpower and resources for the optimal efficiency and performance [7].

Ethiopian customs has significantly reduced its physical intervention and inspection while maintaining its interest of revenue collection and national security by implementing risk management in its processing. Currently the physically inspected goods are less than 40% of the total consignment [9]. This is achieved through application of systematic but mainly manual risk management scheme. Currently, all risk management activities except the final triggering of the risk is performed manually or using general purpose statistical data analysis tools like MS excel. The primary difference between classical statistical method and data mining is in the size of the dataset. As the size of data increases highly it will create challenges that may not be sufficiently solved by statistical techniques alone. This process is obviously labor intensive and inconsistent as the analysis and data collection is done manually it is highly dependent on the effectiveness, skill and knowledge of the officers. On the other side, as the risk

management module in the current system is not robust enough to define risk rules as the complexity level the business demands .

Now ERCA start to replace the existing ASYCUDA++(Automated System for Customs Data) to web based Custom Management System(eCMS), still depends mainly on the existing ways of risk channel assignment based selectivity criteria which is worldwide standard from world custom organization. ERCA still no serious attempt to extend the application of data mining in establishing risk profile of customers and looking into how data mining would fit in the problem solving framework. Data Mining is a technology that use various technique to discover hidden knowledge from heterogeneous and distributed data stored in large databases, warehouses and other massive information repositories so as to find patterns that are valid, novel, useful, and understandable[6]. The data mining tools are capable of not only interpreting extremely large and complex datasets (on the order of thousands of data points or variables) but also extrapolating those relationships as trends and predictions. Data mining has received renewed attention recently because of the convergence of three important trends. First, with increasingly greater volumes of data being collected for various purposes, data mining tools are becoming a necessary part of interpreting the vast amount of data accumulated. Second, the availability and low cost of powerful multiprocessors provide the hardware necessary to manipulate large volumes of data in (near) real time. Finally, powerful new algorithms are being developed that are able to take advantage of the new processing technologies [12].

Some researchers from abroad have studied about custom risk management and custom related areas.

Laporte(46) studied about Risk Management Systems: Using Data Mining In Developing Countries' Customs Administrations.

Li [7] explains, detecting custom declaration frauds with limited examination of imported goods by available scarce resources is posing considerable challenge for effective examination of consignments.

Baştabak [69] investigates the prediction of tariff circumvention using data mining where a total of 13 attributes considered and first KNN classification algorithm then J48 decision tree algorithm were used to classify traders according to their risk level.

Yan-Hai and Lin-Yan [70] to solve the conflict between the number of total transactions and the number of inspection officers, a study was carried out on risk analysis of customs cargo declaration and Q-type cluster method was used to separate the declarations into groups based on their risk level.

Even some local the attempts made regarding use of data mining for ERCAlike Mamo [5] Application of data mining technology to support fraud protection: the case of Ethiopian revenue and custom authority. Master's thesis, Addis Ababa University. And Mezgeb and Berhanu[71] data mining to detect association pattern of customs administration data with market price and currency exchange rate in Ethiopia.

Therefore, the aim of this research is to investigate and demonstrate the potential applicability of data mining techniques in exploring custom risk channel assignment using the data collected from the ERCA information technology department.

.Hence, this research intends to get answers for the following research questions.

- How best can data mining technique support custom risk channel assignment?
- Which data mining algorithm perform best and identify highly predictive features for risk channel assignment?

1.3. Objectives

1.3.1. General Objective

The general objective of this research is to apply data mining techniques in constructing classification model for customs risk channel assignment.

1.3.2. Specific Objectives

In order to achieve the general objective, the following specific objectives were attempted in the present research:

- Assess the existing risk management challenges pertaining to customs risk channel assignment at ERCA
- Select and extract the data set required for analysis from the Ethiopian National Revenue risk management.

- Preprocess data in order to have a cleaned dataset that is suitable for any data mining algorithm.
- Assess and select different classification, algorithms to build a classification model that enable to assign clearance risk channel.
- Build a predictive model that enable to assign clearance risk channel.
- Evaluate the performance of the predictive model constructed by classification algorithms
- Interpret and analyze the results of the selected model with the help of domain expert.

1.4. Methodology

To understand the business operation better and in order to define the research problem properly, review of literature, informal interviews and discussion were conducted with customs professional. For this research, Domain experts are consulted to have insight into the problem domain in order to analysis the data and its structure. Data collection process was carried out from ERCA information technology management department in head quarter. The data was taken from custom risk management database. This phase helped for initial identification of attributes for consideration in the process risk management; On the other hand reviewing books, articles and research papers about data mining and its applications also has been carried out. The potential of data mining in general and particularly data mining applications in risk management has been investigated. By discussing issues of data protection and privacy with participant , the actual data from customs risk management system .This study will employ CRISP-DM process model for the data mining projects life cycle consisting of: understanding the problem domain, understanding of the data, preparation of the data, data mining, evaluation of the discovered knowledge, and use of the discovered knowledge.

The data employed in this research is collected from the risk management system which is managed centrally by ERCA information technology department. For this research, Domain experts are consulted to have insight into the problem domain in order to analysis the data and its structure. The data employed for this study has 18814 records and 9 attributes were selected by consulting domain expert and measure the rank of attributes using Weka. Three algorithms namely J48 Decision tree ,Naïve Bayes and K nearest neighbor was implemented.

1.5. Scope of the Study

The scope of this study was to investigate the potential applicability of classification and prediction techniques in exploring the prevalence of customs risk channel assignment. Though the tool selected for this study, which is Weka software, provides many classification and prediction techniques, this study was restricted to applying only J48 decision tree (DT) ,Naïve Bayes (NB and K Nearest Neighbor(KNN)) techniques.

The scope of this research was also restricted focus only the parts of ERCA related to import duty and tax. In addition, even if ERCA control and manage all import and export related transaction including government offices, international organization, investment manufacturing sector and diplomatic corps, this study considers only the import clearance unit which include commercial and investment section. These sections have the most important to collect duty and apply trade facilitation and control using risk management. Furthermore, as the data source is restricted to the ERCA's database related to commercial and investment transactions. For this research only red, yellow and Green risk channel included.

1.6. Significance of the Study

Data mining tools are capable of not only interpreting extremely large and complex datasets (on the order of thousands of data points or variables) but also extrapolating those relationships as trends and predictions. Data mining has received renewed attention recently because of the convergence of three important trends. First, with increasingly greater volumes of data being collected for various purposes, data mining tools are becoming a necessary part of interpreting the vast amount of data accumulated. Second, the availability and low cost of powerful multiprocessors provide the hardware necessary to manipulate large volumes of data in (near) real time. Finally, powerful new algorithms are being developed that are able to take advantage of the new processing technologies [12].

Though the primary goal and initiatives of this research has been for academic exercise, the discovered patterns (knowledge) can be used by Custom administrators to improve the quality of services. Predicting the risk channel ERCA officers will focus on high risk declaration in order to balance between trade facilitation and controls. So, successful use of the finding of

this research helps to avoid unnecessary delays and wastage of resources by concentrating customs control on high risk consignments and expediting the release of low risk consignments.

1.7 Thesis Organization

This thesis is organized into five chapters. The first chapter deals with the general overview of the study including background, statement of the problem, objectives of the research. The second is devoted to literature review of data mining technology as well as Customs risk management respectively. Chapter three explains the Research methodologies, decision trees, Naïve Bayes and K Nearest Neighbor classifiers as well as the Weka software, used in this study. Chapter four presents the experimentation phase of the study. It comprises training, building and validation of the models. Results of the experiment are also analyzed and interpreted. The last chapter is devoted for the final conclusions and recommendations based on the research findings.

CHAPTER TWO

LITRATURE REVIEW

2.1. DATA MINING

2.1.1. Introduction

It is estimated that the amount of data stored in the world's database grows every twenty months at a rate of 100% [15]. As the volume of data increases, the proportion of information in which people could understand decreases substantially. This reveals that the level of understanding of people about the data at hand could not keep pace with the rate of generation of data in various forms, which results in increasing information gap. Consequently, scholars begin to realize this bottleneck and to look into possible remedies. Current technological progress permits the storage and access of large amounts of data at virtually no cost. Although many times preached, the main problem in a current information centric world remains to properly put the collected raw data to use [18]. The true value is not in storing the data, but rather in our ability to extract useful reports and to find interesting trends and correlations, through the use of statistical analysis and inference, to support decisions and policies made by scientists and businesses [26]. To bridge the gap of analyzing large volume of data and extracting useful information and knowledge for decision making that the new generation of computerized methods known as Data Mining (DM) or Knowledge Discovery in Databases (KDD) has emerged in recent years.

Different scholars provided different definitions about DM. According to Berry and Linoff [16] DM is the process of extracting or “mining” knowledge from large amounts of data in order to discover meaningful patterns and rules. Witten and Frank [15] have also noted that DM is valuable to discover implicit, potentially useful information from huge data stored in databases via building computer programs that sift through databases automatically or semi-automatically, seeking meaningful patterns. DM involves the use of sophisticated data analysis tools to discover previously unknown, valid patterns and relationships in large datasets [25]. These tools can include statistical models, mathematical algorithms, and machine learning methods. Consequently, DM consists of more than collecting and managing data; it also includes analysis

and prediction and use of algorithms that improve their performance automatically through experience, such as neural networks or decision trees.

According to Han and Kamber [19], the major reason that DM has attracted a great deal of attention in the information industry in recent years is due to the wide availability of huge amounts of data and the imminent need for turning such data into useful information and knowledge. DM tools perform data analysis and may uncover important data patterns. The information and knowledge gained can be used for applications ranging from market analysis, fraud detection, and customer retention, to production control and science exploration [19].

DM is an interdisciplinary approach involving tools and models from statistics, artificial intelligence, pattern recognition, data visualization, optimization, information retrieval, high end computing, and others Guo [21]. DM methodology often can improve upon traditional statistical approaches for solving business solutions by finding additional, important variables, by identifying interaction among terms and detecting nonlinear relationships [22]. Models that predict relationships and behaviors more accurately lead to greater profits and reduced costs.

2.2 The Data Mining Process

DM requires massive collection of data to generate valuable information [19]. The data can range from simple numerical figures and text documents, to more complex information such as spatial data, multimedia data, and hypertext documents. Deshpande and Thakare[17] indicated that the data retrieval is simply not enough to take complete advantage of data. It requires a tool for automatic summarization of data, extraction of the essence of information stored, and the discovery of patterns in raw data. With the enormous amount of data stored in files, databases, and other repositories, it is increasingly important to develop powerful tool for analysis and interpretation of such data and for the extraction of interesting knowledge that could help in decision-making.

A typical DM process includes data acquisition, data integration, data exploration, model building, and model validation [17]. Both expert opinion and DM techniques play an important role at each step of this knowledge discovery process.

2.2.1 Data acquisition

The first step in DM is to select the types of data to be used. Although a target dataset has been created for discovery in some applications, DM can be performed on a set of variables or data samples in a larger database called training set to create and model while holding back some of the datasets (test dataset) for latter validation of the model.

2.2.2 Data preprocessing

Once the target data is selected, the data is then pre-processed for cleaning, scrubbing, and transforming to improve the effectiveness of discovery. During this pre-processing step, researchers remove the noise or outliers if necessary and decide on strategies for dealing with missing data fields and accounting for time sequence information or known changes. Then data is transformed to reduce the number of variables by converting one type of data to another (e.g., numeric ones into categorical) or deriving new attributes.

2.2.3 Building model

The third step of DM refers to a series of activities such as deciding on the type of DM operations, selecting the DM algorithms, and mining the data. First, the type of DM operation (classification, regression, clustering, association rule discovery, segmentation, and deviation detection) must be chosen. Based on the operations chosen for the application, an appropriate DM technique is then selected based on the nature of the knowledge to be mined. Once a DM technique is chosen, the next step is to select a particular algorithm within the DM technique chosen. Choosing a DM algorithm includes a method to search for patterns in the data, such as deciding which models and parameters may be appropriate and matching a particular DM technique with the overall objective of DM. After an appropriate algorithm is selected, the data is finally mined using the algorithm to extract novel patterns hidden in databases.

2.2.4 Interpretation and model evaluation

The fourth step of DM process is the interpretation and evaluation of discovered patterns. This task includes filtering the information to be presented by removing redundant or irrelevant patterns, visualizing graphically or logically the useful ones, and translating them into understandable terms by users. In the interpretation of results, the researcher determines and

resolves potential conflicts with previously known or decides redo any of the previous steps. The extracted knowledge is also evaluated in terms of its usefulness to a decision maker and to a business goal.

2.3 Data Mining Tasks

The DM tasks are of different types depending on the use of DM result [19]. Predictive modeling, descriptive modeling, exploratory data analysis, patterns and rules discovery, and retrieval by content are some of the DM tasks.

2.3.1 Predictive modeling

Predictive modeling permits the value of one variable to be predicted from the known values of other variables. Classification, Regression, Time series analysis, Prediction etc. are some examples of predictive modeling. As Tan et al. [23] indicated many of the DM applications are aimed to predict the future state of the data. Prediction is the process of analyzing the current and past states of the attribute and prediction of its future state. Classification is a technique of mapping the target data to the predefined groups or classes. It is a supervised learning because the classes are predefined before the examination of the target data. The regression involves the learning of function that maps data item to real valued prediction variable. In the time series analysis the value of an attribute is examined as it varies over time. In time series analysis the distance measures are used to determine the similarity between different time series, the structure of the line is examined to determine its behavior and the historical time series plot is used to predict future values of the variable.

2.3.2 Classification

Classification is the process of finding a model, which describes and distinguishes data classes or concepts, for the purpose of being able to use the model to predict the class of objects whose class label is unknown [19]. The derived model is based on the analysis of a set of training data (i.e., data objects whose class label is known). Classification problems aim to identify the characteristics that indicate the group to which each case belongs [25]. This pattern can be used both to understand the existing data and to predict how new instances will behave.

DM creates classification models by examining already classified data (cases) and inductively finding a predictive pattern [25]. According to the Two Crows Corporation, these existing cases may come from a historical database, such as people who have already undergone a particular medical treatment or moved to a new long distance service. They may come from an experiment in which a sample of the entire database is tested in the real world and the results used to create a classifier. For example, a sample of a mailing list would be sent an offer, and the results of the mailing used to develop a classification model to be applied to the entire database. Sometimes an expert classifies a sample of the database, and this classification is then used to create the model, which will be applied to the entire database. There are different algorithms that are used for classification purpose such as, decision tree, neural network, genetic algorithm, naïve bayes, etc.

2.3.2.1. Decision tree

A decision tree is a flow-chart-like tree structure where each internal node denotes a test on an attribute each branch represents an outcome of the test and leaf nodes represent classes or class distributions [19]. Decision trees are trees that classify instances by sorting them based on feature values [25]. They are a way of representing a series of rules that lead to a class or value. Each node in a decision tree represents a feature in an instance to be classified, and each branch represents a value that the node can assume. Instances are classified starting at the root node and sorted based on their feature values [26]. In DM, a decision tree is a predictive model, which can be used to represent both classifiers and regression models [24].

A decision tree model consists of a set of rules for dividing a large heterogeneous population into smaller, more homogeneous groups with respect to a particular target variable [24]. The target variable is usually categorical and the decision tree model is used either to calculate the probability that a given record belongs to each of the categories, or to classify the record by assigning it to the most likely class. Decision tree can also be used to estimate the value of continuous variable.

The decision node, branches and leaves are the basic components of a decision tree [27]. Depending on a decision tree algorithm, each node may have two or more branches. For example, CART (Classification and Regression Tree) generates trees with only two branches at

each node. Such a tree is called a binary tree. When more than two branches are allowed it is called a multi-way tree. Each branch will lead either to another decision node or to the bottom of the tree, called a leaf node. By navigating the decision tree you can assign a value or class to a case by deciding which branch to take, starting at the root node and moving to each subsequent node until a leaf node is reached. Each node uses the data from the case to choose the appropriate branch.

Decision trees are generated from training data in a top down general to specific direction [27]. The initial state of a decision tree is the root node that is assigned all the examples from the training set. If it is the case that all examples belong to the same class then no further decisions need to be made to partition the examples and the solution is complete. If examples at this node belong to two or more classes then a test is made at the node that will result in a split. The process is recursively repeated for each of the new intermediate nodes until a completely discriminating tree is obtained. A decision tree at this stage is potentially an over-fitted solution i.e. it may have components that are too specific to noise and outliers that may be present in the training data. As Apte and Weiss [27] indicated, to relax this over-fitting most decision tree methods go through a second phase called pruning that tries to generalize the tree by eliminating sub trees that seem too specific. Error estimation techniques play a major role in tree pruning. Most modern decision tree modeling algorithms are a combination of a specific type of a splitting criterion for growing a full tree and a specific type of a pruning criterion for pruning tree.

The attractiveness of decision trees is due to the fact that, in contrast to neural networks, decision trees represent rules [24]. Rules can readily be expressed so that humans can understand them or even directly used in a database access language like SQL so that records falling into a particular category may be retrieved. Decision tree has its own properties.

The following are some of them:

- Learns with positive and negative examples
- Noise tolerant
- General-to-specific search (reverse for pruning)

- Follows Divide-and-Conquer strategy, which has weaknesses of fracturing and diminishing training data.
- Learns discriminating rules

Decision tree can be implemented with several algorithms. Some of them are J48, ID3, C4.5, CART, etc. J48 is an implementation of C4.5 release 8(3) that produces decision trees [29]. This is a standard algorithm that is used for machine learning. C4.5 is a decision tree-learning algorithm that builds upon the ID3 algorithm as indicated by Lavesson [30]. Amongst other enhancements (compared to the ID3 algorithm) the C4.5 algorithm includes different pruning techniques and can handle numerical and missing attribute values. C4.5 avoids over fitting the data by determining a decision tree, it handles continuous attributes, is able to choose an appropriate attribute selection measure, handles training data with missing attribute values and improves computation efficiency. C4.5 builds the tree from a set of data items using the best attribute to test in order to divide the data item into subsets and then it uses the same procedure on each sub set recursively. The main problem in decision tree is deciding the attribute, which will best partition the data into various classes [29]. The ID3 algorithm is useful to solve this problem.

2.3.2.1.1. Constructing Decision Tree

In chapter 2, section 2.3.1 decision tree presented in detail. So in this section we only describe about constructing decision tree. Decision tree programs construct a decision tree from a set of training sets. The main focus of a decision tree growing algorithm is selecting which attribute to test at each node in the tree. The decision trees are constructed in a top-down fashion by choosing the best and most appropriate attribute each time. An information-theoretic measure is used to evaluate features and select the best attribute, which provides an indication of the “classification power” of each feature. In information theoretic measure, the concept of Entropy and Information Gain (IG) are used by the algorithm. Information gain (IG) is measured as the amount of the entropy (S) difference when an attribute contributes the additional information about the class, whereas Entropy(S) is the sum of the probability of each label times the log probability of that same label [53]. In order to define information gain precisely, we need to define a measure commonly used in information theory, called entropy, which characterizes the

impurity of an arbitrary collection of examples. Given a set S , containing only positive and negative examples of some target concept (a 2 class problem), the entropy of set S relative to this simple, binary classification is defined as:

$$\text{Entropy}(s) = -Pp \log_2 Pp - Pn \log_2 Pn \quad (1)$$

Where pp is the proportion of positive examples in S and pn is the proportion of negative examples in S . The entropy is 0 if all members of S belong to the same class. For example, if all members are positive ($pp= 1$), then pn is 0, and $\text{Entropy}(S) = -1 \cdot \log_2(1) - 0 \cdot \log_2 0 = -1 \cdot 0 - 0 \cdot \log_2 0 = 0$. The entropy is 1 (at its maximum) when the collection contains an equal number of positive and negative examples. If the collection contains unequal numbers of positive and negative examples, the entropy is between 0 and 1.

