ADDIS ABABA UNIVERSITY
SCHOOL OF GRADUATE STUDIES
COLLEGE OF NATURAL SCIENCES
DEPARTMENT OF COMPUTER SCIENCE

Automatic Fraud Detection Model from Customs Data in Ethiopian Revenues and Customs Authority

By: Meriem Muhammed Ali

A THESIS SUBMITTED TO THE SCHOOL OF GRADUATE STUDIES OF ADDIS ABABA UNIVERSITY IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF MASTER OF SCIENCE IN COMPUTER SCIENCE

March, 2013

Addis Ababa

ADDIS ABABA UNIVERSITY

SCHOOL OF GRADUATE STUDIES

COLLEGE OF NATURAL SCIENCES

DEPARTMENT OF COMPUTER SCIENCE

Automatic Fraud Detection Model from Customs Data in Ethiopian Revenues and Customs Authority

By: Meriem Muhammed Ali

Advisor: Sebsibe Hailemariam(PhD)

Approved by Board of Examiners:

Name	Signature
1. Dr. Sebisbe Hailemariam, Advisor	
2. Dr. Fekade Getahun, Examiner	
3. _____	_____

March, 2013

Dedication

To **Tete** and **Uthman** who sacrificed a lot to bring a bright future for me, my sisters and brother.

Acknowledgement

First of all I would like to thank my almighty Allah who gives me the capacity to do this research. Secondly, I would like to express my deepest appreciation and special thanks to my advisor Dr. Sebsibie Hailemariam for his constructive and uninterrupted comments and guidance as well as for his approach, treatment and help at the time of difficulties. Without his guidance and persistent help this study would not have been possible.

I would also like to thank Dr Komminist Woldemariam (FBK, Trento, Italy); he supported me by providing papers which have limited access. Professor Shewandra Hill (Pensulvania University, USA), Professor Ahmed Ali (Addis Ababa University), Tewekel Muhammed and Habtom Gebregzaber aided me by giving guidance for this research.

The staffs of ERCA were fully cooperative in giving me what ever needed information for the research. They sacrificed their valuable working time in answering various questions during the research process. I would like to mention the name of some who play great role in my thesis work especially during problem understanding, data collection, data understanding and data processing: Ato Kebede Lidetu from ERCA head office, Ato Bekri wolchefo and Ato Abdulhakim Abeshi from Addis Ababa Airport branch. Ato Wegayehu Adamu, Ato Teferi Mekonnen, Ato Tefera Measo, Ato Addis Ayele, Ato sisay Asegid, W/t Harra basha, W/t Harira Seid and W/t Akbiret Belayneh from Addis Ababa Legar Branch; and finally, Ato Getachew Shiferaw from information technology development directorate of ERCA.

I owe special thanks and strong appreciation to Ato Haliye Mekonnen, the director of information technology development directorate of ERCA and Ato Nigusie Seid, the software development team leader of ERCA who helped me in providing necessary information and materials which are crucial in this study.

Last but not least, I would like to thank my family and friends for their unlimited support and inspirational encouragement.

Finally, my acknowledgement could not be complete without expressing my grateful and respect for those people who are not mentioned here though they had had contribution for this study.

Table of Contents

List of Figures

List of Tables

List of Annex

Acronyms

AAL:	Addis Ababa Legar
AEO:	Authorized Economic Operators
ASYCUDA :	Automated SYstem for CUstoms Data
ATS:	Automated Targeting System
CART:	Classification And Regression Tree
CHAID:	CHi-squared Automatic Interaction Detector
CIF:	Cost, Insurance and Fright
CPC:	Customs Procedure Code
CRMS:	Customs Risk Management System
DBMS	DataBase Management System
DTI:	Direct Trader Input
EDI:	Electronic Data Interchange
EG:	Extra Goods
ERCA:	Ethiopian Revenues and Customs Authority
FCPM:	Fraud Category Prediction Model
FLPM:	Fraud Level Prediction Model
FPM:	Fraud Prediction Model
FRLPM	Fraud Risk Level Prediction Model
HS-code:	Harmonized System code
IEEE:	Institute of Electrical and Electronics Engineers
ISO:	International Organization for Standardization
IT:	Information Technology
KNN	K-Nearest Neighbor
MC:	Mis-Classification
MD:	Mis-Description
OD:	Origin Difference
RDBMS:	Relational Database Management System
ROC	Receiver Operator Characteristics
SIGTAS:	Standard Integrated Government Tax Administration System

SML	Supervised Machine Learning
SMOTE:	Synthetic Minority Over Sampling TEchnique
SVM	Support Vector Machine
SQL-PL:	Structural Query Language Procedural Language
TIN:	Taxpayer Identification Number
UN:	United Nation
UNCTAD:	United Nations Conference on Trade And Development
USML	UnSupervised Machine Learning
UV:	UnderValuation
VAT:	Value Added Tax
WCO:	World Customs Organization
WEKA:	Waikato Environment for Knowledge Analysis
WTO:	World Trade Organization

Abstract

Customs, which is one of the three wings in Ethiopian Revenues and Customs Authority (ERCA), is established to secure national revenues by controlling imports and exports as well as collecting governmental tax and duties. This research focuses on identification, modeling and analysis of various conflicting issues that Ethiopian customs faces. One of the major problems identified during problem understanding is controlling and management of fraudulent behavior of foreign traders. The declarants' intent to various types of fraudulent activities which result in the need for serious inspection of declarations and at the same time, the huge amount of declarations per day demand significant number of human resource and time.

Recognizing this critical problem of the government, ERCA adopt Automated System for Customs DAta (ASYCUDA). ASYCUDA attempts to minimize the problems through risk level recommendation to declarations using selectivity method that uses five parameters from the declarants' information. The fundamental problem to ASYCUDA risk leveling is, restricting the variables which are used to assign risk level; this may lead to direct the declaration into incorrect channel.

This research proposed a machine learning approach to model fraudulent behavior of importers through identification of appropriate parameters from the observed data to improve the quality of service at Customs, ERCA. In this research, the researcher proposed automated fraud detection models which predict fraud behaviors of importing cargos, in which the problem associated with ASYCUDA risk leveling will be minimized. The models have been built through machine learning techniques by using the past data which was collected from customs data of ERCA. The analysis has been done on inspected cargos records having 74,033 instances and 24 attributes.

Four different prediction models were proposed. The first model is fraud prediction model, which predicts whether incoming cargo is fraudulent or not. The second model is fraud category prediction model, which identifies the specific type of the fraud category among the ten identified categories. The third model is fraud level prediction model, which classifies the fraud level as high or low. The last model is fraud risk level prediction model which is used to classify the risk level of importing cargos into high, medium or low.

Moreover, from the recommendation of IEEE, four best machine learning approaches have been tested for each of the identified prediction models. These are C4.5, CART, KNN and Naive Bayes. Based on the results which are obtained through various experimental analyses, C4.5 is found to be the best algorithm to build all types of the prediction models. The accuracy obtained in the first, second, third and fourth scenarios using C4.5 machine learning algorithms are 93.4%, 84.4%, 89.4%, and 86.8% respectively.

The next best algorithm, Classification and Regression Tree (CART), performed an accuracy of 92.9%, 80.1%, 89.4%, 85.3% for the first, second, third and fourth scenarios respectively.

The researchers observed that both C4.5 and CART perform better for fraud prediction and fraud level classification compared to fraud category and risk level prediction. Moreover, Naive Bayes statistical approach is found to be very poor.

Key words: Fraud prediction, fraud category prediction, fraud level prediction, fraud risk level prediction, classification, machine learning algorithm, ASYCUDA.

CHAPTER ONE

INTRODUCTION

1.1 Ethiopian Revenues and Customs authority

Customs is the main revenue's body of most countries which is responsible to administer the tasks of importing and exporting of goods as well as collecting duties [1, 2]. The duties collected are usually a major source of a country's income. Ethiopian Customs Authority is a governmental agency established to secure national revenues by controlling imports and exports [3]. The Authority protects domestic industry through counteracting contraband activities and held a hand in assisting the economic development of the country. It is responsible for the customs clearance of imported goods as well as tax collection at the customs border. Presently, it is merged with Federal Inland Revenues Authority and Ministry of Revenues; and renamed as Ethiopian Revenues and Customs Authority (ERCA) due to the Business Process Re-engineering [3, 4].

Consequently, the responsibilities of the authority become dramatically expanded. It is now concerned with all aspects of revenues collection including domestic taxes. In addition to this, the authority also responsible in issues like social security, health care, environmental protection, and overall control of foreign transactions covering false indication of origin, illegal foreign exchange transactions and money laundering.

Customs work has two common characteristics: huge amount of cargo transactions and impossibility of examining all goods of the transaction. Hence, facilitating the movement of genuine passengers and cargos is a challenge in customs administration while applying controls to detect customs fraud. On the other hand, international organizations like United Nations Conference on Trade And Development (UNCTAD) push in simplifying clearance and release of genuine passengers and goods while also controlling transactional crime. As stated in negotiation of UNCTAD, it is necessary to balance between these two conflicting issues (i.e. facilitation of service and controlling of fraud) [5].

Furthermore, simplifying customs inspection is recommended and proposed under the revised Kyoto convention of World Customs Organization (WCO) and World Trade Organization

(WTO) trade facilitation negotiation respectively [5, 6]. Using risk management technique is put forward by WTO as a means of facilitating clearance of goods [5].

Given the high volume of transactions in customs, many customs administrations apply risk analysis to determine which goods should be examined and to what extent. In addition to this, risk analysis and risk assessment are processes which are used to identify the most serious risks which need priority action [5].

Due to building risk profile, ERCA also apply risk analysis in order to handle the large amount of transaction through Automated SYstem for CUstoms DAta (ASYCUDA) using selectivity method. One of the main objectives of ASYCUDA is the production of accurate timely statistics by providing an opportunity to captures Import–Export data at source; i.e. following the submission of a declaration to customs as part of the declaration validation and Import-Export processing mechanism [7].

Selectivity method is a technique which is used to classify risks into different level through selectivity module of ASYCUDA. Here, risk refers to the potential for non-compliance and the possibility of event or activities that will have a negative impact up on ERCA's main objective which is revenue collection [7]. In ASYCUDA, there are four classes of risk levels [5, 8, 9]. They are red, yellow, green and blue and described as follows.

- Red: is a color code which is assigned for declarations that contain high risk cargos. Cargos associated with high risk level require physical examination and document checking.
- Yellow: is a color code which is assigned for declarations that associated with medium risk cargo. Documentary checking is required for cargos' declaration targeted to this channel. Moreover, Sample physical examination might be made in addition to document checking.
- Green: is a color code for low risk cargo. A cargo directed to this channel should be permitted for immediate release without any examination (i.e. physical examination and document checking is not required at all for cargos directed to this channel).

- Blue: is a color code which is assigned for cargos' declaration of Authorized Economic Operators (AEO). AEOs are special privileged organizations based on their past history regarding fraud, the exposure to fraud and their great contribution to the country. These organizations are assigned without considering the selectivity method equation which is described in section 2.2.1.2. Document checking at later stage (post audit) is performed for cargos assigned with this color.

This cargo's risk level classification is believed to minimize resources that will be wasted in physical examination of all cargos; by classifying incoming cargo into different risk levels and performing selective examination based on the classification result. The risk level of cargo indicates the possibility of fraudulent activities on certain cargo as well as the impact of fraud on revenue collection.

Selectivity, as a risk management method, is enforced in many countries as a consequence of using integrated customs clearance management system [6]. Presently, more than 90 countries use ASYCUDA for import and export declaration processing [10]. This system is not efficient in supporting risk managements of customs authorities since ASYCUDA's risk management module is forced to be depend on the option of being able to apply and combine simple selection criteria (lists of importers, origins, Harmonized System code (HS-code) or commodity code, transistor, etc.)[6].