The above computation of entropy is in the special case where the target classification is binary. When the target attribute takes values more than two, say k different values, then the entropy of S relative to this k -wise classification is defined as:

$$\text{Entropy}(s) = \sum_{i=1}^k -P_i \log_2 P_i \quad (2)$$

Where p_i is the proportion of S belonging to class i . if the target attribute can take on k possible values, the maximum possible entropy is $\log_2 k$. As entropy is a measure of the impurity in a collection of training examples, information gain is a measure of the effectiveness of an attribute in classifying the training data. It is simply the expected reduction in entropy caused by partitioning/splitting the examples according to this attribute. Say, this attribute is considered to be A , the information gain (S, A) of an attribute A , Relative to a collection of examples S , is defined as:

$$\text{Gain}(S, A) = \text{Entropy}(S) - \sum_{v \in \text{Values}(A)} \frac{|S_v|}{|S|} \text{Entropy}(S_v) \quad (3)$$

where $\text{Values}(A)$ is the set of all possible values for attribute A , and S_v is the subset of S for which attribute A has value v (i.e., $S_v = \{s \in S \mid A(s) = v\}$). Note the first term in the equation for Gain is just the entropy of the original collection S and the second term is the expected value of the entropy after S is partitioned using attribute A . The expected entropy described by this second term is the sum of the entropies of each subset S_v , weighted by the fraction of examples

$\frac{|SV|}{|S|}$ that belong to S_v . Gain (S,A) is therefore, the expected reduction in entropy caused by knowing the value of attribute A. In another way, Gain(S,A) is the information provided about the target attribute value, given the value of some other attribute A. The value of Gain(S,A) is the number of bits saved when encoding the target value of an arbitrary member of S, by knowing the value of attribute A. According to Hamilton et al. [61], the process of selecting a new attribute and partitioning the training examples is repeated for each non-terminal descendant node, this time using only the training examples associated with that node. Attributes that have been incorporated higher in the tree are excluded, so that any given attribute can appear at most once along any path through the tree.

2.3.2.1.2. Rule induction

Rule induction is the process of extracting useful ‘if then’ rules from data based on statistical significance. A Rule based system constructs a set of if-then-rules. It has the form:

IF conditions THEN conclusion

This kind of rule consists of two parts. The rule antecedent (the IF part) contains one or more conditions about value of predictor attributes whereas the rule consequent (THEN part) contains a prediction about the value of a goal attribute. An accurate prediction of the value of a goal attribute will improve decision-making process. IF-THEN prediction rules are very popular in data mining; they represent discovered knowledge at a high level of abstraction [54].

2.3.3 Naive Bayes (NB) Classifier

Naive Bayes classifier is a term in Bayesian statistics dealing with a simple probabilistic classifier based on applying Bayes' theorem with strong (naive) independence assumptions [55]. That is, there are no dependence relationship among the attributes given the value of the class variable [52]. Despite this strong assumption, the algorithm tends to perform well in many class prediction scenarios. Experimental studies suggest that Naive Bayes tends to learn more rapidly than most induction algorithms. Given a new instance, the classifier estimates the probability that the instance belongs to a specific class, based on the product of the individual conditional probabilities for the feature values in the instance. The exact calculation uses Bayes theorem and this is the reason why the algorithm is called a Bayes classifier [56]. In simple

terms, a Naive Bayes classifier assumes that the presence (or absence) of a particular feature of a class is unrelated to the presence (or absence) of any other feature. Depending on the precise nature of the probability model, Naive Bayes classifiers can be trained very efficiently in a supervised learning setting. In spite of their Naive design, Naive Bayes classifiers often work much better in many complex real-world situations than one might expect. The Naive Bayesian classifier is fast and incremental, and can deal with discrete and continuous attributes with excellent performance in real life problems. It has capability to solve also non-linear problems while retaining all advantages of Naive Bayes [55]. Learning a Naive Bayes classifier is straightforward and involves estimating the probability of attribute values within each class from the training instances. Probabilities are estimated by counting the frequency of each discrete attribute values. For numeric attributes, it is common practice to use the normal distribution [60]. Given a new instance, the classifier estimates the probability that the instance belongs to a specific class, based on the product of the individual conditional probabilities for the feature values in the instance. The exact calculation uses Bayes theorem and this is the reason why the algorithm is called a Bayes classifier. The main advantage of using Naive Bayes is that they are probabilistic models, robust to noise found in real data. The Naive Bayes classifier presupposes independence of the attributes used in classification. However, it was tested on several artificial and real data sets, showing good performances even when strong attribute dependences are present. In addition, the Naive Bayes classifier can outperform other powerful classifiers when the sample size is small [59].

3.3.3.1. Naive Bayesian Classification Algorithm

The Naive Bayes classifier can predict class membership probabilities, such as the probability that a given sample belongs to a particular class [66]. This is performed by Naive Bayes classifier as follows. Let T be a training set of samples, each with their class labels. There are k classes, C_1, C_2, \dots, C_k . Each sample is represented by an n -dimensional vector, $X = \{x_1, x_2, \dots, x_n\}$, depicting n measured values of the n attributes, A_1, A_2, \dots, A_n , respectively.

Given a sample X , the classifier will predict that X belongs to the class having the highest a posteriori probability, conditioned on X . That is X is predicted to belong to the class C_i if and only if $P(C_i|X) > P(C_j|X)$ for $1 \leq j \leq m, j \neq i$. Thus we find the class that maximizes $P(C_i|X)$.

The class C_i for which $P(C_i|X)$ is maximized is called the maximum posteriori hypothesis. By Bayes' theorem

$$P(C_i|X) = \frac{P(X|C_i)P(C_i)}{P(X)} \quad (1)$$

As $P(X)$ is the same/constant for all classes, only $P(X|C_i)P(C_i)$ need be maximized. If the class priori probabilities, $P(C_i)$, are not known, then it is commonly assumed that the classes are equally likely, that is, $P(C_1) = P(C_2) = \dots = P(C_k)$, and we would therefore maximize $P(X|C_i)$. Otherwise we maximize $P(X|C_i)P(C_i)$. The class prior probabilities may be estimated by $P(C_i) = \frac{P(C_i,T)}{|T|}$, where $|C_i,T|$ is the number of training samples of class C_i in T .

Given data sets with many attributes, it would be extremely computationally expensive to compute $P(X|C_i)$. In order to reduce computation in evaluating $P(X|C_i)$, the naïve assumption of class conditional independence is made. This presumes that the values of the attributes are conditionally independent of one another, given the class label of the sample (i.e., there are no dependence relationships among the attributes). Thus,

$$\begin{aligned} P(X|C_i) &= \prod_{k=1}^n P(X_k|C_i) \quad (2) \\ &= P(x_1|C_i) \times P(x_2|C_i) \times \dots \times P(x_k|C_i) \end{aligned}$$

The probabilities $P(x_1|C_i)$, $P(x_2|C_i)$, \dots , $P(x_k|C_i)$ can easily be estimated from the training set. Recall that here x_k refers to the value of attribute A_k for sample X .

(a) If A_k is categorical, then $P(x_k|C_i)$ is the number of samples of class C_i in T having the value x_k for attribute A_k , divided by (C_i, T) , the number of sample of class C_i in T . (b) If A_k is continuous-valued, then we typically assume that the values have a Gaussian distribution with a mean μ and standard deviation - defined by

$$g(x, \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x - \mu)^2}{2\sigma^2}} \quad (3)$$

So that

$$P(x_k|C_i) = g(x_k, \mu_{C_i}, \sigma_{C_i})$$

We need to compute μ_{C_i} and σ_{C_i} , which are the mean and standard deviation of values of attribute A_k for training samples of class C_i . In order to predict the class label of X , $P(X|C_i)P(C_i)$ is evaluated for each class C_i . The classifier predicts that the class label of sample X is C_i if and only if it is the class that maximizes $P(X|C_i)P(C_i)$.

2.3.4. The K-nearest Neighbor (K-NN) Algorithm

The K-Nearest Neighbor Algorithm is the simplest of all machine learning algorithms. It is based on the principle that the samples that are similar, generally lies in close vicinity [57]. K-Nearest Neighbor is instance based learning method. Instance based classifiers are also called lazy learners as they store all of the training samples and do not build a classifier until a new, unlabeled sample needs to be classified [58]

The K-NN algorithm is proposed to find out k training samples that are closest to the target object in the training set. Furthermore, determine the dominant category from the k training samples; then, assign this dominant category to the target object, where k is the number of training samples.

Therefore, the principal mechanism of the K-NN algorithm is that all samples have the same characteristics while they are classified in the same category in a feature space, which the category contains the k most neighboring samples. In determining the classification decision, the method determines the category to which the sample belongs only according to the category of the nearest one or several samples. In addition, the K-NN algorithm is only relevant to a very small number of adjacent samples in category decision making. Since the K-NN algorithm mainly relies on the surrounding limited adjacent samples, rather than relying on the method of discriminant domain method to determine the category, thus the K-NN algorithm is more suitable than other methods for the pending sample sets where the class domain crosses or overlaps more. The idea of the K-NN algorithm is demonstrated in Figure 1. In which, X_u belongs to the category (w_1) because four neighboring samples belong to w_1 , only one neighboring sample belongs to w_3 .

The specified implementation process of the K-NN algorithm contains the following six steps,

- 1) Select the k value;

- 2) calculate the distance between the point in the known category data set and the current point;
- 3) Sort in increasing order of distance;
- 4) Select k points with the smallest distance from the current point;
- 5) determine the frequency of occurrence of the category in which k points are located;
- 6) Return to the category with the highest frequency of occurrence of the first k points as the predicted classification of the current point.

The K-NN algorithm needs to calculate the distance between the forecasted data point and the known data point, so as to select the nearest k labeled data, y_1, y_2, \dots, y_k , where y_1 represents the known data point closest to the forecasted point; y_2 represents the known data point that is the second closest to the forecasted point, and so on. Therefore, the short-term load forecasting can be conducted by the K-NN algorithm regression as Equation (1),

$$S_i = \frac{1}{k} * \sum_{j=1}^k S_{y_j} \quad (1)$$

Where s_i represents the i th forecasted value, which is the average value of s_{y_j} ($j = 1, 2, \dots, k$); s_{y_j} represents the forecasted value of the j th closest known data point (y_j).

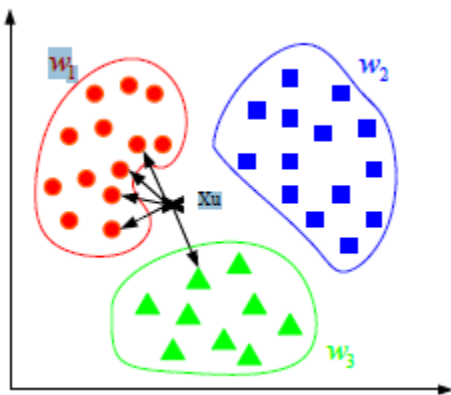


Figure 2.1: KNN proximate algorithm map

Neural networks

An artificial neural network (ANN), often just called a "neural network" (NN), is a mathematical model or computational model based on biological neural networks, in other words, they imitate the way the human brain learns and use rules inferred from data patterns to construct hidden layers of logic for analysis [29].

Neural networks constitute the most widely used technique in DM. As Hajek [28] stated, a neural network is a massively parallel-distributed processor that has a natural tendency for storing experiential knowledge and making it available for use. It resembles the brain in two respects:

- I. Knowledge is acquired by the network through a learning process
- II. Interneuron connection strengths known as synaptic weights are used to store the knowledge.

A neural network is first and foremost a graph, with patterns represented in terms of numerical values attached to the nodes of the graph and transformations between patterns achieved via simple message-passing algorithms [30]. Generally, a neural network can be described as a

directed graph in which each node performs a transfer function of the form

$$y_i = f\left(\sum_{j=1}^n W_{ij} X_j - Q_i\right)$$

Where y_i is the output of the node i , x_j is the j th input to the node, and W_{ij} is the connection weight between nodes i and j . Q_i is the threshold (or bias) of the node. Certain of the nodes in the graph are generally distinguished as being input nodes or output nodes, and the graph as a whole can be viewed as a representation of a multivariate function linking inputs to outputs. Numerical values (weights) are attached to the links of the graph, which parameterize the input/output function and allowing it to be adjusted via a learning algorithm.

Neural network topologies can be divided into feed forward and recurrent classes according to their connectivity [30]. The feed forward neural network was the first and arguably simplest type of artificial neural network devised. In this network, the information moves in only one direction, forward, from the input nodes, through the hidden nodes (if any) and to the output nodes. A neural network is feed forward if there exists a method, which numbers all the nodes in the network such that there is no connection from a node with a large number to a node with a

smaller number. All the connections are from nodes with small numbers to nodes with larger numbers. A neural network is recurrent if such a numbering method does not exist. Contrary to feed forward networks, recurrent neural networks (RNs) are models with bidirectional data flow. While a feed forward network propagates data linearly from input to output, RNs also propagate data from later processing stages to earlier stages.

Learning in ANN's (Artificial neural network) can roughly be divided into supervised, unsupervised, and reinforcement learning. Supervised learning or Associative learning is based on direct comparison between the actual output of an ANN and the desired correct output, also known as the target output. Reinforcement learning is a special case of supervised learning where the exact desired output is unknown. It is based only on the information of whether or not the actual output is correct. Unsupervised learning or Self-organization is solely based on the correlations among input data. No information on "correct output" is available for learning.

According to Larose [31] there are two general categories of neural net algorithms: supervised and unsupervised. Supervised neural net algorithms such as Back propagation and Perceptron require predefined output values to develop a classification model. Among the many algorithms, Back propagation is the most popular supervised neural net algorithm [19]. Unsupervised neural net algorithms such as ART do not require predefined output values for input data in the training set and employ self-organizing learning schemes to segment the target dataset.

For organizations with a great depth of statistical information, ANNs are ideal because they can identify and analyze changes in patterns, situations, or tactics far more quickly than any human mind, as indicated by Guo [21]. Although the neural net technique has strong representational power, interpreting the information encapsulated in the weighted links can be very difficult. One important characteristic of neural networks is that they are opaque, which means there is not much explanation of how the results come about and what rules are used. Therefore, some doubt is cast on the results of the DM.

2.4. Descriptive modelling

The descriptive model identifies the patterns or relationships in data and explores the properties of the data examined [17]. It describes all the data, it includes models for overall probability

distribution of the data, partitioning of the p-dimensional space into groups and models describing the relationships between the variables. Clustering, Association rule discovery, Sequence discovery, Summarization, etc. are some of the examples. Clustering is similar to classification except that the groups are not predefined, but are defined by the data alone [19].

Summarization is the technique of presenting the summarized information from the data. The association rule finds the association between the different attributes. Association rule mining is a two-step process: Finding all frequent item sets, Generating strong association rules from the frequent item sets. Sequence discovery is a process of finding the sequence patterns in data. This sequence can be used to understand the trend.

2.4.1. Clustering

Clustering is a DM (machine learning) technique that finds similarities between data according to the characteristics found in the data and group's similar data objects into one cluster. The objective of clustering is to distribute cases (people, objects, events etc.) into groups, so that the degree of association can be strong between members of the same cluster and weak between members of different clusters [24]. Clustering techniques are employed to segment a database into clusters, each of which shares common and interesting properties [25]. The purpose of segmenting a database is often to summarize the contents of the target database by considering the common characteristics shared in a cluster. Clusters are also created to support the other types of DM operations, e.g. link analysis within a cluster Guo [21]. Clustering tools assign groups of records to the same cluster if they have something in common, making it easier to discover meaningful patterns from the dataset [27]. Clustering often serves as a starting point for some supervised DM techniques or modeling.

Clustering is one of the most useful tasks in DM process for discovering groups and identifying interesting distributions and patterns in the underlying data. Clustering problem is about partitioning a given dataset into groups such that the data points in a cluster are more similar to each other than points in different clusters [28]. For example, segmenting existing insurance policyholders into groups and associating a distinct profile with each group can help future rate making strategies. Clustering methods perform disjoint cluster analysis on the basis of Euclidean

distances computed from one or more quantitative variables and seeds that are generated and updated by the algorithm. You can specify the clustering criterion that is used to measure the distance between data observations and seeds. The observations are divided into clusters such that every observation belongs to at most one cluster.

Clustering studies are also referred to as unsupervised learning and/or segmentation. Unsupervised learning is a process of classification with an unknown target, that is, the class of each case is unknown. The aim is to segment the cases into disjoint classes that are homogenous with respect to the inputs. Clustering studies have no dependent variables. You are not profiling a specific trait as in classification studies. Cluster analysis is related to other techniques that are used to divide data objects into groups. For instance, clustering can be regarded as a form of classification in that it creates a labeling of objects with class (cluster) labels. Clustering techniques are heuristic in nature [27]. Almost all techniques have a number of arbitrary parameters that can be “adjusted” to improve results. Clustering techniques can be divided broadly into two approaches:

- Partitioning clustering approach: - Construct various partitions and then evaluate them by some criterion, e.g., minimizing the sum of square errors. K-Means clustering and expectation maximization (EM) clustering are the two methods of partitioning clustering.
- Hierarchical clustering approach: - Create a hierarchical decomposition of the set of data (or objects). It can be visualized as a dendrogram; a tree like diagram that records the sequences of merges or splits.

Hierarchical clustering approach is further subdivided into agglomerative and divisive.

a) Agglomerative: Start with the points as individual clusters and, at each step, merge the most similar or closest pair of clusters. It is a Bottom Up clustering technique. This requires a definition of cluster similarity or distance.

b) Divisive: Start with one, all-inclusive cluster and, at each step, split a cluster until only singleton clusters of individual points remain. It is a Top Down clustering technique. In this case, we need to decide, at each step, which cluster to split and how to perform the split.

K-Means clustering algorithm

The k-Means algorithm is very widely used to produce clustering of data, due to its simplicity and speed [32]. It is a simple iterative method to partition a given dataset into a user-specified number of clusters [32]. The idea is based around clustering items using centroids. These are points in the metric space that define the clusters. Each centroid defines a single cluster, and each point from the data is associated with the cluster defined by its closest centroid. Techniques for selecting these initial seeds include sampling at random from the dataset, setting them as the solution of clustering a small subset of the data or perturbing the global mean of the data k times. Then the algorithm iterates between two steps till convergence [32]. The first step is Data Assignment. Here each data point is assigned to its closest centroid, with ties broken arbitrarily. This results in a partitioning of the data. The second step is Relocation of “means”. Each cluster representative is relocated to the center (mean) of all data points assigned to it. If the data points come with a probability measure (weights), then the relocation is to the expectations (weighted mean) of the data partitions.