Due to the large amount of cargo transactions in customs, large number of declarations are registered and processed per day. Accordingly, Physical examination of all declarations' cargos cannot be practical due to high amount of cargo transaction and limited amount of resources in customs [11]. Recognizing this critical problem of the government, ERCA adopt ASYCUDA which implement selectivity method, however it has problems regarding risk level assignment. The data size in customs is increasing significantly. However, this large amount of data contains hidden knowledge that is important to identify the fraud behaviors of importing cargo which can be discovered through machine learning techniques.

1.2 Fraud and Fraud Detection

In [12], Fraud can be defined as the misuse of organization's profit without necessarily leading to direct legal consequences. According to Concise Oxford English Dictionary, fraud is defined as

"*wrongful or criminal deception intended to result in financial or personal gain*"[1]. In the case of customs, it may refer as incorrect description of goods, undervaluation of goods, misclassification of goods, under payment of taxes due, invoice falsification and etc. [13]. Usually, these can be occurred when a customer tries to minimize the amount of tax due which depends on various variables such as: incorrect HS-code, country of origin, weight, number of items, Custom Procedure Code (CPC) and others. So that, providing appropriate controlling technique is very important to handle fraud.

Fraud detection is one part of the fraud controlling techniques which can be achieved through automation and helps to reduce the manual parts of screening/checking process [12]. Nowadays, it becomes a central application area for knowledge discovery in databases and machine learning applications, even though it has challenging technical and methodological problems to be applied [1].

There are various fraud detection techniques to recognize patterns of fraudulent transactions. For instance, in [11], a clustering method was proposed to build a fraud detection model in which the model can select interesting subsets which return higher recovery from customs declaration. In [2], association rule based classification technique was proposed to build predictive model which identify higher probability of fraud in customs. It offers an opportunity to detect customs fraud with limited examination of imported goods by available scarce resources through directing the appropriate customs channel. H. Shao et al. [1] proposed classification method for building fraud detection model for identification of fraudulent behavior of customs' declaration. In [14], F. Bonchi et al. proposed a classification-based methodology for planning audit strategies in fraud detection of tax. In [15], A. Kumar and V. Nagadevara proposed application of classification based machine learning techniques for detecting customs fraud.

In this research classification based fraud detection model have been proposed to predict the fraud behavior of incoming cargo which improves the quality of service in customs of ERCA.

The remaining parts of this chapter have been organized as follows. In Section 1.3, the motivation behind this research has been presented. The detailed discussion about the problem

[1] Concise Oxford English Dictionary-10th edition

which necessitated this research is addressed in Section 1.4. Section 1.5 contains the significance of this research. The objective and scope of the research has been presented in Section 1.6 and Section 1.7 respectively. Finally, the methodology including tools used in this research has been described in Section 1.8. 1.9 outlines the organization of this thesis.

1.3 Motivation of the research

The common characteristic of customs work is high amount of cargo transaction and the difficulty of checking all cargos of the transaction [5, 6]. As a result, customs administrations face challenges in facilitating the clearance of cargo in accordance with controlling fraud [5]. This requires effective way of risk management that classifies risks into different levels, and can solve two conflicting issues; controlling fraud and facilitating service. In order to solve this problem, the selectivity method that is used under ASYCUDA has been used though it has various problems. Among this, usually risk levels of cargos' declarations are assigned incorrectly. So that, tax will be uncollected as cargos which have high risk are incorrectly classified to low risk level and leaves custom's zone without any physical examination. On the other hand, when the low risk cargo is incorrectly classified as high risk cargos and physical examination is conducted, resources will be wasted due to incorrect risk level assignment, at the same time the genuine customer/importer might be mistreated or forced to wait for extra longer time unnecessarily. This is due to declarations are directed improperly to wrong customs clearance channels. In addition to this, the selectivity method directs the larger percent of the declarations into red channel and maximizes the amount of cargos which require physical examination. Though the actual channel for larger number of declarations is green, the usual amounts of the declarations which are targeted by the system to red channel are above 70%. As a result, the authority wants to minimize it into 51% (as interview with one of the customs officer). For instance, the declarations' color assignment distribution of ASYCUDA in year 2011 looks as Table 1.1 shows.

Table 1.1: Declaration risk level assignment in year 2011using ASYCUDA's selectivity method

Declaration in:	Red	Yellow	Green	Blue
Number	20358	7510	773	189
Percentage	70%	26.05%	2.68%	0.66%

Furthermore, the researchers also concerned with unsatisfied need of the authority which makes the organization in need of changing the system that currently in use; because of various reasons which are described below [7]:

- Increasing volume of international trade, a technique that process (analyze) huge amount of data is required.
- What are the discriminate variables from customs data of ERCA to predict the fraud behaviors of importing cargos?
- Heavy demand on customs to produce, maximizing identification of fraudulent customs declaration and minimizing customs examination effort.
- Customs wants to minimize human intervention to the system.
- Etc.

Taking these into consideration, ASYCUDA's risk assignment process which has been discussed in Chapter 2 and 3, in which several peoples are involved and also the existence of human intervention into the system, there might be occurred serious problems like hiding and deleting risky importers profile, negligence activities like feeding erroneous information when an officer is being tiresome, during risk metrics are fed to the system manually. In addition to this, due to ASYCUDA's 15% random selection and risk level assignment error, the power of modifying the risk level of customs' declaration (i.e. on the hard copy one) goes to risk management officers. Accordingly, when officers get an opportunity to deal with importers/agents, it might be great source of corruption.

1.4 Statement of the problem

The underlying problem that necessitated this research is the sensitivity and vulnerability of the responsibility which is assigned to ERCA for the existence of fraud in general. In particular, the high level of uncollectable tax is a great problem in this organization according to the interview with the customs officers.

Even though, ASYCUDA has contribution in mitigating problems faced regarding examination of all cargo and attempts to determine the correct risk level of cargos, tax might be uncollected

due to incorrect risk level assignment of cargo since it doesn't use a scientific method to uncover hidden patterns and learn novel knowledge which can help to classify cargos' risk efficiently. Moreover, it doesn't mitigate human intervention to the system. The risk management system in ASYCUDA is subjective because it works based on simple criteria of selectivity [13]. It simply uses a hand coded rule.

The selectivity method which is used currently in ERCA is not an appropriate method due to the organization's particular needs which have been stated in section 1.4. Furthermore, since ASYCUDA's selectivity module is forced on customs and depended on the option of being able to use and combine simple selection criteria (lists of importers, origins, etc.), more informative attributes are not used and more interesting patterns are still hidden due to the analysis tool they are using (i.e. manual analysis and analysis using excel). According to the volume of data they have, the amount of transaction they performed and the performance they required, special analysis technique that exhaustively uses the past data and the inferred knowledge is required.

This research have been answered the following research questions

- Could we find appropriate data from ERCA customs data warehouse for fraud detection purpose?
- What are the classification-based machine learning techniques which can be applied in fraud detection of customs cargo in ERCA to improve the quality of service and minimize problems associated with ASYCUDA risk leveling?
- Could we find appropriate model for fraud detection of ERCA customs operation?

1.5 Significance of the Study

ERCA has a data warehouse that keeps track of the daily transaction collected from different custom branches across the nation. Therefore, building appropriate fraud detection model from the data will be an opportunity to explore hidden knowledge and inform customs' officer appropriate information associated with a specific cargo. At the same time it minimizes the amount of uncollectable tax and duty by minimizing the amount of cargos with fraud from leaving custom's zone without examination. Moreover, it saves the unnecessary wastage of resources by optimizing (maximizing) the number of declarations which are directed to low and medium risk level channel and minimizing high risk level channel. In real situation the amount

of declarations which have fraud are smaller in percent. In addition to this, legitimate customers will be protected from waiting for unnecessary physical examination. Accordingly, the custom authority also earns the goodwill by not causing harassment to the genuine importers through identification of declaration which has not fraud. Last but not least, the study will minimize human intervention to the system by providing fraud detection models.

In general, the benefits of this research are, in terms of both increasing customs duty recovery and saving in wastage of resources for physical examination with associated facilitation of customs clearance service.

1.6 Objective

1.6.1 General objective

The general objective of this research is to build fraud detection models from customs data in ERCA using machine learning approach, that can predicts fraud behavior of importing cargo and classify fraud risks into different levels, in which the quality of customs service in ERCA is improved.

1.6.2 Specific objectives

In order to achieve the specified general objective, this research has undertaken the following specific objectives:

- Identify and organize appropriate dataset for training and testing purpose
- Measure the goodness of the collected data
- Identify the appropriate classification task that would address the business problem
- Identify the appropriate modeling techniques to customs data for the intended classification purpose
- Building the appropriate model using the collected training dataset
- Test the goodness of the model built using the test data set
- Analyze the performance of the model
- Forward possible conclusions and recommendation by addressing future work

1.7 Scope and limitation of the research

This research is conducted based on the data obtained from ERCA, on the customs transaction records covering the period between January 1, 2011 and March 31, 2012 and the import transaction performed in Addis Ababa Lagar (AAL) branch, in which ASYCUDA is utilized more powerfully. AAL can be taken as a representative of the other branches of the customs; 80% percent of the total customs transaction of ERCA is facilitated in this branch. The study has been limited in supporting risk management of ERCA customs (foreign trade for import) by building different cargo's fraud predictor or classifier model using machine learning approach that can help to uncover hidden knowledge based on improved knowledge discovery mechanism and enhances the capability to identify the certainty of fraud without opening the content of cargo.

1.8 Methodology

The methods employ to achieve the stated objectives in Section 1.6 of this research are presented below.

1.8.1 Literature Review

Extensive literature review was conducted to get deeper understanding on fraud detection systems and in particular on machine learning approach to classify/predict customs' declaration fraud as well as to understand the problem domain. Printed materials like books, journal articles, the ERCA training and working manuals as well as electronic materials on the Web were referred for this purpose.

1.8.2 Problem understanding

Understanding of the requirements from a business perspective is the core component in addressing the research objective. In this regard, the researcher should work closely with domain experts through on site observation and interview, and analyze document with the intent of problem understanding. This allows the researcher to define problem and determine the goals of the study, and then converting the knowledge into a machine learning problem definition, and a preliminary plan for controlling of fraudulent operation [16, 17]. In order to have an insight in the overall business objectives, details of the customs authority business process was reviewed

and evaluated from the view of their core business objectives and goal that they would achieve. The existing system was analyzed and gaps were identified.

1.8.3 Data understanding

Usually data understanding begins with an initial data collection and continues with tasks that makes the researcher familiar with the data [16]. Data understanding includes deciding, collecting and analyzing the data needed to model and address the problem. The data quality problems regarding completeness, redundancy, missing values, appropriateness of attribute values, and others were checked.

After the problem and the objective were clearly defined, understanding what kind of data are available and which data is the most appropriate to the problem under consideration for custom fraud detection perspective were conducted. In order to further understand the nature of the data and the attributes at the customs database of ERCA, discussion was made with different officers and documents were analyzed.

1.8.4 Data preparation

All tasks which are undertaken to get the target dataset from the initial raw data are considered in the data preparation phase [16]; it deals about placing the data in a format suitable for building the required models. This is the phase where issues; like data selection, data cleaning and data construction were resolved [18]. After data cleaning which includes checking the completeness of data records, removing or correcting for noise and missing values, the cleaned data were further processed by attribute selection. Though the ideal practice for attribute selection is providing all the attributes in the database for the attribute subset selection tool and letting it to find out those attributes which are the best predictors, blindly including unnecessary attributes can create confusion for the models [17]. Therefore, after the researcher knows the meaning of each attributes, irrelevant attributes were removed and some attributes were transformed into suitable format for model building task. In this study, missing and inconsistency values were handled by cross checking in different sources of the authority's data.

Moreover, as primary objective is detection of fraud from customs declaration data, class label attributes were identified in this phase of the research process.

1.8.5 Modeling

Modeling phase focused in designing and building fraud detection models using the data of foreign trade database. The models should extract and learn new knowledge from existing massive amount of data by employing machine learning technique.

In order to obtain a mechanism which learn automatically and make prediction based on past observation, machine learning methods, techniques and tools play important role [19, 20]. Machine learning is a technique which helps to use computer systems in building predictive models and to improve the efficiency of the system based on the retrieved data from the real time environmental setup [20].

There are a variety of machine learning techniques; each is suitable for discovering a specific type of knowledge. In this study, using machine learning technique, models that can predict/classify cargos' fraud behavior have been built, to systematically select the most appropriate strategy, to cope with fraud risk classification and management problem.

The models which were built using knowledge driven analysis, should predict fraud behaviors of new incoming cargos and insure capability of pattern recognition from extremely large database.

In order to build models which have higher capability of prediction/classification, we did more in algorithm selection that implement machine learning models. Accordingly, four machine learning algorithms which are recommended by Institute of Electrical and Electronics Engineers (IEEE) [21] were selected for initial experiments since it is practically impossible to do experiment in all prediction algorithms that exist nowadays.

These algorithms are C4.5, CART, KNN and Naive Bayes. These algorithms are most influential machine learning algorithms in research community [21].

Based on the comparison of experimental analysis result of these four machine learning algorithms, algorithm which performs best prediction accuracy has been chosen for model building task and further experimental analysis. The performances of the best algorithms have been maximized through further parameter tuning technique.

For validation purpose, two typical testing modes have been used. To select the appropriate testing mode for an algorithm on the given dataset (either k-fold cross validation or percentage split), the researcher performed learning curve analysis which informs the sufficiency of the dataset that in turn recommend the testing mode.

1.8.6 Model Evaluation

Performance of each of the model built should be evaluated. Appropriate model evaluation metrics were designed so as to reach proper analysis of the research finding. Accuracy of the model was taken as the most appropriate metric.

For evaluating the performance of the model, predictive accuracy of the model in accordance with the common machine learning evaluation metrics such as precision and recall were used [20, 22]. Moreover, significant time difference in model construction, Confusion matrix and Receiver Operator Characteristics (ROC) were considered in model selection. Confusion matrix is used to visualize the detailed distribution of correctly and incorrectly classified instances.

The researcher also included performance indicator metrics such as the number of leaves, the size of the tree, the average true positive and false positive rate, and F-Measure values to provide further insight about the goodness of the model built.

1.8.7 Tools

Tools used for this study:

- For preparing the dataset: Microsoft Excel Starter 2010, Microsoft office Access 2007, Oracle DataBase Management System (DBMS) or Structural Query Language Procedural Language (SQL-PL) and Waikato Environment for Knowledge Analysis (WEKA).
- For Building and evaluating the fraud detection models, we have used WEKA-3.67. WEKA-3.67 is the latest stable version of WEKA that implements different machine learning algorithms. According to the comparative study which was done to identify the best freely available tools for classification, the WEKA toolkit has achieved the highest classification performance for large dataset [23].

1.8.8 Environment

- A machine with the highest possible specification is needed to handle the large dataset. The researcher first attempt on the laboratory machine failed due to lack of availability of the required computational resource. To avoid this problem the researcher acquire a CORE i5 machine with 8GB RAM having hard disk capacity of 450GB.

1.9 Organization of the paper

The rest of the thesis document is organized as follows: Chapter two presents literature review and related works. Chapter three presents the problem and data understanding part of the study. Chapter four describes how the final data set has been prepared. Chapter five discusses the design of the methods employed in the constructed models including how testing modes were selected. Chapter six presents methods employed in building the fraud detection models including experimental evaluation of the models as well as the results achieved. Finally, Chapter seven presents the conclusions and recommendation of the researcher.

CHAPTER TWO

LITERATURE REVIEW AND RELATED WORK

2.1 Introduction

In this chapter, the researcher primarily discussed three important and related concepts. The existing system of customs in ERCA has been reviewed in the first section. The second section focused on review of the widely in use fraud detection approaches and/or algorithms. Related works are discussed in the third section.

2.2 Overview of ASYCUDA

To compile this section of the chapter, the researcher used different published and unpublished resources, physical observation and interviews as appropriate. Unpublished documents like training and working manuals of the authority in addition to physical observation and interview were primarily chosen to understand the existing system of ERCA.

Customs of different countries have started using Information Technology (IT) applications like Electronic Data Interchange (EDI), Automated SYstem for CUstoms DAta (ASYCUDA) and Automated Targeting System (ATS) to tackle the growing volume of paper based work [15]. On this line, the Ethiopian customs has already adopted ASYCUDA since 1998 however the current version which is ASYCUDA++ has been deployed on 2004.

ASYCUDA is a computerized and customizable customs management system which includes import and export procedures, and other recognized customs regimes like transit and warehousing. It is developed in Geneva by UNCTAD, and it works in client server environment under Linux operating systems and Relational DataBase Management System (RDBMS) [24].

There are about 90 countries which use ASYCUDA to facilitate customs transaction [10]. Since ASYCUDA is customizable, each country can uses it in their own context to meet their local need.

ASYCUDA takes into account international codes which are common and internationally understandable like commodity code, country code, currency code, etc. and standards established by International Organization for Standardization (ISO), WCO and the United Nation (UN) to

offer its core features. The core features include national configuration, tariff and control tables' maintenance, declaration processing and accounting. Moreover, it also offers a number of important new features such as selectivity module and Direct Trader Input (DTI). The selectivity module is used to predict the risk level of customs declaration whereas DTI allows declarant to have a capability to directly lodge their declarations on the customs server from remote [24].

Since this research is conducted in order to build automated fraud detection model, the researcher focused on the selectivity module of ASYCUDA, in which selectivity method is implemented. However, the selectivity criteria rules which are used in selectivity method may not be identical in different countries in which ASYCUDA is used as ASYCUDA's core features are customizable. The next subsections describe ASYCUDA's selectivity process in detail.

2.2.1 Processes of ASYCUDA selectivity risk level prediction

Selectivity is a technique which is used by customs risk management system through selectivity module in order to identify cargo's risk level. Here, risk refers the potential for non-compliance and the possibility of event or activities that will have a negative impact upon ERCA's main objective which is revenue collection [7]; usually, it is associated with importers and declarant fraudulent activities against revenue collection.

According to the interview, there are about 600 importing cargos' declarations which are declared per day in ERCA, in which number of items might be incorporated. Controlling fraud through examining all the cargos is impossible due to shortage of resource and high amount of transaction in customs. Therefore, customs obliged to use certain risk management technique which helps to indicate the possibility of fraud; in which high probability of fraud will be identified. Risk management replaces random examination of goods and its document with planned and targeted working method and maximizes uses of customs resources [25]. Limiting customs' inspection is also a necessary precondition for effective improvement of international trade transactions; since international trade requires efficient and simple trade formalities, procedures and operations [24].

Hence, ERCA has employed a risk management system through selectivity method which is used under the selectivity module of ASYCUDA. In terms of the cargo itself, the level of intervention

by customs is determined from risk scores assigned by selectivity method of Customs Risk Management System (CRMS) through selectivity module that focuses on identifying consignment in which undeclared, misclassified, undervalued goods and incorrect origin and description might be found.

A selectivity module offers customs with a powerful tool to facilitate the cargos clearance process while improving its control capacity. Based on the selection criteria (including a random rate) which can be maintained at national, regional and local levels, the consignments can be selected for inspection [24].

Figure 2.1 shows the input-process-output description of the selectivity risk leveling process.

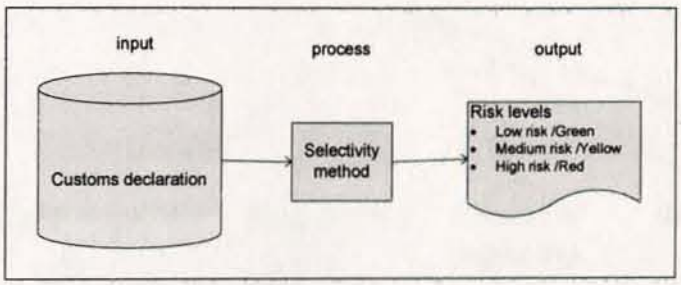


Figure 2.1: Components of selectivity risk leveling (input-process-output)

As can be seen from Figure 2.1 ASYCUDA selectivity risk leveling is composed of three different components. These are the input, the selectivity method process and the output.

The inputs of the selectivity method are taken from the database based on the criteria's value declared for the importing cargo and the weights given for those criteria. First, the inputs (i.e. risk indicators or selectivity criteria and their weights) are produced through various phases such as selection and ranking. Then, the selectivity method processes the inputs in order to generate the final risk level of particular declaration which is declared for particular cargo. The three components of selectivity have been discussed in the following subsections.

2.2.1.1 Selectivity criteria selection and ranking (input of selectivity)

The first component of selectivity process contains the major elements of selectivity which are generated through the following activities:

- Identification of criteria which are available in the declaration document (five criteria are identified in ERCA. These are Tariff / value, Importer, country of origin, declarant and CPC).
- Weighting of the criterion based on their discriminative power for selectivity (high level or attribute weighting).
- Weighting various possible values of the criterion (low level or value weighting).

Figure 2.2 shows the two level weighting (attribute level and value level)

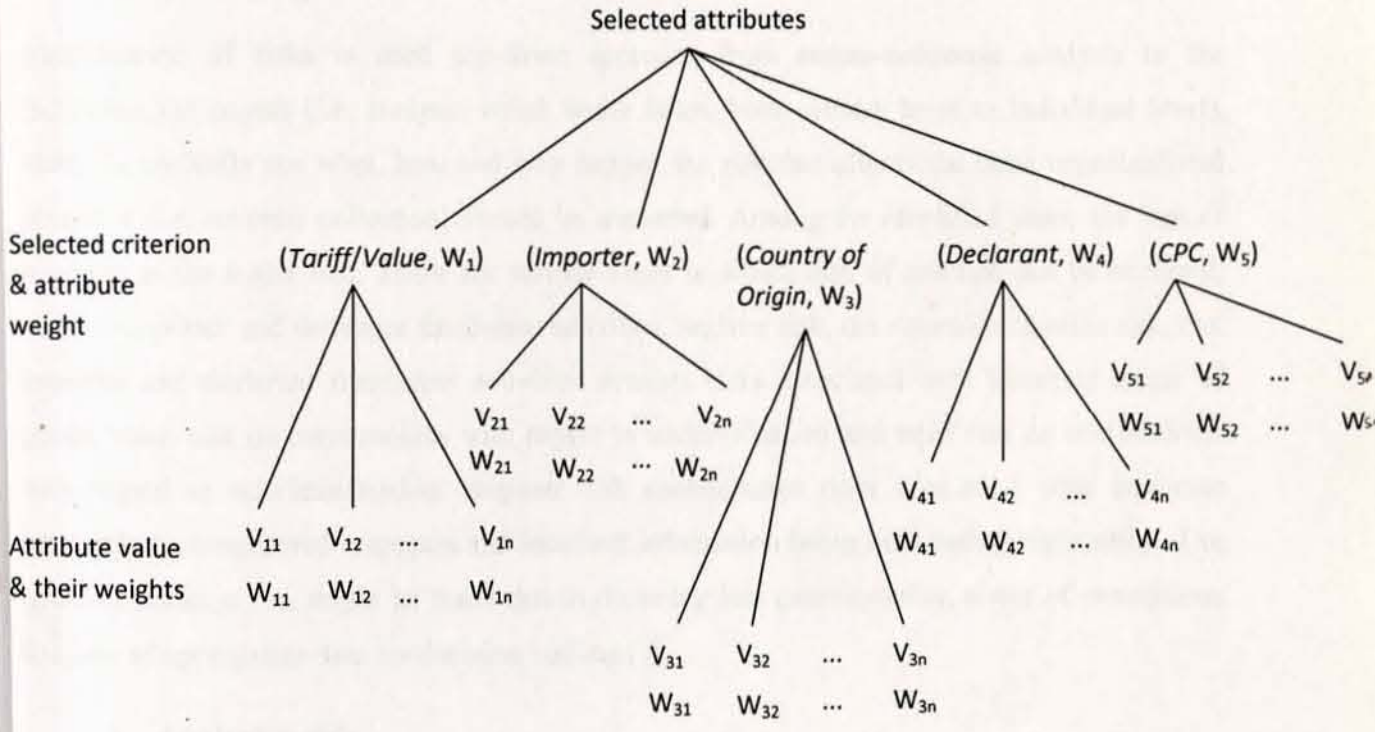


Figure 2. 2: Inputs of selectivity (attribute and value weighting)

Where, n is the number of possible values for the given attribute.