The K-Means algorithm is simple, easily understandable and reasonably scalable, and can be easily modified to deal with streaming data. However, one of its drawbacks is the requirement for the number of clusters, K , to be specified before the algorithm is applied [33].

2.4.2 Association rule discovery

Association rule mining is to find out association rules that satisfy the predefined minimum support and confidence from a given database [42]. The problem is usually decomposed into two sub problems. One is to find those Item sets whose occurrences exceed a predefined threshold in the database; those item sets are called frequent or large item sets. The second problem is to generate association rules from those large item sets with the constraints of minimal confidence. Association rule aims to extract interesting correlations, frequent patterns, associations or casual structures among sets of items in the transaction databases or other data repositories [42]. Given a collection of items and a set of records containing some of these items, association discovery techniques discover the rules to identify affinities among the collection of items as reflected in the examined records [21]. For example, 65 percent of records that contain item A also contain item B. An association rule uses measures called "support" and "confidence" to represent the strength of association. The percentage of occurrences, 65 percent in this case, is the confidence

factor of the association. According to Guo [21], the efficiency with which association discovery algorithms can organize the events that make up an association or transaction is one of the differentiators among the association discovery algorithms. Association rules are widely used in various areas such as telecommunication networks, market and risk management, inventory control etc. There are a variety of algorithms to identify association rules. The most widely used association rule algorithms are Apriori and FP-growth tree. Apriori is an influential algorithm for finding frequent item sets using candidate generation [31]. Frequent-pattern tree, or FP-tree in short is an extended prefix-tree structure storing crucial, quantitative information about frequent patterns. FP-growth method is an efficient and scalable mining for both long and short frequent patterns, and is about an order of magnitude faster than the Apriori algorithm [19].

2.5. Types of Data Mining Systems

There are many DM systems available or being developed. Some are specialized systems dedicated to a given data source or are confined to limited DM functionalities, other are more versatile and comprehensive. DM systems can be categorized according to various criteria [17]. DM systems can be classified according to the type of data source mined. This classification is according to the type of data handled such as spatial data, multimedia data, time-series data, text data, World Wide Web, etc. The other Classification of DM systems are according to the data model. This classification is based on the data model involved such as relational database, object-oriented database, data warehouse, transactional database, etc. Further Classification of DM systems are according to the kind of knowledge discovered. This classification of DM systems based on the kind of knowledge discovered or DM functionalities, such as characterization, discrimination, association, classification, clustering, etc. Some systems tend to be comprehensive systems offering several DM functionalities together. Finally, DM systems can be classified according to mining techniques used. This classification is according to the data analysis approach used such as machine learning, neural networks, genetic algorithms, statistics, and visualization. The classification can also take into account the degree of user interaction involved in the DM process such as query-driven systems, interactive exploratory systems, or autonomous systems [17]. According to Deshpande and Thakare[17], a comprehensive system

would provide a wide variety of DM techniques to fit different situations and options, and offer different degrees of user interaction.

2.6. The Data Mining Models

There are different DM process model standards. The six step Cios et al. (2000) model, KDD process (Knowledge Discovery in Databases), CRISP-DM (Cross Industry Standard Process for Data Mining), and SEMMA (Sample Explore Modify Model Assess), are some of the models that are used in different DM projects.

2.6.1. The six step Cios model

This model was developed, by adopting the CRISP-DM model to the needs of academic research community. The model consists of six steps [50].

1. Understanding of the problem domain: In this step one works closely with domain experts to define the problem and determine the research goals, identify key people, and learn about current solutions to the problem. A description of the problem including its restrictions is done. The research goals then need to be translated into the DM goals, and include initial selection of the DM tools.

2. Understanding of the data: This 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 DM goals. Data needs to be checked for completeness, redundancy, missing values, plausibility of attribute values, etc.

3. Preparation of the data: This 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 will be used as input for DM tools of step 4, 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).The result would be new data records, meeting specific input requirements for the planned to be used DM tools.

4. **Data mining:** This is another key step in the knowledge discovery process. Although it is the DM tools that discover new information, their application usually takes less time than data preparation. This step involves usage of the planned DM tools and selection of the new ones. DM tools include many types of algorithms, such as neural networks, clustering, preprocessing techniques, Bayesian methods, machine learning, etc. This step involves the use of several DM tools on data prepared in step 3. First, the training and testing procedures are designed and the data model is constructed using one of the chosen DM tools; the generated data model is verified by using testing procedures.

5. **Evaluation of the discovered knowledge:** This 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 DM process may be revisited to identify which alternative actions could have been taken to improve the results.

6. **Using the discovered knowledge:** This step is entirely in the hands of the owner of the database. It consists of planning where & how the discovered knowledge will be used. The application area in the current domain should be extended to other domains.

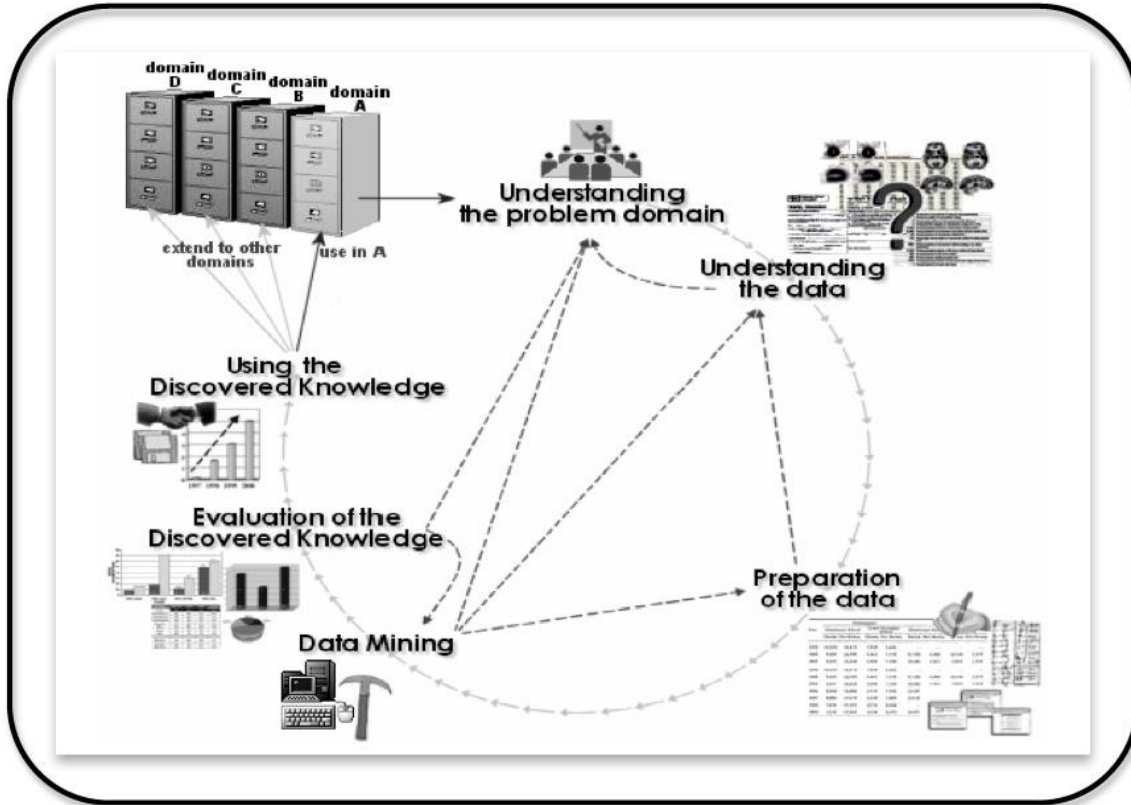


Figure 2.2: The Six Step Cios et al. (2000) process model

2.6.2 The KDD process model

KDD process is the process of using DM methods to extract what is deemed knowledge according to the specification of measures and thresholds, using a database along with any required preprocessing, sub sampling, and transformation of the database as presented by Azevedo and Santos [35]. It is an interactive and iterative process, comprising a number of phases requiring the user to make several decisions. Generally, there are five steps in the KDD process [25]

1. Data selection: This stage consists on creating a target dataset, or focusing on a subset of variables or data samples, on which discovery is to be performed. The data relevant to the analysis is decided on and retrieved from the data collection.

2. Data pre-processing: This stage consists on the target data cleaning and preprocessing in order to obtain consistent data

3. Data transformation: It is also known as data consolidation; in this phase the selected data is transformed into forms appropriate for the mining procedure. This stage consists on the transformation of the data using dimensionality reduction or transformation methods

4. Data mining: It is the crucial step in which clever techniques are applied to extract potentially useful patterns. It consists on the searching for patterns of interest in a particular representational form, depending on the DM objective.

5. Interpretation/Evaluation: This stage consists on the interpretation and evaluation of the mined patterns.

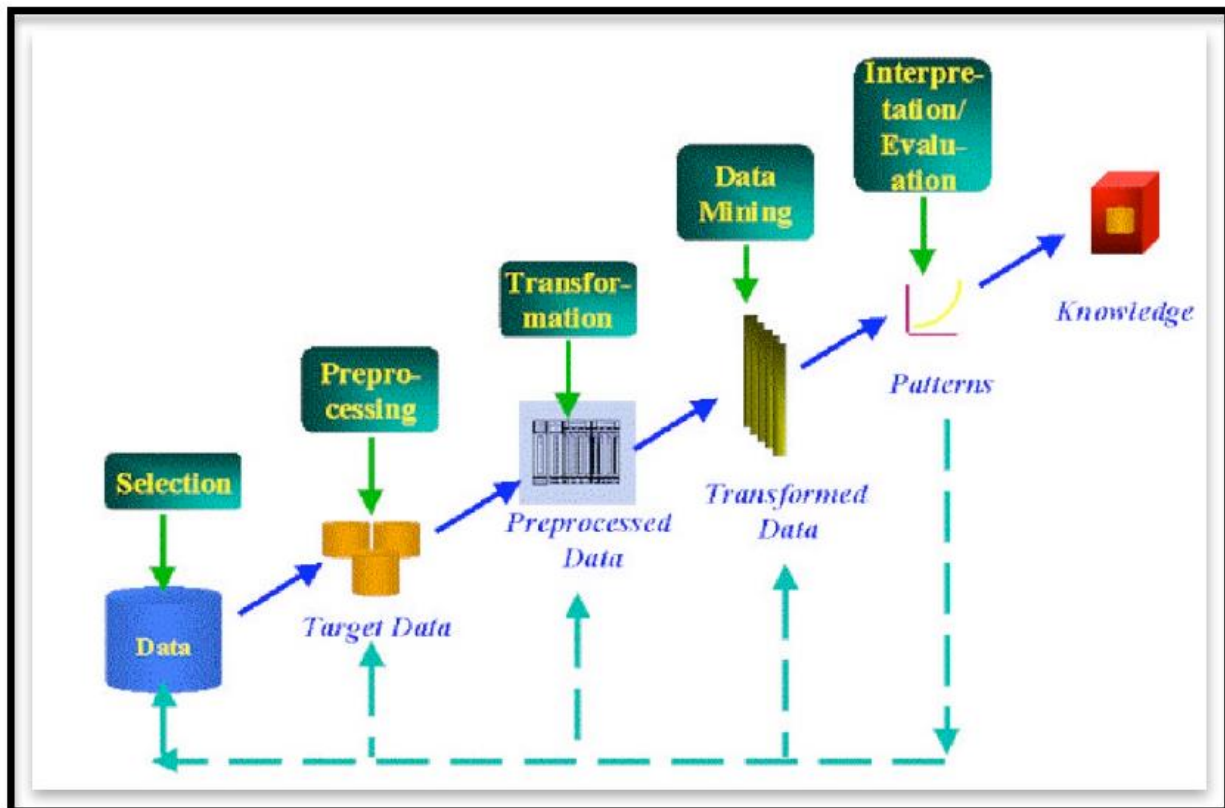


Figure 2.3: The KDD Process

As indicated above, a KDD process involves preprocessing data, choosing a data-mining algorithm, and post processing the mining results. There are very many choices for each of these stages, and non-trivial interactions between them. Therefore both novices and DM specialists need assistance in KDD processes.

2.6.3. The CRISP-DM process

CRISP-DM (Cross Industry Standard Process for Data Mining), process model was first established by four companies in the late 1990s [35]. These were Integral Solutions Ltd. (a provider of commercial DM solutions), NCR (a database provider), DaimlerChrysler (an automobile manufacturer), and OHRA (an insurance company).

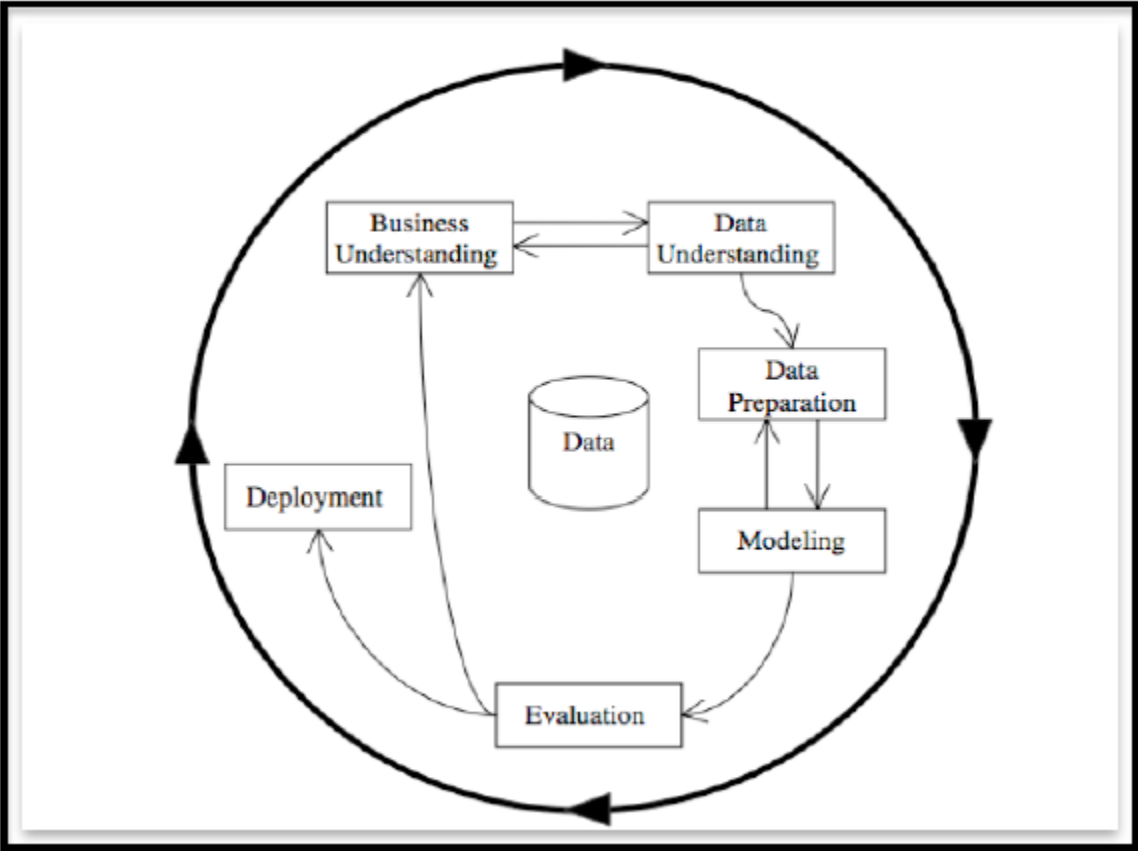


Figure 2.4: The CRISP-DM Process

According to CRISP-DM the life cycle of a data-mining project consists of six phases. The sequence of the phases in the CRISP-DM process is not rigid. Moving back and forth between

different phases is always required. It depends on the outcome of each phase. The CRISP-DM process has six stages.

1. **Business understanding:** This phase focuses on understanding the project objectives and requirements from a business perspective, then converting this knowledge into a DM problem definition and a preliminary plan designed to achieve the objectives.

2. **Data understanding:** It starts with an initial data collection, to get familiar with the data, to identify data quality problems, to discover first insights into the data or to detect interesting subsets to form hypotheses for hidden information.

3. **Data preparation:** It covers all activities to construct the final dataset from the initial raw data.

4. **Modeling:** In this phase, various modeling techniques are selected and applied and their parameters are calibrated to optimal values.

5. **Evaluation:** In this stage the model is thoroughly evaluated and reviewed. The steps executed to construct the model to be certain it properly achieves the business objectives. At the end of this phase, a decision on the use of the DM results should be reached.

6. **Deployment:** 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 customer can use it.

2.6.4. The SEMMA Process

Data mining can be viewed as a process rather than a set of tools, and the acronym SEMMA stands for (Sample, **E**xplore, **M**odify, **M**odel, **and** **A**ssess) refers to a methodology that clarifies this process. The SAS Institute considers a cycle with five stages for the process:

Sample – this stage consists on sampling the data by extracting a portion of a large data set big enough to contain the significant information, yet small enough to manipulate quickly.

Explore - this stage consists on the exploration of the data by searching for unanticipated trends and anomalies in order to gain understanding and ideas.

Modify - this stage consists on the modification of the data by creating, selecting, and transforming the variables to focus the model selection process.

Model - this stage consists on modeling the data by allowing the software to search automatically for a combination of data that reliably predicts a desired outcome.

Assess - this stage consists on assessing the data by evaluating the usefulness and reliability of the findings from the DM process and estimate how well it performs.

The SEMMA process offers an easy to understand process, allowing an organized and adequate development and maintenance of DM projects. It thus confers a structure for its conception, creation and evolution, helping to present solutions to business problems as well as to find de DM business goals [35].

2.6.5. Comparison of SEMMA, KDD and CRISP-DM

Today, research efforts have been focused on proposing new models, rather than improving design of a single model or proposing a generic unifying model. Despite the fact that most models have been developed in isolation, a significant progress has been made. The subsequent models provide more generic and appropriate descriptions. Most of them are not tied specifically to academic or industrial needs, but rather provide a model that is independent of a particular tool, vendor, or application [18]. Santos &Azevedo [35] have summarized the association of the three most popular process model steps as shown in Table 2.1

KDD	SEMMA	CRISP-DM
Pre KDD	-----	Business understanding
Selection	Sample	Data Understanding
Pre processing	Explore	Data preparation
Transformation	Modify	Modeling
Data mining	Model	Evaluation
Interpretation/Evaluation	Assessment	Deployment
Post KDD	-----	

Table 2.1: Comparison of Data mining process model

2.7. Data mining in customs

Data mining is a relatively new field that enables finding interesting knowledge (patterns, models and relationships) in the data sets. It is the most essential part of the knowledge discovery process and has the potential to predict events or to assist in analysis. Data mining has elements of databases, statistics, artificial intelligence and machine learning. Data mining software allows users to analyze large data sets 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 databases. Customs might have such large databases, for instance from the operational import and export systems.

Data mining techniques help us to perform better in risk identification, analyzing and preparing for audits or checks. With data mining Customs can gain time and can perform less checks with the same or even better results. Some of the applicable data mining techniques are shown below. With time series analysis, Customs can look for not known patterns, in order to discover new risks. Explaining a significant increase of trade for a certain product from a certain country of origin makes it possible that new risks will be found. This technique can be used in the stage of risk identification. With predictive modeling, one is able to produce estimations of unknown dependable variables at present or in the past. More commonly, with this method you can predict whether a situation will occur or has occurred. Techniques often used for this method are regression analysis, decision trees and neural networks. Those techniques are also suitable for making selection rules or for improving existing selection rules. With selection rules Customs administrations can select import and export shipments for a check.

2.7.1. Overview of risk management

“The United Nations Conference on Trade and Development (UNCTAD) estimates that the average customs transaction involves 20-30 different parties, 40 documents, 200 data elements (30 of which are repeated at least 30 times) and the re-keying of 60-70% of all data at least once. With the lowering of tariffs across the globe, the cost of complying with customs formalities has been reported to exceed in many instances the cost of duties to be paid. In the modern business

environment of just-in-time production and delivery, traders need a fast and predictable release of goods [44].

To deal with the dilemma of facilitating the trade and controlling at the same time, modern customs administrations are using two main tools: Risk Management and Post Clearance Audits. Despite the limited resources and time constraints, customs administrations, with the help of these tools, can conserve and even improve control effectively while reducing controls.

Trade facilitation ideas, such as risk management and the preferential treatment of trusted operators with a good compliance history, can significantly free resources. These can then be redeployed to target the clandestine cross-border activities [45]. In this regard, Risk Management is one of the most important study areas of customs administrations all over the world.

2.7.2. Risk Management at Customs

The World Customs Organization (WCO) defines “Risk Management” in Glossary of International Customs Terms as “Coordinated activities by administrations to direct and control risk” In addition “Risk Analysis” is defined as “The systematic use of available information to determine how often defined risks may occur and the magnitude of their likely consequences.” [2] US DHS (Department of Homeland Security) Lexicon defines “risk” as the potential for an unwanted outcome resulting from an incident, event, or occurrence, as determined by its likelihood and the associated consequences. It also defines “risk analysis” as the systematic examination of the components and characteristics of risk. On the other hand, “risk management” is defined as a process of identifying, analyzing, assessing, and communicating risk and accepting, avoiding, transferring or controlling it to an acceptable level considering associated costs and benefits of any actions taken [13].

Risk management is a logical and systematic method of identifying, analyzing and managing risks. Risk management can be associated with any activity, function or process within the organization and will enable the organization to take advantage of opportunities and minimize potential losses.

Risk management is successfully applied in the private sector, where insurance, banking, trade and industry find that it creates opportunities to improve business results [2].

As risk-based management concept is applicable in almost every business and governmental area, there is a lot of experience that could be shared with customs issue. From customs point of view, risks include the potentials for non-compliance with customs law such as licensing requirements, valuation provisions, rules of origin, duty exemptions regimes, trade restrictions, and security regulations, as well as the potential failure to facilitate international trade.

Risk management is at the heart of border management efficiency and effectiveness and is the key to achieving the ‘balanced approach’ [43]. Sound risk management is fundamental to effective customs operations, and it would be true to say that all administrations apply some form of risk management, either formal or informal[1].

Risk management as systematic identification and implementation of all measures necessary to limit exposure to customs risk, determines which persons, goods, and means of transport should be examined and to what extent. The high-risk persons, goods and means of transport are subject of high-level controls and interventions; despite of low-risk ones that receive high-level trade facilitation. The risk management process helps customs administrations to focus on priorities and decisions on deploying limited resources to deal with the areas of highest risks.

In general, terms, risk in customs can be defined as the potential for non-compliance with national laws and regulations. In this context, there are safety and security risks that threaten public health and security and there are fiscal risks that can result in loss of revenues. Customs administrations are also responsible for the implementation of non-tariff common commercial policy measures, and any other legislation that is related with customs operations.

In Table 2 some of the risk topics and their relevance with the objectives of customs administrations are shown.

	Revenue Collection	Public Health	Environmental Protection	Fight Against Terrorism	Fair Competition
Non-declared goods	✓	✓	✓		✓
Proper Tariff Classification	✓	✓	✓		✓
Proper Valuation	✓				✓
Proper Country of Origin	✓				✓
Trade Policy Measures		✓	✓	✓	
Proper Customs Procedures	✓	✓	✓		
Intellectual Property Rights (IPR)		✓			✓
Trade Agreements Compliance	✓				✓
Money Laundering				✓	
Environmental Crime		✓	✓		
Smuggling					
Drugs and Precursors		✓			
Weapons of Mass Destruction		✓	✓	✓	
Firearms		✓		✓	
CITES			✓		
Nuclear and Radioactive Materials		✓		✓	
High Customs Duty Goods	✓	✓	✓		

Table 2.2: Origin of Risks for Different Customs Objectives. (COMCEC, 2018)

In conjunction with the increase in foreign economic and commercial relations, the rise and diversity in the illegal movement of goods, vehicles and human beings have also been observed. The increase in the legal and illegal transactions has forced the customs authorities and all the administrations facing such transactions to improve their capabilities. Nowadays in Customs Administrations' data not only come from declarations, but also the results of inspections and from other administrations' data. The information obtained from these various sources makes up the customs information system, which, among other things, allows a risk management system to be constructed and declarations to be directed to the different customs clearance channels [46]. Therefore, the fact that these data are growing day by day requires modern and statistical techniques in order to carry out risk analysis systems more effectively. Data mining is a technique about finding insights, which are statistically reliable, unknown previously and actionable from data. Data mining uses sophisticated mathematical algorithms to segment the data and to predict the likelihood of future events based on past events [47].

Actually, data mining could be predictive and descriptive. Descriptive methods such as clustering and association rule mining extract the general characteristics of the dataset. Predictive methods such as regression make predictions using the existing datasets. Data mining is an influential tool and it uses modern techniques, such as advanced statistical models, visualization,

pattern recognition, fuzzy logic, algorithms and machine learning. Of course, in order for all of these techniques to be successful, the business unit needs to include the necessary information in this process.

As it mentioned before, data mining allows the discovery of knowledge potentially useful and unknown. Whether the knowledge discovered is new, useful or interesting, is very subjective and depends upon the application and the user [49]. That is why, the experience and knowledge of the user (business unit) is vital in this process. In the literature, there are various articles written about fraud detection with data mining. In these articles, different methods for the detection of fraud were proposed and examined. Various approaches can be adopted in data mining studies and of course, the customs data set has a great importance in the selection of these approaches. For example, in Shao et al. [47]’s paper due to the complex customs data, multidimensional criterion data model was used. Four phases were defined in this study. Figure 15 shows the schema of the customs data mining process in Shao et al. [47]’s paper

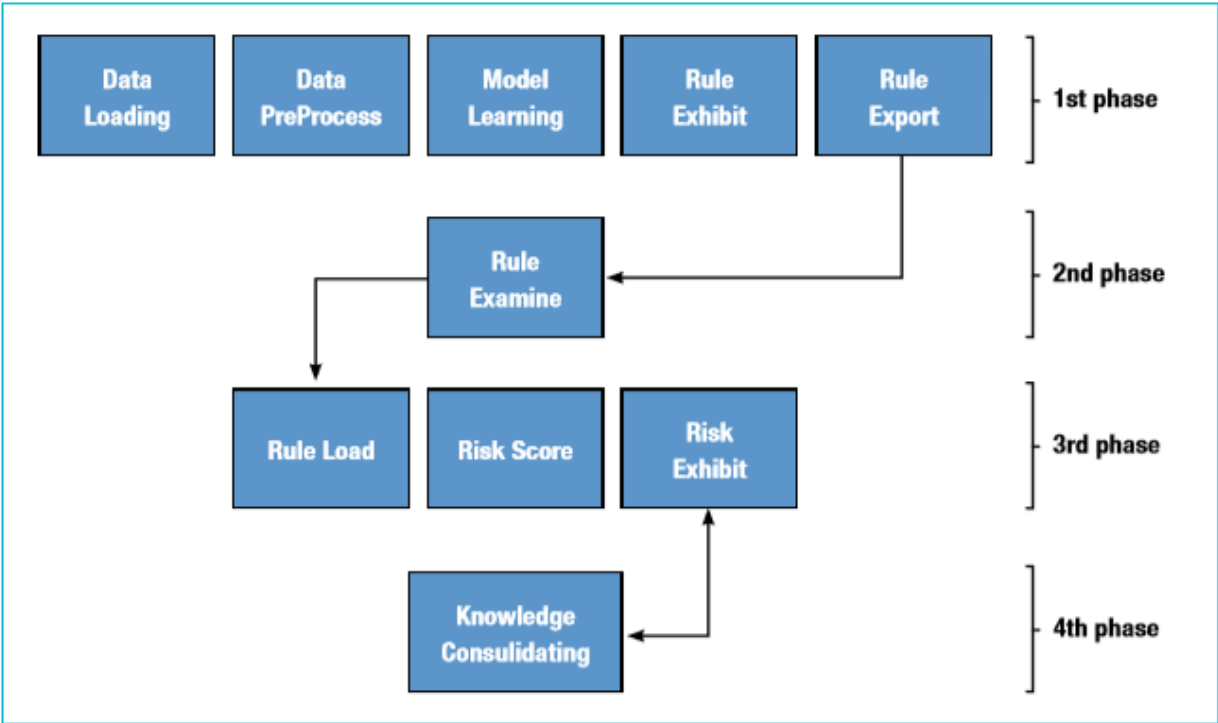


Figure 2.5: Data mining process in customs

For customs, time series analysis, predictive modeling and anomaly detection are some of the most applied data mining techniques in the literature. With time series analysis, which is one of the applicable data mining technique, Customs Administrations can look for patterns. Explaining a significant increase of trade for a certain product from a certain country of origin makes it possible that new risks will be found. Predictive modeling has a great importance on all data mining process. In data mining studies, regressions and decision trees are mostly used in predictive modeling. These approaches are also suitable for making new profiles or for improving existing profiles.

According to WCO, data mining techniques help Customs Administrations to perform better in risk identification, analyzing and preparing for checks. With data mining, Customs Administrations can gain time and can perform fewer checks with the same or even better results.

2.7.3. The Concept of Data Mining and Its Role in Identifying Risks at Customs

In general, risk analysis at customs is built on three pillars. It is valid for all declaration modules including detailed declaration module. All three can be used for both fiscal and safety and security purposes. The first one is named as Fact-Based Risk Analysis. In this process, historical data, which was proven to be non-compliant with the customs legislation or against Anti-Smuggling Law, is used as determining the declarations or vehicles to be channeled to physical control line. The foremost resources can be counted as reports of investigation, data on Anti-Smuggling Database, denunciations, administrative fines and additional tax computations etc. If a potential risk factor has already been violated, the risk factors should be monitored closely. The second one is called Potential Risk Analysis. This method includes assessment of the declaration data and other related resources. The risk indicators on the declarations are determined, and put on the systems as risk profiles. It is potential because it has a probability to break the law or not to be consistent with the regulations, but it is not proven yet. The last pillar is based on Random Selection. Random checks are determined by the system according to a certain volume or count settled on before. The main purpose behind the random selection is to form an understanding that any declaration can be subjected to physical control at any time.

2.8. Risk Preparation/Profiling

A risk profile is the tool a customs office uses to put risk analysis into practice. It is designed to supplement and in many cases replace ad hoc checks on documents and goods by planned working methods. Its actual form may vary from one Member State to another but it must be comprehensive and suited to local conditions. A risk profile may be kept as a dossier or managed by computer, but ease of access by customs officers is paramount. It may be in sections relating to different types of goods. Separate risk profiles may be drawn up for imports and exports. They may also be drawn up for individual products, especially particularly sensitive ones. A risk profile should include a description of the risk area, a risk identification and assessment, risk indicators, checks to be carried out, date of action, results of action taken and evaluation of its effectiveness (stating the indicators used). To remain effective a risk profile must be flexible so that new risks may be identified and gauged, and risks that have been measured and found acceptable may be classed as low. An essential part of a risk profile is continuous review. To remain effective a risk profile must reflect newly identified risks. Risk profile managers must review each profile at regular intervals to ensure that it is always up to date and reflects the latest relevant information (e.g. the latest legislation).

Profiles assist Customs Administrations in making choices since it is generally impossible (and not efficient or necessary) to check all consignments or passengers. Day to day practice has demonstrated that profiles cannot identify all risks. Several reasons exist for this:

- Technical limitations of systems
- Lack of capacity to actually perform the selections

The risk profile contains a description of:

- ✓ The risk area (e.g. drugs/ revenue)
- ✓ Assessment of the risk or possible risk that may be involved;
- ✓ Specific indicators like companies, persons, origin, goods, etc.;
- ✓ The counter-measures to be taken (means of control); and
- ✓ The period that the profile is active for. On the other hand, selection system might be supported by random selection so that unknown risks can be discovered.

From “Risk Areas” to “Risk Profiles”

A risk profile is a document which can be set out in a number of ways but it should be comprehensive and relevant to the traffic throughput in a Customs office.

The risk profile should contain a description of the risk area, an assessment of the risk, the counter-measures to be taken, an action date, the results and an evaluation of the effectiveness of the action taken. A risk profile can be kept in a binder or on a local computer and it should be as accessible as possible to the relevant Customs officers. Risk indicators, on the other hand, are specified selectivity criteria such as: specific commodity code, country of origin, departure country, licensing indicator, value, trader, level of compliance, type of means of transport, purpose of the stay in the Customs territory, financial consequences, or financial situation of the trader/person.

Once established, the profiles along with other information and intelligence will provide a basis for targeting potentially high risk movements of consignments, means of transport, or passengers.

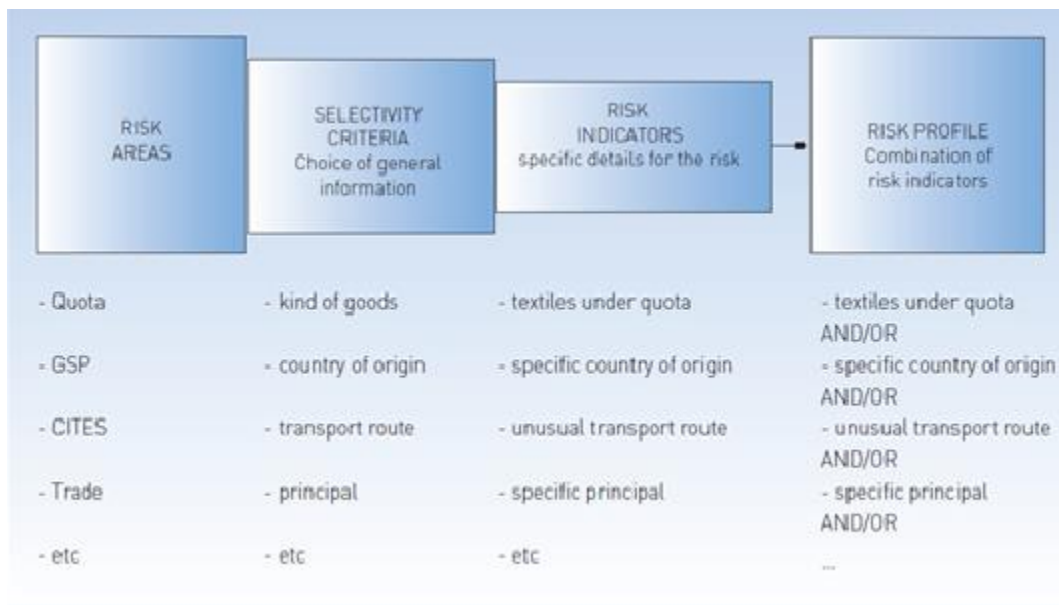


Figure 2.6: Development and Characteristics of a Risk Profile

Steps in developing a profile

A. Collect available data from sources such as:

- Seizure reports
- Intelligence data (Customs investigations reports, available Customs systems, information of other law enforcement agencies, international organizations)
- Cooperation with other law enforcement officers
- Information of trade and industry, shipping companies, stevedores, custom brokers etc.
- Irregularities
- Other signals of Customs officers
- Documents like Bills of Lading, Airway bills, invoices and
- Information available on Internet another open sources.

The collection of recent and pertinent information relating to seizures is crucial. The WCO Customs Enforcement Network (CEN) is a set of tools used to collect, analyses and disseminate information and intelligence mainly on Customs seizures in WCO member countries. The analysis involves examining components of collected information to establish patterns and relationships. Considering the context and the aim of the analysis, identify the main issues by examining the available information/data, such as (if it is drug related): main routes, source countries, suspect ports, risk countries, out of season commodities, mode of transport, concealment methods, etc.

B. Evaluate, structure and chart the data

1. Check and verify the reliability and accuracy of the data,
2. Select a format for the chart which allows you to compare the pertinent data,
3. Itemize data elements on the chart
4. Establish a computer database if feasible

All information collected has no value on its own merit. The usefulness of the information or data is dependent upon its validity and reliability. By evaluating the reliability of the source and accuracy of the information, this information is now ready to be analyzed.

C. Analyze the data

There is no specific structure on how to analyze the data every Customs service has its own method. The Analysis Guidelines as a part of the WCO Risk compendium contains guidelines for this topic.

1. Look for common elements
2. Recognize patterns:
 - a. movement of merchandise (information of counterparts)
 - b. methods of concealment
 - c. conveyances used
 - d. frequently utilized flights
 - e. day/date/time of seizures
 - f. age/sex of violators
 - g. routing of persons/carriers
 - h. origin of contraband

D. Establish, disseminate/activate the profile:

1. Customs profiling system
2. Bulletin board
3. Telephone
4. Mail
5. Briefing

Before a profile is activated in an automated profiling system, consider testing the profile in order to check the impact and outcome

G. Modify profile:

1. Change elements as indicated by feedback
2. Profiles need regular updating.

2.9. Experience from other countries

Risk analysis is indispensable for customs administrations in developing countries. If fewer inspections are to be carried out inspections become more effective. As Laporte, [46] stated, over the past five years, five countries in West Africa: Benin, Burkina Faso, Côte d'Ivoire, Mali, and

Senegal launched projects in this regard with the support of West AFRITAC (Africa Regional Technical Assistance Centre) and the IMF (International Monetary Fund's). For instance, since 2009, the Senegalese customs administration has been working to develop its own system of risk analysis and management. During the course of the first quarter of 2011, all declarations (7,947 in all) were inspected (red, orange, yellow, or green channels) and directed to a channel either based on Downstream SIAR or by using GAINDE's criteria (GAINDE is the French acronym for Automated Management of Customs and Economic Information). Only 56 of the inspected declarations (or 0.7%) were subject to litigation. During the course of the second quarter, 7,633 declarations were inspected (red, orange, yellow, and green channels), or 99.8% of declarations. A total of 60 declarations (or 0.8%) were subject to litigation [2].