According to Figure 2.1, W_i refers to the weights of the attribute (criterion i) where as W_{ij} refers the weight of the i^{th} attribute and j^{th} possible value.

In order to obtain the above elements (inputs) of the selectivity method four phases are performed. These are identifying risk, analyzing risk, assessing and prioritizing risk and

capturing the weight (risk level) of each criterion and the corresponding values into ASYCUDA database.

Risks which affect the basic organizational objective of ERCA (which is revenue collection) are identified during the 1st phase. The identified risks are analyzed due to the concept of likelihood and consequence of risks during the analysis phase. Then, risks assessed and prioritized in order to assign the weights of each criterion. Finally, the risk levels of each criterion (i.e. weight) and the associated values will be assigned which tends to be fed to ASYCUDA database as inputs to selectivity process.

a. Identifying risk

Identification of risks is used top-down approach from macro-economic analysis to the individual tax payers (i.e. analysis which under taken from country level to individual level). Here, theoretically the what, how and why happen the risk that affects the basic organizational objective (i.e. revenue collection) should be answered. Among the identified risks, the loss of revenues is the major one. There are various cases in which loss of revenue can be occurred, such as importer and declarant fraudulent activities, register risk, tax return/declaration risk, etc. Importer and declarant fraudulent activities consists risks associated with incorrect origin of goods, value risk on commodities with regard to undervaluation and tariff risk on commodities with regard to misclassification. Register risk encompasses risks concerned with duplicate registration, unregistered taxpayers and incorrect information being hold during registration. Tax return/declaration risk might be made due to declaring less quantity/price, abuse of exemptions and lack of appropriate data for decision making [7].

b. Analyzing risk

The identified risks are analyzed in order to establish the significance of each risk, and in order to be informed what strategies and resources are needed to manage them. This can be achieved by analyzing the relationship between likelihood of the risk occurring and the resultant consequence, of the risk. The result of this relationship provides the risk level of identified risk, allowing the comparison and prioritizing of all the risk. Likelihood is measured in terms of the probability of risk occurring at all, whereas consequence is measured in terms of the impact that a risk would have on the achievement of the organizational objective. Here, mapping the

identified risks into criteria (attribute in the database) which are directly correlated from is important. Hence, the risks are listed and analyzed with their factors to be occurred particularly. Then the potential risk factors will be considered as risk criteria; such as tariff/value, Custom Procedure Code (CPC), company (importer) that declared the import, country of origin in which the item is made, and declarant that acts as an agent for importers in conducting customs business/works on their behalf [7].

Though this type of risk analysis was used in the early stage of compliance risk management implementation, nowadays ERCA is also using some statistical analysis such as average and standard deviation for analyzing risk.

c. Assessing and Prioritizing risk

Based on the analysis result in the previous phase, the five attributes are selected as criteria since they are assumed to be the major factors for a particular fraudulent activity directly or indirectly.

In order to rank risks of the selected attributes into high, medium or low, the risk will be assessed and prioritized based on the matrix which is done based on the relationship between likelihood and consequence of each criterion. In line with this, the weight of each criterion is determined based on the impact on the primary goal of the revenue authority which is revenue collection and the probability of risk occurrence on specific criterion in relation with the others criteria.

Due to probability and impact of risk associated with particular risk criteria, the weight for each criterion will be varied. Based on the analysis and assessment result, each risk criterion with its corresponding weight or coefficient is presented in Table 2.1.

Table 2. 1: weights of risk criteria

Risk criteria	Weight coefficient
Tariff & value	3
CPC	1
Country of Origin	2
Company	1
Declarant	2

The rate (weight) of the criterion is determined by the relative contribution to the revenue collection which is the main goal of the authority and the relative risk exposure among risk criteria. Tariff-value, which rated 3, is the main factor for revenue collection and it is the most sensitive one, there may be a probability of misclassification of goods into HS-code with minimum value and rate. At the same time, it will have greater impact on revenue collection. The country of origin and declarant are rated 2, the contribution and the exposure is lower than that of the tariff-value one. The company and CPC are rated 1 because the probability of incorrectness in CPC and company is very low even though the impact if it is occurred is very high. After the weight of each criterion/risk element is assessed, the risk rate codes will intent to be captured to the database as presented in Table 2.1.

In addition to this, the values of each criteria are assessed through critical thinking on how those criteria could be a hole for a particular fraud and how could be detected, as well as experiences from past. Based on the assessment, each value of the criteria will be assigned to a particular risk level (weight).

The existing risk assessment methodology of ASYCUDA in ERCA is based on the assessment of selectivity criteria as described in Table 2.2.

Table 2. 2: Risk level assessment methodology for values of each selectivity criteria

No.	Risk criteria	Risk assessment technique on values of criteria
1	Tariff/Value of a commodity	Using statistical method
2	Country of origin	Direct risk assignment
3	Customs procedure code	Direct risk assignment
4	Importer (Company)	By building risk profile
5	Transitor (declarant)	By building risk profile

Due to the limitation of ASYCUDA to learn by its own from past observation, the risk level (weight) of each selectivity criteria and its corresponding values are analyzed, assigned and fed to the database manually. The analysis is based on risk assignment that is determined by domain expert or through statistical calculation from past data or through building risk profile. Weight

and risk level which represent the magnitude of the five attributes and corresponding values can be used interchangeably throughout the literature review part of this research. The five criteria are assessed as presented in Table 2.2 and described in detail as follows.

i. Value-tariff of a commodity

The risk level (weight) of value-tariff is assigned to each commodity which is identified by Harmonized System (HS)-code. The codes are assigned by WCO for each commodity that is exchanged internationally. It has also reserved digits for further classification of commodities in local context. 'Value' and tariff are analyzed independently, and then they will be merged and used as a single risk element. Tariff is duty or tax to be paid for particular imported goods. 'Value' is the summation of price, fright and insurance of goods purchased. It is the source for the tax and duty which will be owed.

Tariff: the weight of tariff is assigned based on statistical analysis of past data. The analysis will be taking place in HS-code of commodities i.e. the number of transaction, contribution to the revenue collection, the possibility of misclassification (confusion for description).

'Value': the weight of 'value' is also assigned for every commodity which is identified with specified HS-code using statistical calculation. Number of transaction and revenue contribution are also considered in assigning the risk level or weight of 'value' for each commodity. In addition to this, the commodity risk level associated with 'value' is determined regarding the most probable 'value' which is used (i.e. declared versus reference 'value'). Reference 'value' means the 'value' of each commodity which is dispatched by the authority within certain interval (period). The reference 'value' will be accepted 'value of a commodity when a commodity has been declared with 'value' less than that of reference 'value'. Whereas in case of a commodity which is declared with 'values' greater than that of reference 'value', the declared 'value' will be an accepted 'value' of a commodity. Commodities which do not have reference 'value', take the declared 'value' as accepted 'value'. Therefore, commodities which are likely accepted the reference 'value' will have high risk.

Finally, the risk level of specific commodity is determined based on the relationship between tariff and 'value' under consideration of the main objective of the organization which is maximizing revenue collection.

ii. Country of origin

Country of origin is the country in which the item is fabricated or made. The risk level of each country is assigned using a direct assignment, based on the quality of specific product that will be manufactured in particular country. Usually high quality goods are used better technology and better raw material. Due to this, the 'value' of goods may raise up. At the same time, the amount of tax and duty due will be high. On the other hand, low quality goods might be on the reverse. Changing country of origin may minimize the 'value' of the commodity and at the same time the amount of tax and duty due. Importers might use this hole to minimize the amount of tax and duty due. So that, countries those produce high quality goods will have low risk level whereas those do not will be assigned to medium and high depending on the vulnerability.

In addition to this, if the country has strong trade policy which protects against the forged invoice, the risk level of that particular country will be low and it will be high or medium if it has not, depending on the magnitude of fraud exposure of the country.

The assessment does not consider the possibility of separation between country of consignment and origin. If the country of consignment and origin is differ, the risk officer will take some adjustment like changing risk level by considering the country of consignment.

iii. Customs Procedure Code (CPC)

CPC is a code that determines the importing procedure of the cargo. It is used to manage various customs formalities or procedures in which goods for home use (commercial goods), investment, donation, diplomatic, etc. are identified. It also indicates the characteristics of the goods and the purpose which the goods are imported. CPC is not most likely exposed to risks since it is cross checkable through declaration's supportive documents. The risk level of CPC is assigned based on direct assignment. It is assigned to medium or low risk level only. For instance, consider three CPC-codes 4000 000, 4100 000 and 4000 415. The first one (4000 000) is the CPC-code for commercial importing declaration. The commercial goods are highly exposed for fraudulent activities. So that the risk level for this CPC-code will be medium. 4100 000 is the CPC-code for manufacturing goods (raw materials) importing declaration. Importers who have manufacturing license are assumed to be genuine importers. So that this CPC-code (4100 000) is assigned to be low risk level. The latter code is the CPC-code for diplomat. Diplomats have special privilege

regarding examination (i.e. it is internationally forbidden to examine diplomat's goods). Accordingly, CPC-code 4000 415 is assigned to be in low risk level.

iv. Importer

The importers profile is one of the five parameters to determine the final control channel line. Every new importer is assigned initially to high risk level. Accordingly, the importers risk profile is maintained constantly in excel format in different branches of customs. Based on the information collected from different customs branches of inland revenues (Standard Integrated government Tax Administration system (SIGTAS)), intelligence and Audit result, the profiles of importers will be maintained in local database which is manipulated in the directorate level. Though the local database which is used for building importers risk profile has about 27 attribute, the profile of importers is built using four attributes; such as *total tax*, *Infraction*, *total shipment* and *number of offence* (frequency of fraud). Then based on these variables the consequence and likelihood will be calculated. Consequence determines the economic impact of the non-compliance, and it is derived as: the ratio between ***total amount of infraction*** to ***total tax collected from an importer*** multiplied by ***hundred*** whereas likelihood indicates the probability of offence by particular importer, and it is calculated as: the ratio between ***number of offence*** to ***total shipment*** multiplied by ***hundred***. Then the result of consequences and likelihood will be changed into high, medium and low risks based on the threshold which is set by the authority. Table 2.3 and Table 2.4 show the thresholds of likelihood and consequence respectively.

Table 2. 3: The three thresholds for likelihood

Threshold	Risk level
<=2	Low
2<x<=5	Medium
>5	High

Table 2. 4: The three thresholds for consequence

Threshold	Risk level
>=5	Low
5<x<=7.5	Medium
>7.5	High

Then, the final risk level for specific importer is assigned based on the matrix that is developed based on the relationship between likelihood and consequence; as presented in Table 2.5.