2.9.1. Data Mining Project at Turkish Custom Authority (TCA)

TCA has been seeking to strengthen its risk management and analysis system, which was established in 2008 with a view to making its customs control operations at sea ports, airports, land borders and inland more effective. TCA evaluates the procedures based on risk analysis with selective methods ensuring that it simplifies legal trade and prevents illicit trade. While it simplifies the procedures for legal and natural persons, who trade legally, it has employed vigorous effective counter measures to combat organized crime. With the Data Mining Project, it will be provided to analyze the big data of TCA with the most efficient and modern methods. The project covers the years of 2018 and 2019.

There will be fundamental contributions of Data Mining Project to TCA's risk analysis system. The whole Project was planned in line with these contributions. The expected contributions are as below.

- Simultaneous access to data from different sources
- Ensuring the effective use of risk scoring systems
- Increasing selectivity in risk analysis by analyzing high-scale data
- Focusing on more risky areas in real time
- Use of advanced techniques such as modeling and reporting with statistical methods and algorithms in Risk Analysis studies

In accordance with the above expectations, all of the tools in Project were provided under the name of Integrated Data Analytics Solution in August 2018. Thus, the Project has started. In addition to Data Mining tools, the Project has six more components. These are; Data Quality, Rule Engine, Real Time Analytics, Text Mining, Social Network Analysis and Dashboards, Inquiry and Reporting tools. What to do within the scope of the project is as follows.

- Take advantage of the international experience and the best country practices (know-how) in the area of risk analysis and integrated data analytics
- The existing risk profiles in the risk analysis system will be analyzed and transferred to the data mining system in a more efficient format various risk rules and models which are based on the best country practices and international experience will be added to the risk analysis system
- Text mining will be done in free text fields such as the Anti-Smuggling Knowledge Base Records
- Detection of anomaly in customs procedures will be done
- Current risk profiles will be examined by machine learning method and improvement proposals will be made
- Data quality studies will be carried out on various data sets
- Predictive models will be created
- Using social network analysis methods, the connection of risky entities with all other entities can be revealed in different dimensions As a result of the project, significant changes are expected in the current risk analysis system.

In this regard, Table 3 presents the key features of the current and strengthened risk analysis system.

Current Risk Analysis System	Risk Analysis System Powered by Data Mining
- Rule Based & Static Risk Profiles	- Rule Based, Static and Dynamic Risk Profiles, Analytical Models
- Ratio and Time Based Targeting (General)	- Ratio, Time and Number Based Targeting (Both General and Specific), Real Time Scoring
- Classical Methods & Tools Used in Analysis Processes	- Faster and Analytical Analysis Process than Classical Methods & Tools
- Process of Analysis that Requires Intensive Time and Labor (Data Quality, Data Integration, etc.)	- Effective analysis of Large Scale Data
- Systematic Constraints in Data Analysis and Small Scale Data Analysis	- Advanced Statistical Methods, Machine Learning, Fuzzy Matching, Anomaly Detection, Social Network Analysis, etc.

Table 2.3: Risk Analysis System Powered by Data Mining in Turkish.

2.10. Related works

Li, and Wang [68] stated that following the analysis on customs inspection result and the exploration on the regularity of non-consistent between customs declaration and actual commodity by use of data mining based on association rules, a classification model is established to predict the risk of commodity through customs clearance and form the reference for customs inspection and monitoring. A certain customs inspection data in total 26613 samples is adopted in this experiment, in which 2500 samples are identified to be inconsistent between customs declaration and actual commodity. The process of classification model establishment is divided into two steps: First is the training stage, the samples of customs inspection data are identified into training samples and testing samples. A classification model is generated by use of known training samples and classification training system based on association rules mining, which is a subset of association rules. The second is the testing stage, testing samples are classified by use of trained classification model and the classification results are evaluated.

Baştabak [69] investigates the prediction of tariff circumvention using data mining where a total of 13 attributes considered and first KNN classification algorithm then J48 decision tree algorithm were used to classify traders according to their risk level. Yan-Hai and Lin-Yan [70] to solve the conflict between the number of total transactions and the number of inspection officers,

a study was carried out on risk analysis of customs cargo declaration and Q-type cluster method was used to separate the declarations into groups based on their risk level.

An example of local studies to be mentioned relating to customs is Mamo's.D [5] study initiated with the aim of exploring the potential applicability of the data mining technology in developing models that can detect and predict fraud suspicious in tax claims with a particular emphasis to ERCA. He has tried to apply first the clustering algorithm, K-Means clustering algorithm is employed to find the natural grouping of the different tax claims as fraud and non-fraud then cluster is used for developing the classification model using J48 decision tree and Naïve Bayes algorithms. The model developed with the 10-fold cross validation with the default parameter values has shown a better classification accuracy of 99.98% on the training dataset. Similarly with the aim to test the applicability of clustering and classification data mining techniques to support CRM activities for ERCA, Biazen [10] tries to segment customers. The K-means clustering algorithm was used for clustering. The classification modeling was built by using J48 decision tree and Artificial Neural Network (ANN) classification with 10-fold cross-validation and splitting techniques. The model which was built using J48 decision tree algorithm shows better performance with 99.95% overall accuracy rate.

Mezgeb and Berhanu [71] study application of data mining to detect association pattern of customs administration data with market price and currency exchange rate in Ethiopia. The association rule method of data mining is used in this paper to generate the interesting pattern from the data.

The results of the experiments carried out using association rules has discovered that the technique of data mining is applicable to generate knowledge from import and export items in custom administration. Algorithms such as Apriori, Tertius, PredictiveApriori and FliteredApriori were used to generate the associations. The implication of this research finding is to clearly identify the association of import-export items with the market price and the effects of those items on the market price and currency rate in Ethiopia.

On the other hand to investigate the applicability of Data Mining in different sectors - Airlines, Banking, HealthCare, and customs different research has been done locally so far. Most of them used clustering and classification techniques with k-Means and decision tree algorithms.

The above reviews of related literatures reveal that a wide variety of issues in Ethiopian revenue and customs authority are making use of the potentials of data mining. The issues for which data mining contributed can be summarized as ranging from industry and institutional. This current research applies data mining for the improvement of decision making in the area of custom risk channel assignment.

Hence, there is a gap to study the data generated from custom risk management to discover patterns that determine custom risk channels. The output of this research gives new patterns, new information and new insight to custom risk management which helps to improve the quality of service provided by Ethiopian customs Authority. The improvement leads balance between trade facility and control of goods. This research fills the gap by creating a predictive model using data mining classification algorithms to discover patterns determining custom risk channel.

In summary, during the literature there are researches that attempt to apply data mining for custom administration but this research is different from the above researches:

- This study conducted directly on custom risk channel assignment
- The study used three color code (classes)
- The study used variables which are very important for Ethiopian custom risk channel assignment
- Finally the study used more data set from Ethiopian custom risk management system.

CHAPTER THREE

RESEARCH METHODOLOGY

3.1 Introduction

Today, the application of data Mining technology in research areas is rapidly increasing in various sectors. This research applies data mining for Ethiopian customs risk. Ethiopian custom authority is among those sectors which are using data extensively. Because, issues related to custom services increasing amount of data stored in this sector. Thus, the application of data mining technology as a research area for the sector is very important. Data mining technology as a tool has its own methods, procedures and techniques to be followed and used in research. These methods, procedures and techniques may be chosen as per the nature of the data and the objectives of the researcher to be used for a specific study. In this research, the CRISPDM (CRoss-Industry Standard Process for Data Mining) methodology is used. It is the most commonly used methodology for developing data mining research projects. This model describes the activities that must be done to develop a data mining research projects. CRISP-DM has objectives such as ensuring quality of data mining research project results; reducing skills required for data mining; capturing experience for reuse; being general purpose (i.e., widely stable across varying applications and robust (i.e. Insensitive to changes in the environment); tool and technique independent and tool supportable. One important factor of CRISP-DM success is the fact that CRISP-DM is industry-tool and application neutral [51]. Hybrid model is a six-step KDD model developed by Cios et al.(2000). It was developed mainly based on the CRISP-DM model by adopting it to academic research. The main differences and extensions include introducing several new explicit feedback mechanisms and in last steps the knowledge discovered for a particular domain may be applied in other domains. This study will employ Cios et al. (2000) KDD process model for the data mining projects life cycle consisting of : understanding the problem domain, understanding of the data, preparation of the data, data mining, evaluation of the discovered knowledge, and use of the discovered knowledge.

3.2 Cross-Industry Standard Process for Data Mining (CRISP-DM) Process Model

CRISP-DM organizes the data mining process into six phases: business understanding, data understanding, data preparation, modeling, evaluation, and deployment. It consists on a cycle that comprises these six stages. These phases help different sectors understand the data mining process and provide a road map to follow while planning and carrying out a data mining project. This model encourages best practices and offers organizations the structure needed to realize better and faster results from data mining [52]. In the following sections, these phases are discussed; including the tasks involved with each phase in connecting with this study..

3.2.1. Understanding the problem domain

3.2.1.1. International Trade and Customs

International trade is essential and a must for economic development. It is the key to the prosperity of nations. Being aware of this fact, trade facilitation takes an important place in every government's agenda. Decreasing international trade costs owing to the efforts for trade facilitation, alongside with the remarkable developments in transportation and communication possibilities, have led to the exponential growth in international trade volume. However, it is observed that the increase in legal trade operations is accompanied with the increase in illegal transactions.

At this point where the international trade has emerged, naturally, the workloads of customs administrations have increased as well. Nowadays customs administrations are dealing with such high number of exports, imports and transits that cannot be compared to the transactions of a few decades ago. Nevertheless, the resources of the customs administrations are not increasing in accordance with the workloads. With the current resources and the number of employees, it is not possible to control every transaction thoroughly. In the meantime, even if we assume that customs administrations would have enough resources, still it would not be economically logical to control every transaction, since control means time and time is money for traders. Given the large number of transactions, customs administrations face the dilemma of facilitating trade to support traders, while detecting non-compliance in order to protect public revenue and public

safety and security. While trade facilitation comes to the fore, supply chain security also becomes an important issue. Over many years, international trade and transport networks and infrastructures have been identified as potential targets for international terrorism and cross-border crime. While customs have always been in charge of controlling international trade in terms of prohibitions and restrictions, the aspect of securing the international trade supply chain has put growing and additional burdens on customs to manage this balance.

“The United Nations Conference on Trade and Development (UNCTAD) estimates that the average customs transaction involves 20-30 different parties, 40 documents, 200 data elements (30 of which are repeated at least 30 times) and the re-keying of 60-70% of all data at least once. With the lowering of tariffs across the globe, the cost of complying with customs formalities has been reported to exceed in many instances the cost of duties to be paid. In the modern business environment of just-in-time production and delivery, traders need a fast and predictable release of goods [44].

To deal with the dilemma of facilitating the trade and controlling at the same time, modern customs administrations are using two main tools: Risk Management and Post Clearance Audits. Despite the limited resources and time constraints, customs administrations, with the help of these tools, can conserve and even improve control effectively while reducing controls. Trade facilitation ideas, such as risk management and the preferential treatment of trusted operators with a good compliance history, can significantly free resources. These can then be redeployed to target the clandestine cross-border activities [45]. In this regard, Risk Management is one of the most important study areas of customs administrations all over the world.

The World Customs Organization (WCO) defines Customs as “the government service which is responsible for the administration of Customs law and the collection of import and export duties and taxes and which also has responsibility for the application of other laws and regulations relating, inter alia, to the importation, transit and exportation of goods. “In Ethiopia, ERCA’s functions include the enforcement of the Customs Proclamation provisions governing the import and export of cargo, baggage and postal articles; the arrival and departure of vessels, aircrafts,

and other means of transport; goods in transit; and the governance of any goods subject to customs control, including rights and obligations of persons taking part in customs formalities.

3.2.1.1. Principles of Customs Operations in Ethiopia

Customs operations involve the administration of customs law relating to the importation, exportation, movement or storage of goods and the collection of duties and taxes. In this regard, customs operations are a key factor for trade facilitation and economic development of a country. For such a crucial sector to function soundly it should stand on principles that guide its course to worthwhile goals. Accordingly, the Ethiopian customs law contains provisions that clearly prescribe the basic guiding principles that have to be applied on customs operations. These guiding principles, which have important implications for the roles of all stakeholders, including the traders themselves, are the following ones:

1. **Self-assessment:**It is the responsibility of importers and exporters or their agents to assess and submit the value of goods to the customs office, which then determines the appropriate duties and taxes to be paid based on the information provided by traders.
2. **Risk management:**ERCA steers its activities through assessing, directing and controlling risks which emanate from the import and export of goods. The purpose is to strike a balance between trade facilitation and controls. Successful implementation of the risk management principle helps to avoid unnecessary delays and wastage of resources by concentrating customs control on high risk consignments and expediting the release of low risk consignments.
3. **Transparency:**Under this principle, ERCA provides relevant information about trade – including the rates of duties and taxes, fees and charges, customs laws and procedures, appeal procedures, etc. through publications and other means. This guide is one example of ERCA’s commitment to enhancing the transparency of its operations.
4. **Accountability:**ERCA clearly defines the duties and responsibilities of each actor in customs operations.
5. **Service orientation:**As a result of the preceding principles, ERCA is committed to creating a conducive environment to provide equitable, expeditious, predictable and reliable services.

6. Prevention of illegal practices by promoting self-compliance:

Under this principle, which is related to risk management and self-assessment, ERCA will seek to prevent illegal practices such as commercial fraud (under-or over-invoicing, wrong description and classification of goods, etc.), smuggling of prohibited and restricted goods, and others, by taking measures that promote self-compliance. Examples of such measures are the provision of information and advice to traders, advance rulings for customs classification, customs valuation and preferential origin, the implementation of post clearance audits, or the use of simplified procedures for authorized traders.

7. Promotion of priority sectors and economic development:

This principle is aimed at the Authority to play its vital role in expediting the economic development of the country by providing special service to priority sectors, such as manufacturing.

From the above we can observe Risk management is one of the core activity in Ethiopian revenue and customs to improve trade facilitation and national security.

At present, ERCA controls are conducted by means of risk-based methods on almost all declarations. As a result of the risk analysis and assessment, declarations are directed to red, yellow, blue and green lines. In addition to the risk-based controls that are applied selectively, “random” controls are also conducted. Customs controls are carried out in this direction. Although, it is always up to the inspection officers in the field to conduct a more detailed control if they detect inconsistency with the documents even if the declaration has been assigned to just document control line (yellow line) by the risk analysis system.

According to ERCA risk management system there are four channels for declaration:

- Red Line is the line on which physical check as well as documentary is conducted.
- Yellow Line is the line on which only documentary check is conducted.
- Blue Line is the line that approved operators benefit in exportation, on which no documentary or physical control is conducted before the clearance. However, after the exportation, the customs office carry out the control of declarations assigned to the blue line on a simpler basis.

- Green Line is the line that only authorized economic operators (AEOs) benefit on which no documentary or physical check is conducted.

Central risk profiles targeting risky shipments, consignments or declarations are put into the system by the Department of Risk Analysis, which Control by Addis Ababa (Head quarter). Branches risk profiles are created by the divisions of risk analysis affiliated to the branch risk team. These profiles are created for a certain period of time. The risk profiles are revised as needed, and if necessary profiles are extended for an agreed time period. If the profile is decided to be no longer appropriate then they are terminated, but retained in the “history” file of risk profiles.

Risk Management is a logical and systematic method that identifies, analyzes, resolves, and monitors the risks involved in any activity or process. Risk analysis is the use of the available information in order to determine how often the identified risks can occur and the size of possible outcomes [14]. Within customs and border protection, risks are assessed and managed at three levels: strategic, operational and tactical. Strategic risks are the high level whole of agency border risks articulated in the agency’s annual plan and annual risk plan, as well as the government’s broader strategic border management plan. These risks are: terrorism; the unauthorized or irregular movement of people; biosecurity threats; the movement of prohibited and restricted goods; unlawful activity in the maritime zone; and the nonpayment of border related revenue (customs duty, taxes and charges).

ERCA implemented risk analysis that uses selectivity model and the analysis methodology is based on 15 criteria. From which 8 criteria are based on the core particulars of the customs declaration: customs value, customs classification (tariff), country of origin, consignment, the CPC (customs procedure code), special certificate, the company and those of the customs clearing agents. The remaining 7 criteria are Car list, Diplomatic list, Government list, 5% random selection (from Yellow and Green declaration), Yellow (Raw material and chemical) list, Authorized economic operator (AEO) list, and Manufacturing (Especial privileged company) list.

3.2.2. Data understanding

After problem domain identified the data that was explored to conduct the study is from the Risk management database which is managed by ERCA information technology department at head quarter. In information technology department there is one team to analysis and provide data for any individual or company for research and study.

As mentioned in chapter 1, section 1.3, the main objective of this research project was to investigate the potential applicability of data mining techniques for risk channel assignment. Classification and prediction techniques of data mining technology were used in extracting useful and interesting patterns and relationships among features of the dataset. After understanding the problem and the goal of the data mining task is defined, the researcher can easily select and understand the data that would be relevant for the intended purpose.

The data understanding phase starts with an initial data collection and proceeds with activities in order to get familiar with the data, to identify data quality problems, to discover first insights into the data or to detect appropriate subsets of the data being collected. Before starting the actual data mining task, we should be able to clearly define our problem and also have a good understanding of our data to be used for the data mining task.

The initial data collection involves selecting a representative section of the data which is likely suitable and reliable for the objectives stated. For this research, Domain experts are consulted to have insight into the problem domain in order to analysis the data and its structure. Data collection process was carried out from information technology management department in head quarter. The data collected and used in this study included custom risk data which has from 18814 records. Each record of the data contained the following information was initially identified: Declaration No, Declaration year, branch, company name, TIN number, Declarant name, Declarant code, Hs code, Goods descriptions, risk level, reason for change and country of origin. From These information 9 attributes were by consulting domain experts and by ranking attributes using Weka.

The following table shows attributes selected for the experiments.

Attribute Name	Type
Company Name	Nominal
TIN Number	Nominal
Declarant Name	Nominal
Declarant Code	Nominal
Hs code	Numeric
Goods descriptions	Nominal
Country of origin	Nominal
Country of consignment	Nominal
Risk Level	Nominal

Table 3.1: Attributes selected for experiments

There are different algorithms for data classification and prediction in data mining like decision tree induction, Bayesian classification, rule-based induction, the neural network support vectormachines, and *k*-nearest neighbor classifiers. However, for this particular research problem, classification algorithms such as decision tree induction (J48), Naive Bayes classifier and *k*-nearest neighbor (K-NN) classifiers are selected. The proposed method of this study was based on the classification and prediction task of data mining. A decision tree is one of learning algorithms which poses certain advantages that make it suitable for discovering the classification rule for data mining applications. The naive Bayes classifier is also proved to be very effective on many real data applications and The K-NN algorithm is a robust classifier which is often used as a benchmark for more complex classifiers such as Artificial Neural Network (ANN). Moreover, operations of data mining tasks require specific software that contain these techniques. For the purpose of this research, the Weka software was used. Generally, this chapter elaborates the methods, procedures, techniques, software, and the nature and preparation of data which was used in this research

3.2.3 Data Preparation

The data collected for this research from ERCA information technology department custom risk management database. The dataset initially had 14 attributes and 18814 records but after the preprocessing stage, it was reduced to 9 attributes for building the predictive model. The data was extracted to Microsoft Excel for preprocessing purpose. It was then later converted to arff format which is compatible with Weka software. After data is loaded in Weka, Automatic operations by filters like NumericToNominal are applied on three numeric attributes so that all the 3 attributes have nominal value. The data was selected by consulting domain experts which type of data is more important for risk channel assignment.

One of the most important tasks in doing research related to data mining is preparing the data in a way that is acceptable for the specified data mining tools, techniques and tasks. The purpose of this stage is to cleanse the data as much as possible and to put it into a form that is suitable for use in subsequent stages. CRISP-DM defines certain tasks in this phase which are very specific to "structured" data stored in a database like select data, clean data, construct data, and integrate data. Thus, after the data was collected, the researcher prepared the data in such a way that the data was appropriate to the requirements of the selected data mining tasks and the specific data mining tool. At the time of data preparation, the researcher inspected the relevance of individual attribute values and types, quantity and distribution of missing values and noisy data. After that, all constraints related to the collected data were avoided using different mechanisms as per the requirements of the selection techniques. Thus, data preprocessing is the main task that was performed to alleviate the aforementioned data set constraints.

3.2.3.1 Data preprocessing

All raw data sets which are initially prepared for data mining are often large; many are related to humans and have the potential for being messy. Real-world databases are subject to noise, missing, and inconsistency due to their typically huge size, often several gigabytes or more. If there is much missing, irrelevant and redundant information present or noisy and unreliable data, then knowledge discovery during the training phase is more difficult. Moreover, data mining tools may not accept dataset which is not error free. It is well known that data preparation and filtering steps take considerable amount of processing time in data mining tasks. Data

preprocessing is commonly used as a preliminary data mining practice to overcome these constraints. It transforms the data into a format that will be easily and effectively processed by the data mining software which in turn understandable for the users. There are a number of data preprocessing techniques which include: Data cleaning; that can be applied to remove noise and correct inconsistencies, outliers and missing values. Data integration; merges data from multiple sources into a coherent data store, such as a data warehouse or a data cube. Data transformations; such as normalization which can improve the accuracy and efficiency of mining algorithms. Data reduction can reduce the data size by aggregating, eliminating redundant features. The data processing techniques, when applied prior to mining, can significantly improve the overall data mining results. The data collected from information technology department, which was employed for the purpose of this study, suffers from different constraints. These include missing values, encoding inconsistency in various attribute values and unrecognized characters. These constraints resulted in difficulties to perform and fulfill predefined data mining objectives and tasks. Thus, to construct an optimal model, clean and automated data should be prepared. Thus, to achieve the best performance for a selected data set, preprocessing is an important process in data mining tasks [65].

Therefore, the researcher of this study was used necessary data preprocessing techniques as needed depending on the performance of the data for the mining process. For example, various missing value techniques can be used for handling missing data for existing databases and for data left unknown during or not applicable during entry [64]. These include

- Ignoring the instance: The record containing the missing value attribute is
- Ignored/ omitted. This results in loss of a lot of information.
- Manual Replacement: Manually search for all missing values and replace them with appropriate values. Mostly, these are done when the replacing missing values are known.
- Using a global constant: Replacing all missing values with some constant like “unknown” or “?”.
- Using attribute mean/mode: Replacing the missing values with mean or mode of non-missing values of that attribute or of same class.

- Using the most probable value: Replacing by the most probable value, using decision trees or Bayesian methods.
- Expectation maximization (EM) method: It proceeds in two steps. First step compute the expected value of the complete data record likelihood and the second step, substitute the missing values by the expected values, obtained from the first step, and maximize the likelihood function. These steps are continued until convergence is obtained.

The other task that was performed in preprocessing stage was class label balancing using the class balancer method from the preprocess package in Weka. This was done so as to resample and balance the number of class labels for the attributes used as target class in model building process. The target classes were 'RiskLevel'. Risk level class has four labels which are 'Red', 'Yellow', and 'Green'. The labels are imbalanced in size which directed the model to predict for those high in number. The class labels 'YELLOW', 'RED', 'Green' of the 'Risk Level' attribute class had also the size 3493, 5809, 9512 respectively. After the class balancer method was used in conjunction with the filtered classifier, the classes became balanced and only the training data will be reweighted so each class has the same total weight. The test data will be left unchanged and this avoided the majority class bias.

Figures depicted below show the distribution of class labels before and after class balancer was used for target class.

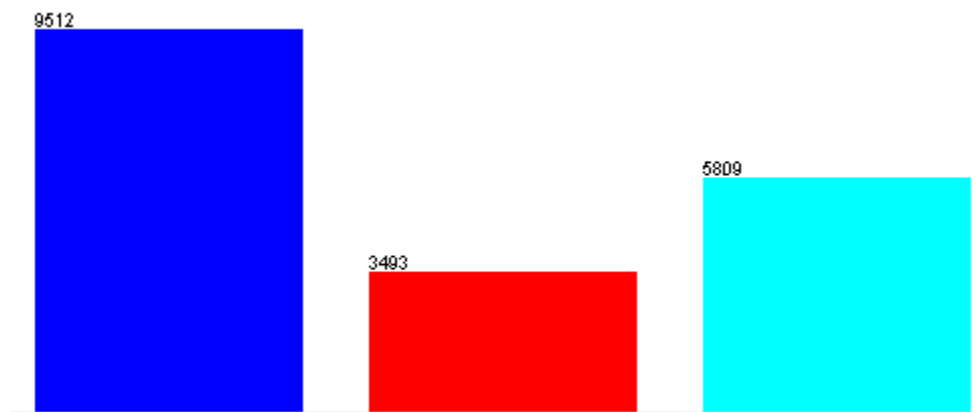


Figure 3. 1: Statistic about class labels distribution in a data set based on 'Risk Level' as a target class before 'Class Balancer' was used.

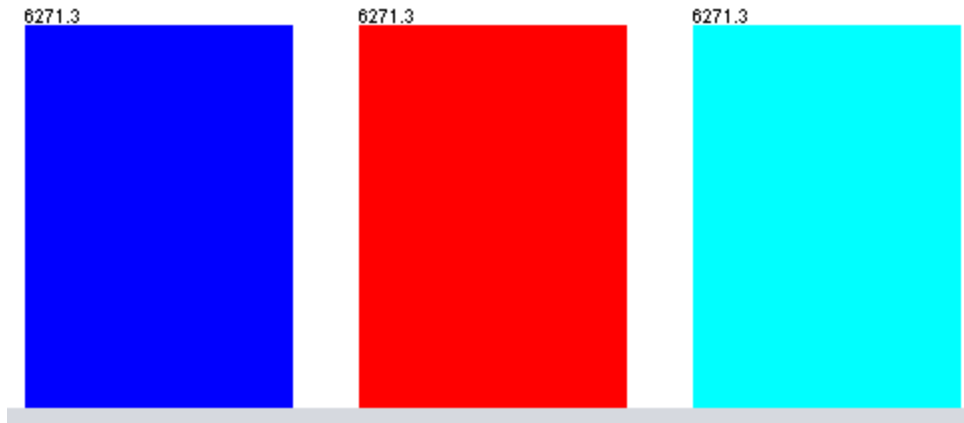


Figure 3.2: Statistic about class labels distribution in a data set based on ‘Risk Level’ as a target class after ‘Class Balancer’ was used.

3.2.3.1.1 Attribute Selection

The classifiers in Weka are designed to be trained to predict a single ‘class’ attribute, which is the target for prediction. Some classifiers can only learn nominal classes; others can only learn numeric classes; still others can learn both. Weka classifiers need predefined classes in order to train and build classification models [67]. Unless they are given the target attribute by the data miner, the last attribute is taken as a target class by default. It is possible to choose any attribute name as a target class no matter its position in the list while running the program. Therefore, the training attribute should be pre-classified so that the data mining algorithms know what the user is looking for. Therefore, “Risk Level” were used as a target class for this study. And three attributes changed from numeric to nominal according to their importance of risk channel assignment these attributes are declaration number, declaration year and harmonized code (Hs code). After getting the rank the last ranked attributes such as declaration year and officer name are removed to implement the experiments for this research.

Removing unwanted attributes, irrelevant for the research goal should be considered in the construction of the final data set. Theoretically, some classifiers such as decision trees determine relevant attributes for classification automatically using the concept of information gain or entropy without manual efforts. However, it is important to exclude those attributes that are not relevant for analysis in order to simplify the tasks performed before model building is started. To

decide on the relevant attributes for this study, a discussion with domain experts, reviewing of various materials and using Weka futures to rank attributes was made. Below the figure shows the rank of attributes.

```
=== Attribute Selection on all input data ===  
  
Search Method:  
    Attribute ranking.  
  
Attribute Evaluator (supervised, Class (nominal): 9 Risk Level):  
    Information Gain Ranking Filter  
  
Ranked attributes:  
1.3094   6 Goods Description  
1.0536   2 TIN  
1.0192   1 Company Name  
0.538    5 HS Code  
0.4508   3 Declarant Name  
0.4493   4 Declarant Code  
0.0996   8 Country of Consignment  
0.0857   7 Country of Origin  
  
Selected attributes: 6,2,1,5,3,4,8,7 : 8
```

Figure 3.3: Rank of attributes

Finally for this research the following attributes selected based on domain expert and rank of attributes: Company name, TIN number, Declarant name, Declarant code, Hs code, Goods descriptions, country of consignment, risk level and country of origin. These information are very important to assign custom risk and most of the rules applying on these attributes. In fact, Ethical considerations also restricted from using sensitive data type.

3.2.3.1.2 Data Protection and Privacy Issues

Prior to the conduct of any data collection and discussion, attempts were made to address the data protection and privacy issues by explaining the main objectives of the research. The researcher was introduced as a MSc. student with an official support letter from the office of the student support department, St. Mary's University. It was tried to make sure that the data being taken will not involve with personal information's and will not be disclosed for third party.

3.2.4 Training and Building Models

One of the important tasks to be performed at this step is selection of relevant software that supports the data mining techniques which are to be employed in the study. To perform any data mining task, selection of appropriate tools and techniques is an important task which may be initiated after the definition of problem to be solved and the related data mining goals. Various types of tools may be used for data mining task by different researchers. Selection of appropriate data mining tools and techniques depends on the main task of the data mining process. Moreover, the accessibility and familiarity of these tools for the researchers can be also another factor for selection. Depending on these factors, Weka and Microsoft- Excel were very effective tools for conducting this research. In conducting this research Weka software version 3.9 was employed for reasons of accessibility and familiarity.

The Weka software is developed by researchers at the University of Waikato in New Zealand. “Weka” stands for the Waikato Environment for Knowledge Analysis. The system is written in Java which is widely available for all major computer platforms [63]. It provides extensive support for the whole process to implement data mining, including preparing the input data, evaluating learning schemes statistically, and visualizing both the input data and the result of learning. Weka includes a variety of tools for preprocessing a dataset, such as attribute selection, attribute filtering and attribute transformation, feeding into a learning scheme, and analyze the resulting classifier and its performance. Weka is organized in packages that correspond to a directory hierarchy. It consists of four graphical user interface modules available to the user .These are: Explorer, Experimenter, Knowledge Flow and Simple Command-line interface. The explorer of Weka interface is the main module for visualizing and preprocessing the input data and applying machine learning algorithms to it. Data visualization in the explorer panel minimizes visual interpretation of complex relationships in multidimensional data. Graphics tools are used to illustrate data relationships. Before the data is loaded and used by the explorer interface, it should be first stored in spread sheet or database and changed in to a name called dataset. Tasks like loading of data, data preprocessing, attribute selection, data visualization and using different learning algorithms, such as classification, clustering and association rule

extraction are accessible the interface of the explorer. It allows us to provide a uniform interface to these learning algorithms, along with methods for pre- and post-processing and for evaluating the result of learning schemes on any given dataset. The Knowledge flow interface is another approach for accessing and using the same functionality with explorer but with a drag-and-drop style of Knowledge flow module. Experimenter interface is used to test and evaluate machine learning algorithms. The last but not the least interface is a simple command line interface which is used as an interface for typing commands [56]. The ways of using Weka by researchers may vary. One way of using Weka is to apply a learning method to a dataset and analyze its output to extract information and knowledge about the data. Another is to apply different learners and compare their performance in order to choose one for prediction [63].

Thus, for the purpose of this study, the pre-processed data set was loaded on Weka machine learning environment and each of the chosen algorithms were run one by one.

3.2.4.1. Data Formatting

The application of the data to Weka required that some preprocessing be undertaken. The dataset produced in Excel for the statistical processes were copied and then converted to .csv file format to allow them to be applied to Weka. The .csv file extension allowed initial analysis to be conducted, with later conversion to be taken in to an arff format (with .arff extension) data file for the experimental outcome to be saved. In arff data format, the internal name of the data set should be stated using the symbol '@'. Attributes should also be defined by preceding the symbol '@' and then with their relevant values and data types. The rest of the dataset consists of the token @data, followed by comma separated values for the attributes. Fig.3.1 depicts the sample of machine understandable format of the dataset in Weka employed for this study.

risk126.arff														
relation: risk125														
No.	1: Dec No	2: Dec Year	3: Branch	4: Company Name	5: TIN	6: Declarant Name	7: Declarant Code	8: HS Code	9: Goods Description	10: Country of Origin	11: Country of Consignment	12: System Risk Level	13: Manual Risk Level	14: Reason
	Numeric	Numeric	Nominal	Nominal	Nominal	Nominal	Nominal	Numeric	Nominal	Nominal	Nominal	Nominal	Nominal	Nominal
1	9.0	2018.0	AAA	MIDROC GOLD ...	000...	Girma G/Egziab...	37328/1198	8.4818E7	VALVE ASSY	DE	NL	Yellow	Green	has a pre
2	10.0	2014.0	AAA	SALINI CONST...	000...	BESHADA BITW...	0024425	8.42129...	FUEL FILTER KIT	IT	IT	Red	Yellow	importer
3	14.0	2018.0	AAA	United Nations ...	000...	MESELE BUSHI...	07329/913	3.00490...	MEDICAL SUPPL...	NL	NL	Yellow	Green	diplomati
4	22.0	2018.0	AAA	MIDROC GOLD ...	000...	GETASETEGN ...	37328/528	4.00922...	HOSE FUELS	SE	NL	Yellow	Green	has previl
5	24.0	2018.0	AAA	MULUGETA AG...	003...	SAMUEL FISSIE...	1250/1139	8.5362E7	SCHNEIDER BRE...	IT	AE	Yellow	Red	descripto
6	39.0	2016.0	AAA	OMAR HUSSIE...	000...	Tirshet Demissi...	1264/349	7.11790...	ARTIFICIAL BRAC...	TH	TH	Yellow	Red	country of
7	43.0	2015.0	AAA	DAVIS & SHIRT...	000...	GASHAW HUSS...	12177/781	8.41370...	GRUNDFOS PUMP	DK	DK	Yellow	Red	high risk
8	46.0	2016.0	AAA	SHINTS ETP GA...	004...	ASHENAFI MUL...	0939/731	9.60719...	NY ZIP	VN	VN	Red	Green	row mater
9	54.0	2018.0	AAA	GOAL PRINTIN...	000...	ABRAHM SIRAH...	0229/294	4.8201E7	CBE LOGO PRINT...	CN	CN	Green	Red	descripto
10	56.0	2016.0	AAA	ADAMA DEVEL...	000...	Tirshet Demissi...	1264/349	8.41490...	IMPELLER ITEM N...	DE	CZ	Yellow	Green	green righ
11	58.0	2016.0	AAA	THE SUGAR C...	001...	Asmamaw Birk...	0838/908	8.4141E7	VACUUM PUMPS	IN	IN	Red	Yellow	choice the
12	68.0	2018.0	AAA	GALLICA FLOW...	000...	AMANUEL GOB...	1035/1008	4.8211E7	PCS OF PLAIN TH...	KE	KE	Red	Yellow	voucher
13	71.0	2014.0	AAA	DASHEN BREW...	000...	TESFAYE TAFE...	1237/637	9.03289...	PROXIMITY DETE...	ID	DE	Red	Yellow	importer
14	73.0	2018.0	AAA	HABESHA BRE...	000...	SIELESH KASSI...	111997/1083	8.42091...	STANDARD AND ...	DE	DE	Red	Green	AEO
15	74.0	2014.0	AAA	HUAWEI TECH...	000...	Teka Gebremes...	0158/742	8.51718...	MICROWAVE TEL...	CN	HK	Yellow	Green	re-export
16	77.0	2018.0	AAA	SUR CONSTRU...	000...	Asmamaw Birk...	0838/908	8.40991...	ROD ASSY CYLIN...	KR	KR	Green	Yellow	has not pr
17	78.0	2016.0	AAA	THE PAN AFRV...	001...	AYALEW YIMER...	118109/718	3.822E7	LABORATORY FC...	US	AE	Yellow	Green	diplomati
18	88.0	2014.0	AAA	HUAWEI TECH...	000...	Teka Gebremes...	0158/742	8.51718...	POWER1000 HYB...	CN	HK	Red	Yellow	re-export
19	91.0	2016.0	AAA	MIDROC GOLD ...	000...	GETASETEGN ...	37328/528	8.4818E7	VALVE MANIFOLD...	US	NL	Red	Green	green righ
20	93.0	2014.0	AAA	AMBO MINERAL...	000...	KEBEDE TESS...	0660/540	3.92049...	PRINTED BOPP ...	MU	MU	Red	Yellow	importer&
21	94.0	2018.0	AAA	SAMSON GEBR...	001...	Ephrem Hailu B...	168543/811	8.5068E7	BATTERY ITEM N...	CN	ZA	Yellow	Red	descripto
22	95.0	2014.0	AAA	HUAWEI TECH...	000...	Teka Gebremes...	0158/742	8.51718...	PATCH CORD, FC...	CN	HK	Red	Yellow	re export
23	97.0	2014.0	AAA	AMBO MINERAL...	000...	KEBEDE TESS...	0660/540	3.92049...	PRINTED BOPP ...	MU	MU	Red	Yellow	importer&
24	97.0	2015.0	AAA	NORDIC CLINI...	003...	SOLOMON TES...	0016/255	8.5115E7	GENERATOR	GB	AE	Yellow	Red	different d
25	100.0	2014.0	AAA	HUAWEI TECH...	000...	Teka Gebremes...	0158/742	8.51718...	SINGLE RAN TEL...	CN	HK	Red	Yellow	re-export
26	107.0	2018.0	AAA	AFRICA UNION ...	001...	Helina Bushe L...	643653/1132	3.00490...	VACCIN	KE	KE	Yellow	Green	diplomati

Figure 3.4: Sample data set to convert ARFF format

3.2.4.2. Algorithms Deployed

This software package encompasses different techniques and algorithms. However, in this study only three techniques were used. One was J48 which is a decision tree classifier and used for decision tree construction. The second was Naive Bayes, where the estimation of the likelihood is performed by means of the simplistic (naive) assumption that an attributes is independent of each other, given the class and the third is The K-NN algorithm is proposed to find out k training samples that are most closest to the target object in the training set. Furthermore, determine the dominant category from the k training samples; then, assign this dominant category to the target object, where k is the number of training samples.