Table 2. 5: Matrix of the consequence and likelihood to assign risk level of importers

Consequence	Likelihood			
		Low	Medium	High
	Low	L	M	M
	Medium	M	M	H
	High	H	H	H

v. Declarant

Declarant, that acts as an agent for importers in conducting customs works on their behalf regarding importing process. Everything what is done in the risk analysis and assessment of importer should be considered on assigning risk level of declarant also.

According to the discussion with the risk management team, since the result of the analysis using local database required further manual investigation regarding the remaining attributes, the local database software which helps to build importers and declarants profile could not be integrated with ASYCUDA system. So, the result of the assessment is fed to ASYCUDA manually.

d. Capturing the weight of criteria(attribute) and value

During this phase, the weights of each criterion's and values regarding each criterion which have been analyzed, assessed and assigned in the previous sub-phases (Section b and c) are fed to ASSCUDA database. So, each criterion will have the assigned weight: high (3), medium (2), and low (1). Furthermore, each criterion's value will have any of the three values which indicate the risk level of specific object in the specified criteria; such as 2(high), 1(medium) or 0(low) risks.

These values will be fed to ASSYCUDA database. These are the final inputs to apply selectivity method.

2.2.1.2 Selectivity method

In this component of selectivity process, the final selectivity color is triggered through the selectivity method under selectivity module of ASYCUDA. The selectivity criteria are assigned based on the information declared for the attributes which are selected as risk criteria; during cargo declaration. Based on this, the selectivity method grades the declared import into red (high risk level), yellow (medium risk level) or green (low risk level).

The selectivity method calculates the mean using the weights of the criteria and the weights of their values, through propagating the given weight of each criterion as coefficient for criterion's value weight (using equation (2.1)). Then based on the given threshold (which is fixed by the authorized body) the final risk level values will be assigned, in which the final color code which targets declaration's control channel will be determined.

$$\bar{x} = \frac{\sum_{i=1}^{M} w_i x_i}{\sum_{i=1}^{M} w_i} \qquad (2.1)$$

Where x_i is the risk level for a criterion's value(value level weight), w_i is the weight for a specific criterion i (attribute level weight) , M is the number of the criteria considered (i.e. five in this context) and $\bar{x}$ represent the weighted mean value that is used to determine risk level code and color code for final clearance channel. Then, the result of the equation will be compared with the thresholds which have been set by the authority and the risk level value (0, 1 or 2) will be set based on the interval that $\bar{x}$ falls. The threshold for each final risk levels is a secret for the authority.

2.2.1.3 Risk level prediction

In the output component, the risk level of certain cargo's declaration will be predicted. As a result the ASSYCUDA selectivity color code (control channel line) will be assigned as stated in Table 2.6.

Table 2. 6: Selectivity code with corresponding color code and risk level

Value (risk code)	Risk level	color code for risk level
2	High	Red
1	Medium	Yellow
0	Low	Green

The value 2 is interpreted as high risk level and the red control channel line (color code) will be assigned. If the value is 1, it is interpreted as medium risk level and the yellow color code will be assigned. If it is 0, it indicates low risk level and the green color code will be assigned.

Consignments assessed as high risk are subjected to full physical inspection. On those consignments determined to be medium risk, either only document checking is performed or physical checking of some samples might be made in addition to document checking. Consignments assessed as low risk do not require physical inspection and document checking at all (immediate release is permitted).

2.3 General approaches of fraud detection

Today fraud detection is an interesting research area in different sectors since fraudulent activities are enlarged in type and size. There are various approaches of fraud detection applications. In this research two fraud detection approaches were reviewed. These approaches are knowledgebase (rule-based) and machine learning (Supervised-example based) approaches.

2.3.1 Rule-based(knowledgebase) approach

Knowledgebase approach deals about constructing a fraud detection application based on the knowledge of domain expert in which the domain expert design the rules based on his/ her prior knowledge of fraud. Since rule-based approach is a white box approach, it can be possible to reason out why particular instance is identified as fraudulent however, expert who has prior knowledge of fraud is required to design the rules [26].

J. Hollmen argued that rule-based fraud detection approach allows a domain expert to use their skill in formulating the knowledge and generating rule-based model which is fully understandable, and any possible risk provided by such a system may be understood by a set of rules that identifies the risk [27]. However, M. Sternberg and R. Reynolds stated its management problem when rules are complex [28]. Rule-based systems do not consider situational change

unlike machine learning systems which adapt the new environment if once devised for the task [27].

L. A. Digiampietri et al. argued that rule based approach has advantage of using the knowledge of the experts to construct a model that evaluate any operation however rules needs to be updated frequently to deal with new fraudulent behaviors [29].

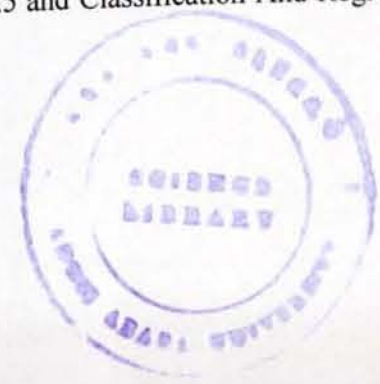
This was our motive to propose the machine learning system to build fraud detection model since ASYCUDA is implemented based on rule-based approach which uses if...then rule to predict the risk level of a declaration associated with a cargo even the weights of the criterion and criterion values are considered.

2.3.2 **Machine learning (supervised) approach**

The convergence of using computer and communication creates a society who uses information as basic needs. There are a lot of data locked in different database. Yet they couldn't be processed to extract previously unknown and potentially useful information. Therefore, the amount of data being collected in databases become far exceeds our ability to analyze the data without the use of automated analysis techniques. Building computer program which sift through the massive amount of data and extracts most possible useful patterns that will generalize to make accurate prediction on the future data is required. Machine learning is the appropriate technique for this purpose [30].

Machine learning deals about finding and describing structural patterns which is used to predict new examples based on observation from data, using tools which help to explain the data and make prediction from it. The data is taken as set of examples that will be used to extract the structural pattern. Learning is interpreted as actual description of the structure of a data which is used for classifying new examples. This structural description is used to explain and understand the patterns of examples on the given dataset as well as future prediction of new instances. This is one of the advantages of machine learning over statistics [30].

There are various supervised machine learning algorithms for prediction problem. Most of the Supervised Machine Learning (SML) algorithms are statistical SML algorithms which include Naive Bayes and decision tree algorithms such as C4.5 and Classification And Regression Tree



(CART). There are non-statistical machine learning algorithms such as Artificial Neural Network (ANN), K-Nearest Neighbor (KNN), Support Vector Machine (SVM), etc.

IEEE International Conference on data mining (ICDM) identified top ten machine learning algorithms from most common and influential machine learning algorithms on machine learning communities. Among the ten algorithms we emphasized on the five supervised machine learning algorithms which are used for classification purpose. These algorithms are C4.5, CART, KNN, Naive Bayes and SVM. The rest are unsupervised machine learning algorithms or couldn't be used for classification purpose [21].

Classification is the process of extracting a model (or function) that describes and distinguishes data classes. The model is used to predict the class of objects whose class label is unknown. The constructed model is based on the analysis of a set of training data that incorporates set of data objects whose class label is known [31].

2.3.2.1 Machine Learning Algorithms

The brief descriptions of the five recommended classification algorithms are presented below.

The C4.5 classification algorithms [31]

C4.5 is one of the most popular decision tree algorithms. Decision tree is a common way to represent information in a flowchart-like tree structure, where each internal node (non-leaf node) denotes a test on an attribute value, The outcomes of the test are represented by branches of each node and class labels are represented by leaf nodes (or *terminal node*).

The decision tree is implemented based on a greedy algorithm which constructs decision tree in top down recursive divide and conquer manner. The tree is constructed by selecting the attribute which minimizes the information needed to classify the tuples in the resulting partitions and the expected number of tests needed to classify a given tuple. Different attribute selection measure is used to select attribute which provide the least impurity in the partition such as information gain, gain ratio and gini index. The selected splitting criterion using attribute selection measure method should make the resulting partitions at each branch as "pure" as possible. A partition is pure if all of the instances in it belong to the same class.

Decision tree is popular because of various reasons. Construction of the tree does not require any domain knowledge therefore it is appropriate for exploratory knowledge discovery. Handling high dimensional data, classification performance, understandability of its knowledge representation and its simplification and fast response in learning and classification phases are reasons which make decision tree more popular. Decision tree algorithms are used for many applications such as fraud detection and medical diagnosis however; successful use may depend on the data at hand. The decision tree induction has important feature in which when new training data is given, the decision tree which is acquired from learning on previous training data will be restructured rather than learning a new tree from scratch.

C4.5 is an algorithm used to generate a decision tree for classification purpose. It is the successor and an improvement of ID3 (Iterative Dichotomiser), a basic decision tree algorithm. C4.5 uses *gain ratio* as attribute selection measure and incorporates various enhancements to ID3 such as handling missing, continuous and discreet values in addition to categorical values, being robust in the existence of noise data and pruning of irrelevant branches after tree construction and improving in computational efficiency. Gain ratio which considers the probability of each attribute value is an improvement of information gain which leads to bias towards the attribute with many values. C4.5 is adopted in different researches conducted for classification purpose [1, 14, 32].

Attribute selection method is a procedure to select the attribute that best discriminates the given tuples whose class label is missing. The process of decision tree generation which perform repeatedly splitting on attributes, is equivalent to partitioning the initial training set into smaller training sets repeatedly until each leaf is pure or some stopping criteria is met. During attribute selection, attribute with the highest gain ratio will be selected.

Given an attribute A that has n distinct values, the data tuples S can be split into n sub data set S_i where S_i have the i^{th} distinct value of A.

Gain ratio of an attribute A can be calculated as equation (2.2) shows.

$$GainRatio(A) = \frac{I(S) - Entropy(A)}{-\sum_{i=1}^{n} \frac{|S_i|}{|S|} \log_n \frac{|S_i|}{|S|}} \qquad (2.2)$$

The description is presented in equation (2.3) and equation (2.4).

$$I(S) = -\sum_{i=1}^{n} \frac{|C_i|}{|S|} \log_n \frac{|C_i|}{|S|} \quad (2.3)$$

$$Entropy(A) = \sum_{i=1}^{n} \frac{|S_i|}{|S|} I(S_i) \quad (2.4)$$

Where n is number of classes. |X| is the total number of tuples in X dataset.

Classification And Regression Tree (CART)

CART is another decision tree based machine learning algorithms. In CART, binary tree construction is performed in a way that each internal node has exactly two outgoing edges. The binary splitting is performed using *gini index* attributes selection measure. Gini index (intelligent IBM miner) is used to compute the weighted sum of the impurity of each partition. The one which has less impurity will be the splitting criteria on current node. An important feature of CART is its ability to generate regression tree which predicts real number [33]. In Gini index attribute selection measure used to select the attribute that maximizes the reduction in impurity of the attribute (or, equivalently, has the minimum Gini index) as the splitting criteria [31]. Equation (2.5) shows how Gini index of attribute A is calculated.

$$\Delta Gini(A) = Gini(S) - Gini_A(S) \quad (2.5)$$

Where, Gini(S) measures the total impurity of the data tuple S and evaluated as equation (2.6)

$$Gini(S) = 1 - \sum_{i=1}^{n} \left(\frac{|C_i|}{|S|}\right)^2 \quad (2.6)$$

$Gini_A(S)$ is a measure of impurity if S is split based on the value/values of attribute A. It is defined as equation (2.7) shows.

$$Gini_{A \in V}(S) = \frac{|S_1|}{|S|} Gini(S_1) + \frac{|S_2|}{|S|} Gini(S_2) \quad (2.7)$$

Where n is the number of classes, |X| is the number instances in X dataset and S_1 is the set of tuples that fulfill the constraint on attribute A whereas S_2=S-S_1.