3.2.5 Evaluation

At this stage, models which appear to have high quality are built. Before proceeding to final deployment of the model, it is important to more thoroughly evaluate the model, and review the steps executed to construct the model, to be sure it properly achieves the business objectives. A key objective is to determine if there is some important business issue that has not been sufficiently considered. At the end of this phase, a decision on the use of the data mining results should be reached [51]. Thus, for the purpose of this study, various models were built and

evaluated. Finally, the models which are considered best by the researcher were used for further stages of the study. The evaluation was made based on selected outputs provided by models from each classifier. For the test options, cross-validation with number of 10 folds and 66 percent splitting option.

In 10 folds cross validation approach, the entire data set is divided into 10 mutually exclusive subsets (or folds) or partitions with approximately the same class distribution as the original data set (stratified). Each fold is used once to test the performance of the classifier that is generated from the combined data of the remaining 9 folds, leading to 10 independent performance estimates. In each of the 10 iterations, 1 fold is used as test (holdout) sample while the remaining 9 are used for model building. For methods comparison studies with relatively smaller data sets, the k-fold types of experimentation methods are recommended. In essence, the main advantage of 10-fold (or any number of folds) cross-validation is to reduce the bias associated with the random sampling of the training and holdout data samples by repeating the experiment 10 times, each time using a separate portion of the data as a holdout sample [62].

In addition to 10 folds cross validation 66 percent splitting option was selected for model building. 66 percent splitting option divided the data into 66 percent for training data and the remaining 34 percent for testing data. For this particular study using 10 folds and 66 percent splitting test option by using default and modified parameter.

Note down the results/critical outputs on each algorithm were also the main task that was performed in the experimental process. Then, the results of each of the selected algorithms were summarized.

3.2.6 Deployment

The creation of the model is not the end of the project. Even if the purpose of the model is to increase knowledge of the data, it will be necessary to organize the knowledge extracted, as well as to present it in a useful way to the customer. 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 is the customer, 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 customer to understand up front what actions will need to be carried out in order to actually make use of the created models [51]. For the purpose of this research, results of the study are reported and recommendations based on the findings of the study are given.

CHAPTER FOUR

EXPERIMENTATION

4.1 Experimental setup

As the goal of this study is to detect custom risk using data mining techniques as classification technique was adopted to develop a predictive model. The models were built with three different supervised algorithms i.e. Decision Tree Classification Algorithm, Bayesian Classifier and K nearest neighbor using Weka 3.9 machine learning software. The data analysis and classification was carried out using Weka software environment. Weka provides three options to partition the dataset in to training and test data. These are: preparing distinct files for training dataset and test dataset; cross validation with possibility of setting variety number of folds (the default was 10 fold) and percentage split. For this study 10-fold cross validation and 66 percentage split have been used. 10 folds cross validation options was selected with the intention to be free from bias during dataset partitioning for training and testing and 66 percent splitting option used for comparison for the model.

As it is explained in chapter 3, cross validation mode, the data is divided into some number of partitions of the data, in this case, 10 approximately equal proportions, and each in turn was used for testing while the remainder was used for training. This process repeats 10 times and at the end, every instance has been used exactly once for testing. Finally the average result of the 10 fold cross validation is considered [63]. Therefore, from a total records of 18,814, which were used for model building using 16,933 instances (90%) of these records were used to build(train) models and the remaining 1,881(10%) of the dataset were used to test the performance of the models.

On the other hand, using 66 percent splitting option from a total of 18,814 records which were used for model building 12,417(66%) of these records were used to build(train) models and the remaining 6397(36%) of the dataset were used to test the performance of the model. The models were also evaluated and compared for their classification and prediction performance using modified parameters.

For this study there are 18814 data set and three classes namely Red, Green, and Yellow. The classes are with different size to make balance between the classes we used the class balancer in conjunction with the filtered classifier on Weka preprocessing; only the training data will be reweighted so each class has the same total weight. The test data will be left unchanged and this avoided the majority class bias.

4.2 Model Building and Result Analysis

4.2.1 Decision Tree Model Building

A decision tree is a classifier in which previously unobserved records can be fed into the tree. At each node it will be sent either left or right according to some test. Ultimately, it will reach a leaf node and be given the label associated with that leaf. From the results of the decision tree classifier, it is possible to generate interesting rules. In fact, decision tree methods are often chosen for their ability to generate understandable rules in addition to their classification and prediction capabilities. Using J48 Decision Tree classifier, various experiments were performed for J48 decision tree. The first one was with the default settings of the program and others by modifying 'minNumObj' and 'confidenceFactor' and binary splits parameters.

The modification of these parameters has reduced the size of the tree and number of leaves. The size of the tree and number of leaves of the second trial was also large. In fact, it was possible to reduce the size of the tree and number of leaves below this size. However, further modification of the afore mentioned parameter settings to reduce the size of the tree and number of leaves below the size mentioned above resulted in decreasing overall accuracy of the model.

Different experiments using 10 fold cross validation by changing the parameters for the target class were executed. The first model which was performed with the default setting scored less accuracy as compared with the second that was carried out by modifying the same parameter settings mentioned above.

Similarly different experiments using percentage split option by changing parameters were executed. The first model which was performed with the default setting scored less accuracy as

compared with the second that was carried out by modifying the same parameter settings mentioned above.

Hence, the researcher selected the results of the second trials for both of the test options and experiments as a working model for further stages of the study. The selected rules and results of a confusion matrix from the selected trials are presented below. Results which show overall accuracy, size of trees and number of leaves for some trials carried out using Decision tree classifier before and after modification of the aforementioned parameters are presented in appendix part.

The Following Figure shows the default parameters setting of Weka.

batchSize	<input type="text" value="100"/>
binarySplits	<input type="button" value="False"/>
collapseTree	<input type="button" value="True"/>
confidenceFactor	<input type="text" value="0.25"/>
debug	<input type="button" value="False"/>
doNotCheckCapabilities	<input type="button" value="False"/>
doNotMakeSplitPointActualValue	<input type="button" value="False"/>
minNumObj	<input type="text" value="2"/>
numDecimalPlaces	<input type="text" value="2"/>
numFolds	<input type="text" value="3"/>
reducedErrorPruning	<input type="button" value="False"/>
saveInstanceData	<input type="button" value="False"/>
seed	<input type="text" value="1"/>
subtreeRaising	<input type="button" value="True"/>
unpruned	<input type="button" value="False"/>
useLaplace	<input type="button" value="False"/>
useMDLcorrection	<input type="button" value="True"/>

Figure 4. 1: Default Run option on Weka

Experimentation 1:

The following table shows the experiment using 10 folds cross validation by modifying the default parameters using “Risk Level” as target class. The modified parameters are “binarysplits=True”, “minNumObj=2” and “confidence factor”=0.25.

		Predict			Total	Score (accuracy rate)
		Green	Yellow	Red		
Actual	Green	4516	836	919	6271	72.01 %
	Yellow	387	5806	77	6270	92.6%
	RED	229	681	5361	6271	85.49%
	Total					83.36 %

Table 4.1: **Output from J48 Decision Tree classifier based on ‘Risk Level’ as a target class (using 10 folds cross validation)**

The model which was performed with the modified setting mentioned above provides better result and number of leaves and tree as comparing with the default run option. The data provided to the program 15,684(83.36%) records were classified correctly and the remaining 3,130(16.64%) were classified incorrectly. However, the size of a tree and number of leaves generated from the second experiment were also very large and complex. The size of the tree and number of leaves were 68 and 135 respectively. The results of this table also indicates That 4,516 records were predict correctly as green class whereas 5,806 records were correctly predicted as yellow and 5161 records were predict correctly as Red. While 836 and 919 records were incorrectly classified as Yellow and Red classes respectively but their actual class was classified as Green, 387 and 229 records incorrectly classified as Green but their actual class should be Yellow and Red respectively. Finally 681 and 77 instances were incorrectly classified as Yellow and Red classes but their actual classes were Red and Yellow classes respectively.

Experimentation 2:

The following table shows the experiment using 66 percent split option by modifying the default parameters with “Risk Level” as target class. The modified parameters are “binarysplits=True”, “minNumObj=2” and “confidence factor”=0.25.

		Predict			Total	Score (accuracy rate)
		Green	Yellow	Red		
Actual	Green	1791	176	135	2102	85.2 %
	Yellow	165	1907	50	2122	89.87%
	RED	119	175	1894	2188	86.56%
	Total					87.21 %

Table 4.2: **Output from J48 Decision Tree classifier based on ‘Risk Level’ as a target class (using 66% split option)**

The model which was performed with the modified setting mentioned above provides better result and number of leaves and tree as comparing with the default run option. The data provided to the program 5592(87.2%) records were classified correctly and the remaining 820(12.79 %) were classified incorrectly. However, the size of a tree and number of leaves generated from the second experiment were also very large and complex. The size of the tree and number of leaves were 68 and 135 respectively. The results of this table also indicates That 1791 records were predict correctly as green class whereas 1907 records were correctly predicted as yellow and 1894 records were predict correctly as Red. While 176 and 135 records were incorrectly classified as Yellow and Red classes respectively but their actual class was classified as Green, 165 and 119 records incorrectly classified as Green but their actual class should be Yellow and

Red respectively. Finally 175 and 50 instances were incorrectly classified as Yellow and Red classes but their actual classes were Red and Yellow classes respectively.

One can see from the above two experiments result using J48 Decision tree classifier using 66 percent splitting option has better accuracy which is 87.21 %. This result was selected to compare with Naïve Bayes and K Nearest neighbor classifiers.

4.3.2 Naive Bayes (NB) Model Building

The second data mining technique employed in this study was Naive Bayes classifier. To build this model, the same software package (Weka software) that was used for decision tree model building is employed. The test option used in this experiment was also 10-fold cross validation and 66 percent splitting option. As it is discussed in chapter 3, so as to start building a model to a specific dataset, there is usually a need to prepare the dataset in a form which is suitable for the particular data mining technique and software. An attempt has been made to clean and preprocess the data for Naive Bayes (NB).

In order to carry out an experiment (classification of records of this dataset) using the Naive Bayes classifier of the Weka software, the researcher used the same dataset and target classes which are employed for building decision tree model so far. Experiments with this classifier were also conducted with and without modification of parameter settings target class. The modified parameter was “usesuperviseddescritization” parameter from its default value of “False” to “True” value. This parameter help us to change the numeric value to nominal without using Weka filter. The modified parameter setting has better accuracy than default settings.

The modification of the aforementioned parameter settings improved the accuracy performance of Naïve Bayes classifier for both test option as compared with the accuracy resulted from the default settings of the program. The model provided with an accuracy of 85.18% and 84.94 % using 10 folds cross validation and 66 percent test option respectively. The trials were conducted by changing “usesuperviseddescritization” parameter. Therefore, the results of the model using 10 folds cross validation were selected to compare the model with decision tree and K nearest neighbor. The confusion matrix output of the models for “Risk Level” as target class with 10 folds cross validation and 66 percent splitting option depicted below in table 4.3 and 4.4.

Experimentation 3:

The following table shows the experiment using 10 folds cross validation split option by modifying the default parameters using “Risk Level” as target class. The modified parameter is usesuperviseddescritization=True.

		Predict			Total	Score(accuracy rate)
		Green	Yellow	Red		
Actual	Green	5653	428	191	6272	90.13%
	Yellow	761	5129	381	6271	81.79%
	RED	545	484	5243	6272	83.59%
	Total					85.18 %

Table 4.3: Output of Naive Bayes classifier based on” Risk Level” as a target class (using 10 folds cross validation).

As it is shown in table 4.5, the model which was performed with the modified setting mentioned above provides better result as comparing with the default run option. The data provided to the program 16026 (85.18 %) were correctly classified by NB classifier to build a model with 10 folds cross validation for splitting option and 2788 (14.82 %) records were classified incorrectly .Results in table 4.5 also depict that 428 and 191 of records were misclassified as Yellow and Red classes respectively while they actually be Green class also 761 and 545 records were incorrectly classified as Green but actual classes were Yellow and Red respectively. On the other hand, 484 and 381 instances were misclassified as Yellow and Red classes while they should be Red and Yellow classes respectively.

Experimentation 4:

The following table shows the experiment 66 percent splitting option by modifying the default parameters using “Risk Level” as target class. The modified parameter is `usesuperviseddescritization=True`.

		Predict			Total	Score(accuracy rate)
		Green	Yellow	Red		
Actual	Green	1925	115	63	2103	91.54%
	Yellow	280	1686	156	2122	79.45%
	RED	221	150	1816	2516	72.18%
	Total					84.64 %

Table 4.4: **Output of Naive Bayes classifier based on “Risk Level” as a target class (using 66 percent split option).**

As it is shown in table 4.6, the model which was performed with the modified setting mentioned above provides better result as comparing with the default run option. The data provided to the program 5427 (84.64 %) were correctly classified by NB classifier to build a model with 66 percent splitting option and 985 (15.36 %) records were classified incorrectly. Results in table 4.4 also depict that 115 and 63 of records were misclassified as Yellow and Red classes respectively while they actually be Green class also 280 and 221 records were incorrectly classified as Green but actual classes were Yellow and Red respectively. On the other hand, 150 and 156 instances were misclassified as Yellow and Red classes while they should be Red and Yellow classes respectively.

From the two experiments using naïve Bayes with 10 folds cross validation has a better accuracy than 66 percent splitting option. For further comparison of the model with decision tree and K nearest neighbor we used this result.

4.3.2 K Nearest Neighbor (KNN) Model Building

The third data mining technique employed in this study was Nearest Neighbor classifier. To build this model, the same software package (Weka software) that was used for decision tree and Naive Bayes for model building is employed. The test options used in this experiment were also 10-fold cross validation and 66 percent splitting option. As it is discussed in chapter 3 so as to start building a model to a specific dataset, there is usually a need to prepare the dataset in a form which is suitable for the particular data mining technique and software. An attempt has been made to clean and preprocess the data for Nearest Neighbor.

In order to carry out an experiment (classification of records of this dataset) using the K Nearest Neighbor of the Weka software, the researcher used the same dataset and target classes which are employed for building decision tree and Naive Bayes model so far. Experiments with this classifier were also conducted with and without modification of parameter settings for both target classes. The modified parameter was KNN parameter from its default value of “1” to different number.

The default parameter settings for both 10 folds cross validation and 66 split option had better performance of Nearest Neighbor classifier for target class as compared with the accuracy resulted from the modified settings of the program.

Therefore, the results of the model with default setting for both test options were selected for the purpose of this study. The confusion matrix output of the models using 10 folds cross validation and 66 percent splitting options are depicted below in table 4.5 and 4.6.

Experimentation 5:

The following table shows the experiment 10 folds cross validation splitting on K Nearest Neighbors using default parameters settings and “Risk Level” as target class.

		Predict			Total	Score(accuracy rate)
		Green	Yellow	Red		
Actual	Green	6006	183	82	6271	95.77%
	Yellow	341	5645	285	6271	90.02%
	RED	249	231	5791	6271	92.35%
	Total					92.71 %

Table 4.5: **Output of Nearest Neighbor classifier based on “Risk Level’ as a target class (using 10 folds cross validation).**

As we can see from Table 4.9 the resulting confusion matrix shown above, the nearest neighbor algorithm using 10-fold cross validation scored an average accuracy of 92.71% using default parameter. This result shows that out of the total 18,814 training datasets 17442 (92.71 %) records are correctly classified, while 1372 (7.29 %) of the records are incorrectly classified. Similarly this experiment has shown 183 and 82 records wrongly classified Yellow and Red classes respectively while the actual class was Green. 341 and 249 instances were classified incorrectly as Green while the actual classes were Yellow and Red respectively. Lastly 231 and 285 records were classified as Yellow and Red classes while the actual classes were Red and Yellow classes respectively.

Experimentation 6:

The following table shows the experiment 66 percent splitting option on K Nearest classifier by using default parameters settings and “Risk Level” as target class.

		Predict			Total	Score(accuracy rate)
		Green	Yellow	Red		
Actual	Green	2066	69	36	2171	95.16%
	Yellow	129	1806	136	2071	87.2%
	RED	100	97	1909	2106	90.65%
	Total					91.05 %

Table 4.6: Output of Nearest Neighbor classifier based on “Risk Level’ as a target class (using 66 percent splitting option).

As we can see from Table 4.10 the resulting confusion matrix shown above, the nearest neighbor algorithm using 66 percent split option scored an average accuracy of 91.05% using default parameter. This result shows that 5780 (91.05 %) records are correctly classified, while 568 (8.95 %) of the records are incorrectly classified. Similarly this experiment has shown 69 and 36 records wrongly classified Yellow and Red classes respectively while the actual class was Green. 129 and 100 instances were classified incorrectly as Green while the actual classes were Yellow and Red respectively. Lastly 97 and 136 records were classified as Yellow and Red classes while the actual classes were Red and Yellow classes respectively.

One can see from the above two experiments result using K nearest neighbor classifier based on “Risk Level” as target class using 10 folds test option on default parameter setting has better

accuracy which is 92.71 %. This result selected to compare with decision Tree and Naïve Bayes classifiers results.

4.4 Performance Evaluation of the models

Evaluation is one key point in any data mining process. It serves two purposes: the prediction of how well the final model will work in the future and an integral part of many learning methods, which help to find the model that best represents the training data. One of the objectives of this study was to compare and evaluate the techniques which were used in the study, such as decision tree, Naïve Bayes classifiers and K nearest neighbor to select the one, which performs the best.

In the series of experiments, evaluation of models is done based on performance/accuracy of models and confusion matrix, discussion with the domain expert and based on the soundness of the rules generated. It is easy to learn that all the three classifiers are performing well but there is a difference on their accuracy.

The evaluation was performed on the results of the aforementioned parameters for the Decision tree and Naïve Bayes classifiers but for K Nearest Neighbor result using default parameters. From the three classifiers K Nearest Neighbor has better performance for this particular dataset.

For comparison three best models were selected from Decision tree, Naïve Bayes and K Nearest Neighbor classifiers based on their scored accuracy. The selected models summarized experimental results has showed in the table below.

Table 4.7: comparison of best accuracy between J48, Naïve Bayes and K Nearest Neighbor.

Parameters		Classifiers		
		J48 Decision Tree(66 percent split option)	Naïve Bayes	K Nearest Neighbor (KNN)
Accuracy		87.21%	85.18 %	92.71 %
Precision	Green	0.863	0.813	0.910
	Yellow	0.845	0.849	0.932
	Red	0.911	0.904	0.940
Recall	Green	0.852	0.902	0.958
	Yellow	0.898	0.820	0.900
	Red	0.866	0.836	0.923
F-measure	Green	0.856	0.855	0.934
	Yellow	0.871	0.834	0.916
	Red	0.888	0.869	0.932
True-positive Rate	Green	0.852	0.902	0.958
	Yellow	0.898	0.820	0.900
	Red	0.866	0.836	0.927
False-positive Rate	Green	0.066	0.103	0.047
	Yellow	0.052	0.073	0.033
	Red	0.044	0.045	0.029
ROC Area	Green	0.937	0.963	0.974
	Yellow	0.932	0.949	0.948
	Red	0.0832	0.966	0.964
Correctly classified instances		5592	16043	17442
Incorrectly classified instances		820	2771	1372
Time to build model (in sec.)		0.01 seconds	0.05 seconds	0.03 seconds

Table 4.13 shows the Accuracy, Precision, Recall and F-Measure results achieved by the Naïve Bayes, Decision Tree and K nearest neighbor classifiers. The table also depicts the True-positive , False-positive rates , ROC Area , time taken in building models and number of correctly and incorrectly classified instances by each classifier. These results show that K Nearest Neighbor classifiers outperformed the Naïve Bayes and Decision tree classifiers for most of the metrics used for evaluation. Therefore, K Nearest Neighbor classifier tends to learn more rapidly for this particular dataset. However, the time taken to build a model by the Decision Tree classifier was smaller than Naïve Bayes and K Nearest Neighbor tree classifier. That is, the Naïve Bayes classifier tends to learn more rapidly for the given dataset.