Support Vector Machine (SVM)

SVM is a machine learning algorithm for the classification of both linear and nonlinear data. J. Han and M. kamber in [31] stated that SVM is a promising (highly accurate) algorithm for classification. However, the authors argue that its extreme slow processing (even for the fastest SVMs) enforces not to be used for large dataset. The author also recommended that many classification approaches such as artificial neural network and SVM must be used to data represented in a single table since they expect data from a single table. SVM algorithm works as follows.

In order to transforms the original training data into higher dimension, non linear mapping is used. Viewing input data as sets of vectors in an *n*-dimensional space, an SVM will construct a separating hyperplane (decision boundary which separates the instances of one class from other) using essential training instances in that space, and searches the one which maximizes the *margin* between the two data sets. The hyperplane that has the largest distance to the neighboring data points of both classes is the optimal separating. SVM can be used for classification and prediction which can be applied in number of areas such as handwritten digit recognition, object recognition, and speaker identification, as well as benchmark time-series prediction tests [31].

Naïve Bayes

Nave bayes is a statistical classifier which uses a simplified version of bayes formula to classify the new instance to particular class. The prior and posterior probability of each class is calculated, given the attribute values presented in the instance; the class with the highest probability will be assigned to the novel instance [32].

$$\hat{C} = \underset{C_i}{argmax}\, p(C_i) \prod_{j=1}^{n} p(V_j|C_i) \qquad (2.8)$$

Equation (2.8) shows the Naive Bayes formula which assumes that the effect of each attribute's value of an instance on a given class is independent of each other.

Naive Bayes classifier is a simple but effective classifier which can be applied in various application of information processing including information retrieval and natural language processing [31].

K-Nearest Neighbor (KNN)

KNN is a classification algorithm which works based on learning the analogy, comparing a given test instance with training instances that are similar to it and finding the k most similar instances (where k is number of instances). The training instances which are denoted by n number of attributes are represented by n dimensional spaces. Each instance in training data represents a data point. In such a way all training instances will be stored in an n- dimensional pattern space. Classification of new instance in this algorithm will be carried out, the classifier searches for the pattern space for k instances that are closest to the unknown instance. Euclidian distance measure [31] is used to select the nearest value of numeric attributes. If the attributes are nominal, 0 represents the distance if the attribute values of the instances are the same and 1 if they have different attribute value. Equation (2.9) shows the distance measure employed in KNN algorithm.

$$D(x,y) = \sqrt{\sum_{j=1}^{n} f(x_j, y_j)} \qquad (2.9)$$

Equation (2.9) gives the distance between two instances of x and y where x_i and y_i are the value of the i^{th} attribute of x and y instances respectively. $f(x_i, y_i)= (x_i-y_i)^2$ if x_i and y_i are numeric whereas for nominal valued attributes $f(x, y)=1$ if the attribute value of x_i and y_i are different and $f(x, y) = 0$ if they are the same value[32].

2.3.2.2 Performance Evaluation techniques

Evaluating learning algorithms' performance is a basic aspect in machine learning [32]. There are various evaluation techniques to evaluate the performance of a classifier such as Accuracy, True Positive(TP) Rate, False Positive (FP) Rate, Precision, Recall, F-Measure, ROC, confusion matrix, Time taken to build model, Number of Leaves and Size of the tree. Each metrics has strength and weakness [34]. The metrics are not also applicable to all Machine learning algorithms. The appropriate metrics should be selected.

Accuracy

The accuracy of a classifier is a basic performance measure which computes the percentage of test set tuples that are correctly classified by the classifier. [31]. Accuracy of a model can be calculated as equation (2.10).

$$Accuracy(\%) = \frac{Count\ of\ Truely\ Classified\ Test\ set}{Total\ Count\ of\ Test\ Set} \times 100\% \qquad (2.10)$$

True positive, True Negative, False Negative and False Positive

These performance metrics are more appropriate for a two class classification problem where the 1st is positive and the alternate is labeled to negative.

True positive indicates the number of correct positive predictions (classifications); true negative is the number of correct negative predictions; false positive is the number of incorrect positive predictions; and false negative is the number of incorrect negative predictions [22].

Precision, Recall and F-measure

Precision: denotes how many of the test data correctly classified by the classifier from the total test and computed as TP/TP+FP.

Recall: deals about how many of the actual correct value classified correctly and computed as TP/TP+FN. Recall and precision are commonly used performance evaluation metrics in information retrieval. Recall, true positive rate and sensitivity are defined in the same manner but used in different domains.

F-Measure: is a weighted measure of both precision and recall, which is computed as $\frac{2PR}{P+R}$, where P is precision and R is recall.

True Positive and False Positive rate

True Positive Rate (TPF): indicates how many of the actual positive class is correctly classified. It is computed as TP/TP+FN.

False Positive Rate: how many of the actual negative class incorrectly classified as positive. It is computed as FP/FP+TN.

Area under Receiver Operator Characteristic (ROC)

An ROC curve shows the trade-off between the proportion of positive tuples that are correctly classified (the true positive rate) and the proportion of negative tuples that are incorrectly

classified as positive (the false-positive rate) for a given model [31]. The curve is defined as a graph which is drawn by using a model's true positive rate as the y coordinate and its false positive rate as the x coordinate, under all possible score thresholds [35].

The accuracy of a model can be assessed by measuring the area under ROC. The closer the area into 0.5, the less accurate the corresponding model is and it denote the random prediction. The closer the area into 1.0 the higher the accuracy it has and the model with perfect accuracy will have an area of 1.0. The model with excellent prediction will have an area of 0.9 [31].
Figure 2.3 shows an ROC curve with excellent prediction (i.e. Area under ROC=0.9685).

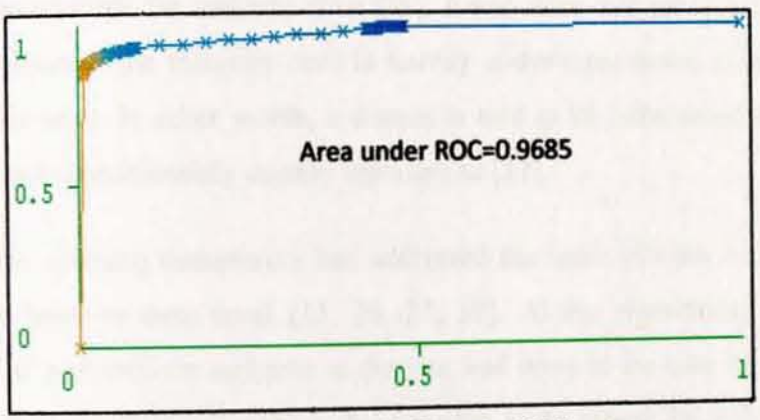


Figure 2. 3: Sample ROC curve

S. Wu and P. Flach [34] argue that ROC is a performance evaluation metrics and recently used tool to select the possible model in machine learning application since ROC is not sensitive for imbalanced class distribution.

Confusion matrix: Confusion matrix is a more specific performance indicator which will have NxN table grid where the row headings are actual class labels and columns are predicted class labels. A cell at i^{th} row and j^{th} column indicates the number of instances predicted as class j where they are actually from class i. One can derive different performance indicators during experimental analysis stage from this matrix.

Time taken to construct a model: refer to the amount of time which is used to build a model. The shorter the time, the more desirable it is.

Number of Leaves and Size of the tree: can be used as performance evaluation metrics for a classifier when a classifier is constructed based on decision tree algorithm. A decision tree which

has large number of leaves will have over fitting problem and tree size (height) is larger implies model will take more time to classify a tuple whose class label is unknown.

2.3.3 Issue of training data

Training data must be checked for its goodness in terms of class distribution and sufficiency. These two basic issues are briefly discussed in the following subsections.

2.3.3.1 Techniques for mitigating class imbalance problem

Scholars observed that class imbalance may affect negatively the performance of a classifier by creating bias towards the majority class [36]. A two-class data set is said to be imbalanced when one of the classes (the minority one) is heavily under-represented in relative to the other class (the majority one). In other words, a dataset is said to be imbalanced if the classes on a given dataset are not approximately equally represented [37].

The machine learning community has addressed the issue of class imbalance in two ways: in algorithmic level or data level [15, 36, 37, 38]. At the algorithmic level, solutions include adjusting the probabilistic estimate at the tree leaf level in the case of classification trees [15], modifying the classifier i.e. making the classifier to be robust for imbalanced data set through assigning cost for different cases of the data set [39] or it may be training a classifier by majority class and letting the classifier to identify the minority class by detecting them as anomalies [38].

The data level balancing deals about balancing the imbalanced dataset through resampling the dataset. Resampling refers to the process of changing the existing probabilities of the majority and minority class in the training dataset by decreasing or increasing the number of instances in the majority or minority class respectively [38]. There are two types of resampling techniques: under sampling and over sampling. Under sampling is the process of reducing the number of instances in majority classes whereas oversampling is increasing the number of instances in the minority classes; until the classes are approximately equally represented.

Data level imbalanced dataset handling technique is recommended by Japkowicz in [38], it is much easier to do some change (make resampling) on the data than to change an already working algorithm. More over Maloof in [40] stated that resolving the imbalanced dataset problem in

algorithm level and in data level showed the same experimental result, however resampling is a simple and attractive choice than modifying the algorithm.

Though both data level balancing techniques can be used to solve the issue of class imbalance, they have their own drawbacks. Under-sampling may throw out potentially valuable information, whereas over-sampling artificially increases the size of the data set and as a result it increases the computational burden of the learning algorithm as well as the required memory size [36].

Over sampling is better where accuracy is a great issue and memory is sufficient to process a given oversampled dataset. There are different types of over sampling techniques; such as random over sampling and Synthetic Minority Over Sampling TEchnique (SMOTE). Random oversampling increases the number of instances in the minority class by duplicating the existing minority class instances. SMOTE is a resampling technique by increasing the number of instances in the minority class by generating new instances rather than replicating the existing instances [37]. A thorough discussion of different over sampling techniques is beyond the scope of this thesis.

Random over sampling may cause over fitting problem since the existing instances are duplicated. In the case of SMOTE, since the instances are synthetically generated and the generated instances are not the exact copy of the existing instances, over fitting problem cannot be occurred.

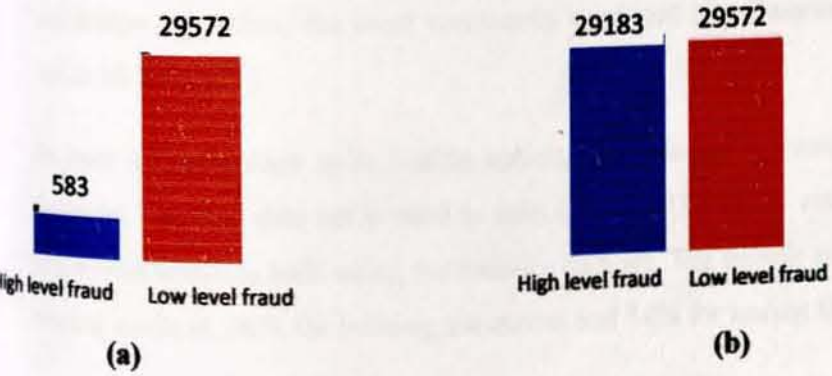


Figure 2. 4: Example of SMOTE Analysis (a) before SMOTE (b) after SMOTE

For instance Figure 2.3(a) shows the heavily under-represented class distribution of a dataset which is represented in ratio of 16:84. Figure 2.3 (b) shows the class distribution after SMOTE analysis was applied.

2.3.3.2 Techniques to check the goodness of the Size of a dataset

The data is said to be sufficiently large, if a model cannot acquire any knowledge even if more examples are added. The sufficiency of the dataset can be checked by using learning curve analysis. Learning curve is sketched by using the data size and the performance obtained as "x" and "y" coordinates respectively. It is the analysis which is done by increasing the training data size starting from certain percent of the total training dataset and check whether it converges or not. If it converges the data is said to be sufficient. If the performance is increasing while the data size is increased the data is said to be insufficient because the model is not stop acquiring new knowledge from the added data and needs more examples to converge. The issue of insufficient data size can be solved either by adding more data or through using appropriate testing mode selection.