CHAPTER FIVE

CONCLUSION AND RECOMMENDATION

5.1 Conclusion

The purpose of this study was to investigate the potential applicability of data mining techniques in exploring the Custom Risk. In doing so, 18,814 sample data were initially collected from Ethiopian Revenue and custom Authority. The size of class labels for both selected target classes are balanced for model building process. The particular data for the research was taken from the data stored total records of 18814. The total numbers of attributes used in the study were 9. The study was conducted based on the data mining steps or processes discussed in chapter 3.

The methodology employed consisted of steps such as identifying data sources and business understanding, data understanding, data preparation, model building and testing. However, since a data mining task is a cyclic process, these steps were not followed strictly in forward order only. Rather, there was a need to go back and forth among these different steps.

To build models with both classifiers, attribute such as “Risk Level’ was used as target classes. The models were also evaluated and compared for their classification and prediction performance as well as the soundness of the rules extracted from decision trees generated and the best performing models of these classifiers were then chosen. Both the classifiers were built by using the attributes such as “Goods Description”, “TIN”, “Company Name”, “Hs Code”, “Declarant Name”, “Declarant Code”, “Country of consignment “ and “ Country of origin ” and “Risk Level”. 10-fold cross validation and 66 percent splitting option was used as a test option in the model building process.

Various experiments were made iteratively by making adjustments on parameter settings of a program to come up with more understandable and meaningful results and models. When “Risk Level” was used as a target class for the J 48 Decision Tree approach with 66 percent testing option, the model with more accurate results than 10 folds cross validation. There was some better result in accuracy and number of trees and leaves with the model performed with modified parameter settings of the programs. The modified parameters are “binarysplits=True”, “minNumObj=2” and “confidence factor”=0.25. The model which was performed with

these modified parameter provides better result comparing with the default run option. At the beginning of the model using the default parameters of the program, this model had an accuracy rate of 85.91% which is less than from an accuracy of 87.21% that resulted due to parameter settings modification.

However, the structure of the decision tree was very large and difficult to understand. By modifying the aforementioned parameters, it was possible to reduce the size of the tree and number of leaves from 9636 and 9636 to 440 and 879 respectively. Therefore, the second trial was used as the working model for this study.

For the Naïve Bayes approach, the classification accuracy performance also has shown some variations before and after parameter settings modification. Two experiments using 10 folds cross validation and 66 percent splitting test option were conducted on Risk Level as target class. The experiment using 10 folds cross validation split option by modifying the default parameters has a better accuracy as compare with 66 percent splitting option.

Similarly, For K Nearest Neighbor two experiments were conducted using 10 folds cross validation and 66 percent test option using Risk Level. The experiment 10 folds cross validation splitting on K Nearest Neighbors using default parameters settings score a better accuracy of 92.71%. This result is better if we compare with the result from 66 percent splitting option.

The results of this study demonstrate that K Nearest Neighbors have shown the best accuracy performance using 10 folds cross validation test option. Additionally K Nearest Neighbor has better results for both 10 folds cross validation and 66 percent splitting option than Decision tree and naïve Bayes.

In general, the results from this study were helpful and encouraging. The encouraging results obtained from this study indicate that data mining is really a technology that should be considered to support custom risk. Data mining improve custom clearance efficiency and service level affects the overall trade efficiency, investment flow, employment level even regional economic development.

5.2 Recommendation

Although this research work is conducted mainly for academic purpose, the researcher believes that the findings of this study can be used for further exploration and Investigation of custom risk by concerned bodies and organizations. That means, application of data mining technology in custom for risk management is an important research area so as to improve the services being provided as well as to protect illegal goods enter to the country.

In the way of doing this study and on the basis of the findings of the research work, the researcher has come up with a sort of tasks that need more attention for the future work. Thus, the researcher makes the following recommendations based on the results of this study.

- Since this study has used a small percentage of the data which comprises only a five years data of custom risk to build Naïve Bayes, decision tree and k nearest neighbor models, it is better to build more comprehensive models by using more additional data from various sources like goods inspection results, custom intelligence and administration decision. Most of these documents
- The model cannot support first time customer there should be away to include them.
- Although encouraging results were obtained from this study, particularly, using Naïve Bayes, k nearest neighbor and decision tree there might be a probability to obtain more accurate and better performing results using other classification and prediction techniques which were not used by the researcher due to time constraint. Therefore, it is recommended that these classifiers should be applied and proved to this data.
- Generally data mining applications in custom can have tremendous potential and usefulness. The effective use of information and technology is crucial for custom to stay competitive in today's complex environment. The challenges faced when trying to make sense of large, diverse, and often complex data source are considerable. In an effort to turn information into knowledge, Ethiopian revenue and custom should implement data mining technologies to make balance between trade facilitation and control of goods.

Reference

- [1] Revised Kyoto Convention 1999; Keen 2004; eds De Wulf&Sokol 2004 Changala et al, 2015.
- [2] WCO Guideline for post clearance audit. (2012).WCO
- [3] Federal NegaritGazeta of the federal democratic republic of Ethiopia (2008). Proclamation No. 587/2008, page 4123.
- [4] Ethiopian revenue and custom authority. www.erca.gov.et, access date, August 2018.
- [5] Mamo, D. (2013). Application of data mining technology to support fraud protection: the case of Ethiopian revenue and custom authority. Master's thesis, Addis Ababa University.
- [6] Cios K.J. (2007). Data Mining: A knowledge discovery approach. Springer, New York: USA.
- [7] Li, Y., Shu, C., Wang, Y. (2011). Study on management mechanism and risk control of china e-port. Advances in information sciences and service sciences, volume 3, number7.
- [8] Mikuriya, (WCO) a holistic risk-based compliance management approach.
- [9] ERCA Risk management Document and guidelines (2018).
- [10] Biazen, B. (2011). Knowledge discovery for effective customer segmentation: the case of Ethiopian revenue and customs authority. Master's thesis, Addis Ababa University.
- [11] Cios et al (2000)KDD(Knowledge Discovery in Databases).
- [12] Martikainen, J. (2009). Data mining in tax administration: using analytics to enhance tax Compliance. Master's thesis, information systems science, AALTO university school of economics.
- [13] (Data Mining Report DHS, 2006) process model.
- [14] Baştabak, B. (2012). A data mining framework to detect tariff code circumvention in Turkish customs database.Master's thesis, school of informatics of the Middle East technical university.
- [15] Witten, I and Frank, E 2000, Data mining: Practical Machine Learning Tools and Techniques with Java Implementations, 2nd edn, Morgan Kaufmann publishers, San Francisco.
- [16] Berry, M and Linoff, G 2000, 'Mastering Data Mining: The Art of Science of Customer Relationship Management', John Willy and Sons Inc, New York.

- [17] Deshpande, SP and Thakare, VM 2010, 'Data Mining System and Applications: A Review', International Journal of Distributed and Parallel systems (IJDPS), Vol.1 (1), pp. 32-44.
- [18] Kurgan, A and Musilek, P 2006, 'A Survey of Knowledge Discovery and Data Mining Process Models', The Knowledge Engineering Review, Vol. 21(1), Cambridge University Press, pp. 1-24.
- [19] Han, J and Kamber, M 2006, Data Mining: Concepts and Techniques, second edition, Morgan Kaufmann Publishers, San Francisco.
- [20] Hand, D, Mannila, H and Smyth, P 2001, Principles of Data Mining, A Bradford Book, The MIT Press Cambridge, Massachusetts London, England.
- [21] Guo, L 2003, 'Applying Data Mining Techniques in Property & Casualty Insurance,' USA. <http://www.casact.org/pubs/forum/03wforum/03wf001.pdf> Access date: December 4, 2010.
- [22] SAS Institute Inc. 1999, Data Mining in the Insurance Industry: Solving Business Problems Using SAS® Enterprise Miner™ Software, SAS Institute Inc.
- [23] Tan, P, Steinbach, M and Kumar, V 2009, Introduction to Data Mining, 3rd edition, Pearson Education, New Delhi.
- [24] Hajizadeh, E, Ardakani, D and Shahrabi, J 2010, 'Application of Data Mining Techniques in Stock Markets: A Survey', Journal of Economics and International Finance, Vol. 2(7), pp. 109-118.
- [25] Two Crows Corporation 2005, Introduction to Data Mining and Knowledge Discovery, 3rd edition, Two Crows Corporation, Potomac: U.S.A.
- [26] Phyu, N 2009, 'Survey of Classification Techniques in Data Mining', Proceedings of the International Multi Conference of Engineers and Computer Scientists, Hong Kong.
- [27] Qiu, M, Davis, S and Ikem, F 2004, 'Evaluation of Clustering Techniques in Data Mining Tools', Issues in Information Systems, Vol. 5 (1), pp. 254-260.
- [28] Guha, S, Rastogi, R and Shim, K 1998, 'CURE: An Efficient Clustering Algorithm for Large Databases', Proceedings of the ACM SIGMOD Conference, Seattle, USA.
- [29] Meera, G and Srivatsa, SK 2010, 'Adaptive Machine Learning Algorithm (AMLA) Using J48 Classifier for an NIDS Environment', Advances in Computational Sciences and Technology, Research India Publications, Vol. 3(3) pp. 291-304.
- [30] Lavesson, N 2003, 'Evaluation of Classifier Performance and the Impact of Learning Algorithm Parameters', Master Thesis in Software Engineering, Department of Software Engineering and Computer Science, Blekinge Institute of Technology, Sweden

- [31] Wu, X Kumar, V Quinlan, R Ghosh, J Yang, Q Motoda, H McLachlan, J Ng, A Liu, BYu, S Zhou, Z Steinbach, M Hand J and Steinberg, D 2007, 'Top 10 Algorithms in Data Mining', Springer-Verlag, London.
- [32] Wu, X Kumar, V Quinlan, R Ghosh, J Yang, Q Motoda, H McLachlan, J Ng, A Liu, B Yu, S Zhou, Z Steinbach, M Hand J and Steinberg, D 2007, 'Top 10 Algorithms in Data Mining', Springer-Verlag, London.
- [33] Pham, DT Dimov, SS and Nguyen, CD 2005, 'Selection of K in K-means Clustering', Journal of Mechanical Engineering Science, Vol. 219, pp. 103-119.
- [34] Kotsiantis, S and Kanellopoulos, D 2006, 'Association Rules Mining: A Recent Overview', GESTS International Transactions on Computer Science and Engineering, Vol.32 (1), pp. 71-82.
- [35] Azevedo, A and Santos, F 2008, 'KDD, SEMMA AND CRISP-DM: A Parallel Overview', IADIS European Conference Data Mining, Portugal, pp. 182-185.
- [36] Fayyad, U Piatetsky-Shapiro, G and Smyth, P 1996, 'Knowledge Discovery and Data Mining: Towards a Unifying Framework', Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD96), Portland, OR. AAAI Press.
- [37] Apte, C and Weiss, M 1997, Data Mining with Decision Trees and Decision Rules, Future Generation Comp.System, New York www.research.ibm.com/dar/papers/pdf/fgcsapteweiss_weiss_cover.pdf , Access Date: December 20, 2010.
- [38] Hajek, M 2005, 'Neural Networks', <http://www.cs.ukzn.ac.za/notes/NeuralNetworks2005.pdf> Access Date: December 25, 2010.
- [39] Singh, Y and Chauhan, S 2005, 'Neural Networks in Data Mining', Journal of Theoretical and Applied Information Technology, pp. 37-42.
- [40] Jordan, I and Bishop, M 1996, 'Neural Networks', ACM Computing Surveys, CRC Press, Vol. 28(1), pp. 73-75.
- [41] Larose, T 2006, Data Mining Methods and Models, John Wiley & Sons Inc. Publisher, Hoboken: New Jersey.
- [42] Kotsiantis, S and Kanellopoulos, D 2006, 'Association Rules Mining: A Recent Overview', GESTS International Transactions on Computer Science and Engineering, Vol.32 (1), pp. 71-82.
- [43] Widdowson, D. (2012). Risk Based Compliance Management: Making It Work in Border Management Agencies. World Bank Publications. Washington. 7

- [44] COMCEC (Standing Committee for Economic and Commercial Cooperation of the Organization of Islamic Cooperation). (2018). Facilitating Trade: Improving Customs Risk Management Systems in the OIC Member States.
- [45] Grainger, A. (2011). Developing the Case for Trade Facilitation in Practice. *World Customs Journal*, 5 (2), 65-76
- [46] Laporte, B. (2011). Risk Management Systems: Using Data Mining In Developing Countries' Customs Administrations. *World Customs Journal*, 5(1), 17-27.
- [47] Oracle, (2008) What is Data Mining?, Oracle Data Mining Concepts, 11g Release 1 (11.1).
- [48] Shao, H., Zhao, H. & Chang, G. (2002). Applying Data Mining to Detect Fraud Behavior in Customs Declaration. *Proceedings of the First International Conference on Machine Learning and Cybernetics*, vol.3, no., pp. 1241-1244.
- [49] Zaïane, O. R. (1999). Chapter I: Introduction to Data Mining *CMPUT 690 Principles of Knowledge Discovery in Databases*.
- [50] Cios, K and Kurgan, L 2005, 'Trends in Data Mining and Knowledge Discovery', In Pal, N.R., and Jain L.C. (Eds.), *Advanced Techniques in Knowledge Discovery and Data Mining*, Springer Verlag, London, pp. 1–26.
- [51] Mariscal, G., Marban, O. and Fernandez, C. (2010). A survey of data mining and knowledge discovery process models and methodologies. *The Knowledge Engineering Review*, 25(2), 137-149.
- [52] Shearer, C. (2000). The CRISP-DM Model: The New Blueprint for Data Mining. *Journal of data warehousing*, 5 (4), 13-18.
- [53] El-Telbany, M., Warda, M. and El-Borahy, M. (2006). Mining the classification Rules for Egyptian Rice Diseases. *The international Arab journal of information Technology*, 3(4), 303-305
- [54] Wasan, K., Bhatnagar, V. and Kaur H. (2006). The Impact of Data Mining Techniques on Medical Diagnostics. *Data Science Journal*, 5(19), 119-124.
- [55] Bhargavi, P. and Jyothi, S. (2009). Applying Naive Bayes Data Mining Technique for Classification of Agricultural Land Soils. *International Journal of Computer Science and Network Security*, 9(8), 118-119.
- [56] Soman, T. and Bobbie, P. (2005). Classification of Arrhythmia Using Machine Learning Techniques. Southern Polytechnic State University (SPSU), S. Marietta Parkway, Marietta, USA.

- [57] T. M. Cover and P. E. Hart, "Nearest Neighbor Pattern Classification", IEEE Transactions on Information Theory, vol. 13, No. 1, pp. 21-27, 1967.
- [58] J. Han and M. Kamber, "Data Mining Concepts and Techniques", Elsevier, 2011.
- [59]. Ferrari, D. (2005). Mining Housekeeping genes with a Naive Bayes classifier. Master of Science School of Informatics, University of Edinburgh. Five Children in Rural Upper Egypt. Journal of tropical pediatrics, 46, 283-284.
- [60]. Shailja (2009). Classifying Web Services with and without Association Rules. Thesis submitted in partial fulfillment of the requirements for the award of degree of Master of Engineering in Computer Science & Engineering, Thapar University, Patiala.
- [61] Hamilton, H., Gurak, E., Findlater, L. and Olive, W. (2011). Overview of Decision Trees, Rudjer Boskovic Institute.
- [62]. Delen, D. and Sirakaya, E. (2006). Determining the Efficacy of Data-mining Methods in Predicting Gaming Ballot Outcomes. Journal of Hospitality & Tourism Research, 30(3), 317-322.
- [63] Witten, I., Frank, E., Trigg, L., Hall, M., Holmes, G. and Cunningham, S. (2000). Practical Machine Learning Tools and Techniques with Java Implementations. San Francisco: Morgan Kaufmann Publishers.
- [64]. Pedarla, P. (2004). E-Intelligence form design and Data Preprocessing in Health Care. The University of Waterloo in fulfillment of the thesis requirement for the degree of Master of Applied Science in Systems Design Engineering, Waterloo, Ontario, Canada.
- [65]. Nasereddin, H. (2009). Stream Data Mining. International Journal of Web Applications, 1(4), 183-188.
- [67] Bouckaert, R., Frank, E., Hall, M., Kirkby, R., Reutemann, P., Seewald, A. and Scuse, D. (2010). A Weka Manual for Version 3-6-2. University of Waikato, New Zealand, 14-21.
- [68] Li, Y., Shu, C., Wang, Y. (2011). Study on management mechanism and risk control of China e-port. Advances in information sciences and service sciences, volume 3, number 7.
- [69] Baştabak, B. (2011). Application of informatics technologies into customs: origin and tariff code diversion impacts and identification problem. International journal of e-business and e-government studies, vol 3, no 1.
- [70] Yan-Hai and Lin-Yan (2005) Risk analysis of customs cargo declaration and Q-type cluster method was used to separate the declarations into groups based on their risk level.
- [71] Mezgeb and Berhanu (2015) data mining to detect association pattern of customs administration data with market price and currency exchange rate in Ethiopia.

Appendices

Appendix A:

This result shows some of the trials before and after parameter settings modification using J48Decision tree classifier for target class “Risk Level”.

Test option: 10 folds cross validation trials

Trials	minNumObj	ConfidenceFactor	binarySplit	sizeofTree	No of leaves	accuracy
1	Default(2)	Default(2)	False	9636	9626	81.58%
2	25	0.025	True	135	68	83.36%
3	55	0.025	True	77	39	79.78 %
4	35	0.0025	True	103	52	81.85 %
5	40	0.0025	True	95	48	81.01 %
6	55	0.0025	True	73	37	79.73%
7	2	0.25	True	440	879	87.21%

Appendix B:

This result shows some of trials before and after parameter settings modification using J48decision tree classifier for target class “Risk Level”.

Test option: 66 percentages split trials

Trials	minNumObj	ConfidenceFactor	binarySplit	sizeofTree	No of leaves	accuracy
1	Default(2)	Default(2)	False	9078	9064	81.49%
2	2	0.25	True	839	420	87.3%
3	25	0.025	True	135	68	82.28 %
4	35	0.0025	True	103	52	80.76 %
5	40	0.0025	True	95	48	79.69 %
6	55	0.0025	True	73	37	78.66%