There are two popular testing modes: k-fold cross validation and percentage split testing modes.

The k-fold cross validation is a widely used experimental testing option where the dataset is randomly divided into k mutually exclusive or disjoint blocks of instances, then the model is trained using k-1 blocks and the remaining block is used to test the performance of the model; this process is repeated k times. Finally, the recorded measures are averaged. In K-fold cross validation algorithm, the most commonly used and experimentally tested for optimality is the value 10 for k [23].

In case of percentage split testing option, the data set is randomly divided into two disjoint datasets. The first data set is used to train (construct) a model and the second is used to evaluate the model which is built using the training data set. The default splitting percentage values in this testing mode is, 66% for training the model and 34% for testing the model [23].

K-fold cross validation is recommended for insufficient data because, in this testing mode the final reported result is the averaged of K experiments. Either of k-fold cross validation or percentage split testing option can be used if the data is sufficient.

2.4 Related works

The rapid growth of business volume posed customs to have shortage of resource's administration while the rapid development of world trade and economy required the higher level of efficiency in facilitating customs clearance. On the other hand, the rapid growth of information technology makes the customs database to be vast in volume, rapidly growing and having complex relationship; which required special analysis tool which analyze the vast data and attempts to be used for future prediction rather than using manual analysis. Accordingly, customs authorities face challenges of detecting customs declaration's frauds with limited examination of imported goods by available scarce resources [2].

Though researches on fraud detection through machine learning techniques were done in different countries of customs such as China [1, 2, 11, 41], India [15] and Brazil [29], in the case of Ethiopia, there is no published research which is officially conducted in the circle of scientific literatures and publications yet. Though researches done on the custom data of a given country might be used as indicator for another country, it is not possible to directly apply findings of one country to another country. This is primarily because different countries have different concern on custom service, the fraud detection techniques used in different countries are different based on the type of the data used and the importers' intent to fraud differs from country to country.

According to the review of the researcher on related works, several fraud detection and prediction researches were conducted by scholars based on customs data of various countries such as India, China and Brazil. Most of the researches reviewed, share common problem which is large amount of customs transaction and limited amount of resources to control fraud by examining all cargos of the transaction. The reviewed researches can be broadly classified based on the general approaches they have used into four categories.

The first one is classification based fraud detection and prediction [1, 14, 15,] which attempts to build a classification model from customs relation based observational data. Moreover, since tax audits are very expensive in both financial and human resources, [14] proposed a classification approach for fraud detection of tax which attempts to identify the subjects which return high recovery.

The second approach is clustering based fraud detection approach which categorizes the given instances into different clusters. In [11], cluster method of fraud detection proposed to solve the confliction between the rapid increasing of cargo quantities while the custom's inspection force is limited. The proposed solution was, classifying cargos into seven different clusters, and thus the custom's inspection force will focus on the cargos' cluster which is identified as high risk level.

The third approach is outlier based fraud detection which identifies suspicious operations through outlier detection. It works by comparing each new customs transaction with the previous operations to identify those that might be considered as fraudulent [29].

The last approach is association rule based classification which is proposed in [2] and [41] in order to construct a model which identifies commodities risk level based on past observation for examination of limited goods with higher probability of fraud.

During related work analysis, the researcher observed that, all the four approaches have the same role which is maximizing organizational benefit while minimizing the cost in accordance with controlling fraud. However, scholars may consider the type of the data used for analysis during approach and/or algorithm selection. In particular, the type of the data (i.e. whether the instances are labeled with fraudulent or non-fraudulent class label attribute or not) determines the type of machine learning techniques (Supervised Machine Learning (SML) or unsupervised machine learning (USML)).

In case of SML, the process of model building is to explain the value of target attribute in terms of others independent attributes and later on the model will be asked to predict or classify new instances with unknown target attribute. In case of USML, the process consists of expecting the system to identify patterns or relationships or regularities in the data, those discriminate different structures of the given data. Scholars argue that classification is a typical SML technique whereas association rules and clustering are the typical unsupervised machine learning techniques [14].

In [14], since the target attribute was not given, the target attribute was formed through derivation from the given attributes in order to apply the classification approach (SML

techniques). [1] and [15] also used SML techniques. However, [2, 11, 41] adopted USML techniques which doesn't need a target attribute.

In [29], an outlier based fraud detection approach is adopted since the number of fraudulent instances is highly under represented from normal (non-fraudulent) instances. However, there is also another approach to solve this class imbalance problem that is used by [15] at algorithm level and data level and [14] which is balancing the imbalanced classes at data level. Once the unbalanced data problem is solved the predictive fraud detection models were built using classification approach (Decision tree and ANN algorithm in [14] and [15] respectively). But in case of outlier based fraud detection approach, the imbalanced data distribution is taken as fundamental prerequisite [29].

Another important aspect the researcher observed from analysis of the related work is the data that have been used to conduct fraud detection and prediction. In [2], 26613 inspected cargos records in which 2500 (9.4%) records are identified as inconsistent between the declaration and actual commodity (i.e. fraudulent) is used for experimental analysis. Attributes such as enterprise type, commodity classification, trade partner, commodity flow and trading method are attributes used to build the association based predictive model. However, the researchers stated that there was relevant attribute like commodity type which must be included to improve the performance.

In [1], various attributes such as commodity ID, address, corporation ID, commodity, producing area, wrap-mode, corporation, credit, trail and the credit, were used for fraud detection model construction. In [29], weight and price are the stated attributes which are used in the outlier analysis, however, any additional information about the data (number of attribute used, list of attributes and number of instances) are not specified in the article.

In [11], the data sources were the year 2002 of invoice of customs, book of introduction, reduced tax resources and enterprise resources. Total sample of 8615 instances with 12 attributes was used for analysis. The attributes are CODE_TS and CODE_NAME which are identified as mark variance and IM_ITEM_COUNTS, EX_ITEM_COUNTS, IM_USD_PRICE, EX_USD_PRICE, IM_DUTY_ALL, IM_MANU_PRICE, MANU_DUTY_FRE, EX_MANU_USD_PRICE, RED_USD_PRICE and RED_DUTY (identified as analysis variance).

According to [14], the initial dataset for building predictive model to identify subjects (companies) which provide higher recovery consists of 80643 and 175 attributes. From this dataset only 4103 (5%) instances are audited companies records which were used for analysis. However, after data cleaning and attribute selection task were performed 3880 instances and 20 attributes are left for analysis. One of the attributes i.e. 'recovery' represents the amount of tax evaded and obtained out from audit. More over derived attributes like audit cost and actual recovery were used for analysis. The data which was finally used for analysis is also characterized by significant imbalanced on class distribution which is represented in ratio of 82:18.

In [15], the data source contains 13 tables and about 300 attributes which pertaining all importing cargo information. Data concerned with Electrical & Electronic goods with the selected 23 attributes which assumed to be relevant was used for building the predictive model. This data was taken from Air cargo of Bangalore.

According to [14], the initial dataset for building predictive model to identify subjects (companies) which provide higher recovery consists of 80643 instances and 175 attributes. From this dataset only 4103 (5%) instances are audited companies records which were used for analysis. However, after data cleaning and attribute selection task were performed 3880 instances and 20 attributes are left for analysis. One of the attributes i.e. 'recovery' represents the amount of tax evaded and obtained out from audit. More over derived attributes like audit cost and actual recovery were used for analysis. The data which was finally used for analysis is also characterized by significant imbalanced on class distribution which is represented in ratio of 82:18.

The other important issue which is observed during the analysis of related works is the data preparation task which makes the data suitable for the required fraud detection task.

[1] and [2] argue that customs data is complex and requires efficient data preparation techniques to achieve the required performance.[1], after reviewing various fraud detection researches such as [14] concluded that data preparation and modeling are crucial tasks on fraud detection system compared to the performance gained using modifying different algorithms. As we observed in

the reviewed papers, scholars used different data preparation technique to obtain suitable data for the intended purpose.

To improve the accuracy of the classifier and make its running time acceptable, [1] applied cleaning the data, descritizing continuous attribute, constructing a star schema and preparing the data through cluster analysis. According to [2] and [41], since the continuous attributes (such as price) in the original data have many distinct values, the values of this attribute descritized to be suitable for rule generation task. In [11], data preparation task includes selecting the relevant attributes, sampling and clearing the data to be suitable for intended clustering task.

The major data preparation task which is applied in [15] was solving the class imbalance problem using two different techniques: at algorithm level and at data level. At the algorithm level, solution includes adjusting the probabilistic estimate of classification trees at the tree leaf level. At the data level, solution includes resampling the data through under sampling the majority class or over sampling the minority class.

According to [14], the target dataset prepared by constructing derived attribute like actual recovery, dividing the pre-classified data (the target attribute is already known) into training which is used to build predictive model and testing dataset to evaluate the performance of the built classifier. Audit cost which is provided based on the estimation of domain expert is used to generate the actual recovery. Actual recovery is a derived attribute which is obtained by subtracting the audit cost from recovery. Attribute 'actual recovery' which discriminate tuples of the classes takes a value zero and one in case of fraud is detected and not detected respectively. This attribute is formulated as if actual recovery is greater than 0, the target class will be 1 and it will be 0 other wise. Data cleaning which consists of removing noisy tuples (i.e. tuples with excessively deviating values) and rows with many null values and attribute selection (column removing) were among data preparation activities which were made in [14]. The sufficiency of the dataset is checked through incremental sampling approach (learning curve analysis) which was performed through experiments of randomly generated subsets of the dataset (10%, 20%, 33%, 50%, 66%, 90%) of the total dataset.

As can be seen in Table 2.7 algorithms such as C4.5 [1, 15], C5.0 [14], CHAID [1], ANN [15], apriori [2, 41] are machine learning algorithms which are used for fraud detection purpose.

However, decision tree algorithms such as C4.5 and C5.0 is the common, simple and fast machine learning algorithm which is commonly used as compared with others.

In [14], parameter tuning and over sampling technique were used to improve the performance (i.e. predictive accuracy) of the model, by reducing the misclassification rate towards the majority class.

Table 2. 7: Description of approaches in reviewed related work

Approaches	Machine learning algorithms	Tools used	Performance evaluation metrics
Classification	• Decision tree algorithms (C4.5, C5.0 and CHAID) • ANN	Not specified	• Accuracy • Size of tree • confusion matrix • misclassification rate
Association rule based classification	Apriori	Not specified	• Accuracy • confidence • support
Clustering	Not specified	SAS.10.0	Not specified
Outlier based	Not specified	Not specified	Not specified

Accuracy is a basic performance measure which is used in the papers reviewed in [1, 2, 41]. [2] and [41] used support and confidence to select interesting rule which is generated by priori algorithm. Moreover, in [1], size of tree (i.e. the small maximum number of branches from a node) is used to select and evaluate the model in addition to predictive accuracy of the model. [14], adopted performance evaluation metrics such as confusion matrix and misclassification rate in addition to other non-domain specific evaluation metrics such as profitability.

This research focused on identifying the gap in the existing system (ASYCUDA selectivity). The review on the related work provides significant insight to the researcher on various issues to be considered in the design stage of the research process. The vast majority of approaches available allow the researcher to systematically select feasible set of approach where the comparative study on approaches are based on the four IEEE recommended machine learning algorithms [21].

CHAPTER THREE
PROBLEM AND DATA UNDERSTANDING

3.1 Problem Domain Understanding

One of the phases in building predictive model from data is, understanding the problem domain. Without a deep understanding of the problem domain, no one can achieve the desired objective even they use highly sophisticated tools, models and techniques [42]. Once an in-depth knowledge in the problem domain obtained, data analysts can clearly set the objectives and attempts to be made to achieve the defined goals [17]. On this section, the researcher focuses on understanding the study objectives and requirements from a business perspective, and then converting this knowledge into machine learning problem definition, and a preliminary plan designed to achieve the objectives [16].

Domain experts were consulted to have a clear understanding of the problem domain. The domain experts communicated include individual in each sub division that follows the **declaration clearance flow** and other related staffs, some of these are agent (declarant), face-vet officer, risk management officer, examiner, assessor and verifier, database administrator and programmer. In addition to this, the problem and data understanding task has been supported by various user manuals, ASYCUDA++ technical documentation and declaration documents. Declaration is a source document for customs data; it is a form which provides full information about the importing cargo and it is filled by importer's agent (declarant). Incoming cargos will be permitted to be imported inside the country after the declaration associated with that cargo is processed (registered, assessed and paid).

3.2 Work flow for declaration clearance

There are various steps to be followed in customs clearance and control procedures of importer declaration. In all customs clearance offices of ERCA, all the customs procedures starting from registration to clearance of goods are taking place in a similar way. They have defined similar customs organizational structures which have identical customs sections and desks. They also use the ASYCUDA software to undertake customs clearance of declarants' document and

physical examination [43]. Figure 3.1 depicts the flow of processes in customs clearance and procedures of ERCA.

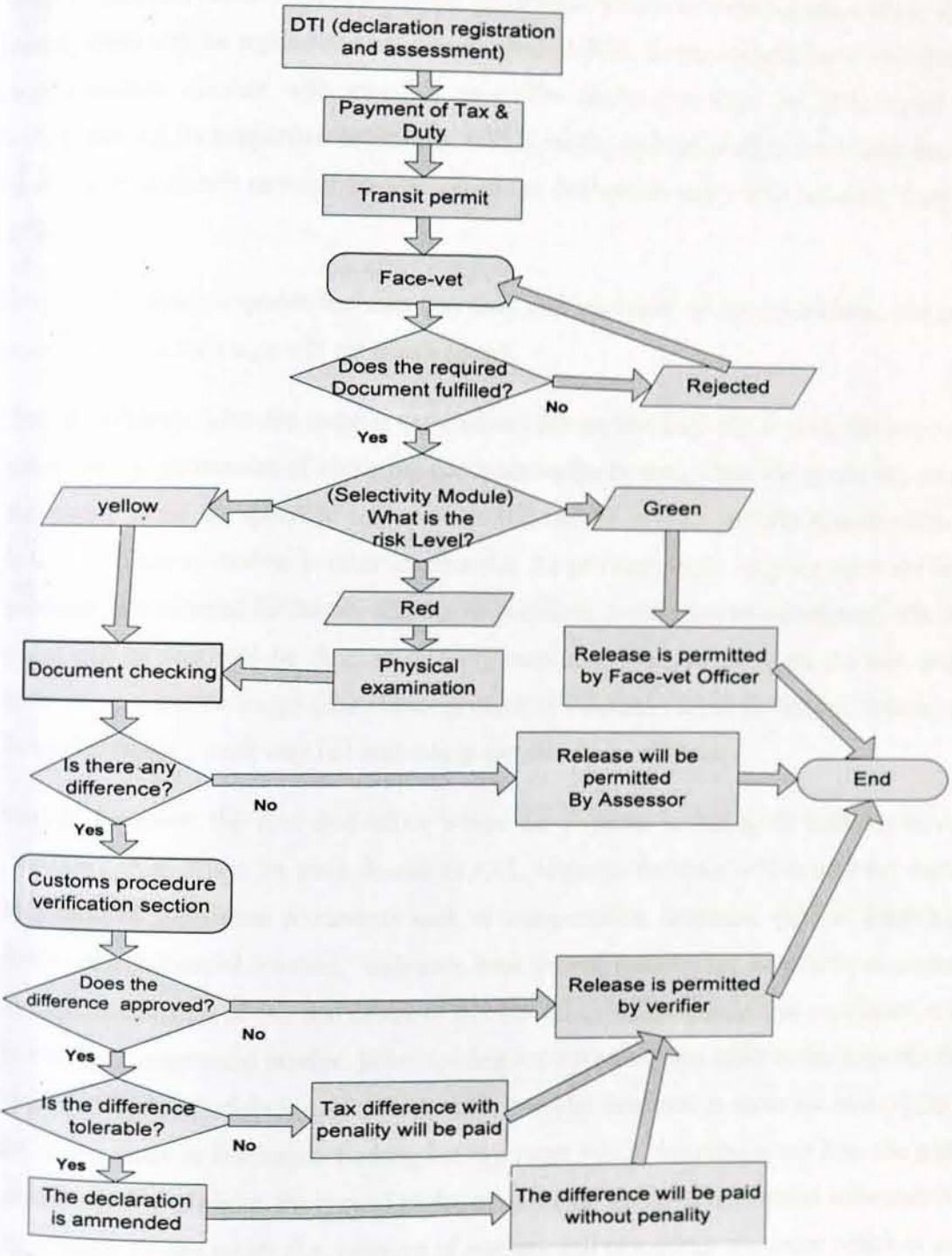


Figure 3.1: Flow of work for customs clearance

Details of the various steps of the customs clearance work flow is given below.

Step 1: Direct Trader Input (DTI): the declaration which has detailed information about the consignment will be registered and assessed through DTI. Every declaration is identified by its own reference number with specified year. The declaration must be filled based on the information on its supportive documents. DTI is an out sourced module from customs. So, the importer or declarant can register and assess the declaration using DTI remotely from her/his office.

Step 2: Payment: importer/declarant pay duty and tax based on the information declared and assessed so that the cargo will get transit permit.

Step 3: Transit: After the cargo is declared and the tax and duty due is paid, the cargo will get transit permit (permission of importing goods across the border). Then, the goods can rout inside the country along the specified path only, till it is arrived to AAL customs branch; unless it has special privilege to destine in other location (i.e. the privilege might be given since the imported goods are raw material for factory and can be examined in the importer warehouse). The cargo is sealed and its seal will be checked in every customs station which is on the way and some information about the cargo (like scanning machine indication result in 'Millea' station) will be forwarded through track way bill with stamp approval in every station.

Step 4: Face-vet: the first desk/office where the importer or his agent contacts to customs clearance office. When the track destine to AAL, importer/declarant will deliver the declaration with required supportive documents such as transportation document (bill of loading), price document (commercial invoice), insurance, bank permit, packing list, certificate of origin, track way bill, declaration of fact and others to face-vet officers. Documents like certificate of origin, packing list, commercial invoice, bill of loading are delivered from seller to the importer through bank. Certificate of origin is a document which provides information about the country, in which the item is made or fabricated. Packing list is a paper which describes about how the goods are packed, number of packs, the type of packs, etc. Commercial invoice contains information about the payment for the goods (i.e. payment of receipt). Bill of loading is a paper which is used for loading goods. It is prepared by transportation company when the fright payment is made. The insurance paper and the bank permit paper respectively show that the goods are insured and

foreign currency is permitted to purchase the goods. Declaration of fact is the description of goods prepared by importer.

In this phase, the face-vet officers perform manual checking of the declaration. The officer roughly checks obvious errors like filled values that doesn't match with supportive documents and whether the declaration has been fulfilled the required attachments or not, whether the goods are importing within appropriate customs procedure or not and whether the documents which must be attached in original copy is delivered or not. If the declaration fulfilled the requirement, it will be accepted by officers for further clearance processes otherwise it will be rejected. If the document is accepted, the face vet officer will make warehouse allocation and checks the risk level of the declaration which is assigned automatically by ASYCUDA selectivity module. The assigned risk level will be approved by risk officers (step 5). Accordingly based on the approval, the face–vet officer can permit release (if the declaration is approved to be green channel) or the document will be transferred to assessor or examiner if it is yellow or red respectively.

Step 5: Risk assignment approval: Due to random targeting and lack of trust under ASYCUDA, the risk level of the declared goods which is assigned by ASYCUDA should be approved by risk band officers. The officers may change the risk level of the declared goods on the hard copy of the declaration by considering different parameters like sensitivity of goods and country of consignment. After the assignment is approved by the risk officers, the declaration will be sent to the channel in which clearance process is takes place.

Step 6: Physical examination: examination may takes place by opening, counting, and checking the goods against the data inputted in the declaration and the attached document.

Step 7: Validation and document examination: the declared declaration is checked against the attached document and even it might be checked against the physically imported goods by taking samples. Here, the correctness of the tariff classification, the valuation and the appropriate use of CPC will be verified. If there is any problem (infraction) regarding the document or the cargo itself, the document will be passed to 'customs verification procedure section' (step 8) otherwise the 'exit note' or clearance will be given.

Step 8: Verification: this step will continue if examiner or assessor in step six and seven respectively believed that the declaration is associated with certain fraud. So that, in verification

section, investigation will be made to verify the assumption made (i.e. if there is contradiction between the declaration and the actual cargo). After verification, one of the three possible values will be expected: clearance, amendment or difference. Clearance will be given if the fraud is not existed. If fraud which is tolerable by the organization is obtained, amending the incorrect value of the declaration with the correct value will be made. The infraction due to this type of fraud should be paid without penalty. If fraud which is not tolerable is existed, amendment of the declaration cannot be made. Rather the infraction with duplicate amount of penalty should be paid and the difference obtained will be recorded on local data of the branch which is maintained for internal use rather than updating the information on ASYCUDA database. How fraud is determined whether it is tolerable or not has been discussed in section 3.3.3.

3.3 Data collection and understanding

The data understanding task begins with an identification of initial dataset and continues with activities in order to be familiar with the data, to identify data quality problems, and to detect interesting subsets which will be used to build the intended fraud detection models [16].

The data understanding is one of the main phases in building predictive model from data. Since this research is conducted over customs wing of ERCA, the foreign trade database which contains 31 tables is used as a base data. This database holds several variables which include general information about the declaration, duties and tax, financial, procedural, trade operators, transportation, flow, valuation and other details. The local data which is maintained for internal (department) communication purpose in AAL branch is also identified as one of the data sources.

3.3.1 Data collection

As stated in section 1.6.1, the goal of this research is to build fraud detection models from customs data of ERCA which predicts the fraud behaviors of importing cargos and can classify fraud risks of cargo into different levels. The data source is identified as the customs data of the authority. Since customs have a very huge database, it is important to select the data set which will be appropriate in building the fraud detection models. In order to select the interesting subset of the data, problem understanding task through observation and interview with customs officers

in different positions as well as in different branches of ERCA was conducted, which was an input to select the required subset data from the huge data.

Based on the knowledge of problem understanding, sample data was taken and some analysis which was important to select the appropriate data set and relevant attributes were applied on it. Since the research focused on building models which can predict the fraud behaviors of incoming cargos based on the knowledge learnt from available dataset, the researcher decided that the data set must include the records of the cargos which have been physically inspected. In this line, though the authority has centrally administered database, the researcher couldn't get enough information which is used to identify the cargo whether it is examined physically or not from these central database. Hence, the researcher collected the inspected cargos' records which are recorded for office use (we can call it "local data") in addition to the data in the central database (ASYCUDA data). But the local data doesn't have full information regarding different features of the cargos. Hence, the inspected cargos' full information therefore has been obtained using appropriate data integration procedures.

3.3.2 Data integration

Data coming from different sources should be integrated systematically considering consistency, integrity, and related issues. In this research to build appropriate dataset for the business a two level data integration procedure: high level and low level data integration were conducted. These are described in details as follows.

a. High level (data source level) integration

The responsibility of the high level data integration is, obtaining the declaration information of physically inspected cargos which is registered in year 2011 and partial year of 2012. The integration also considered the version of declarations, whether it is the one declared by importer or changed after inspection.

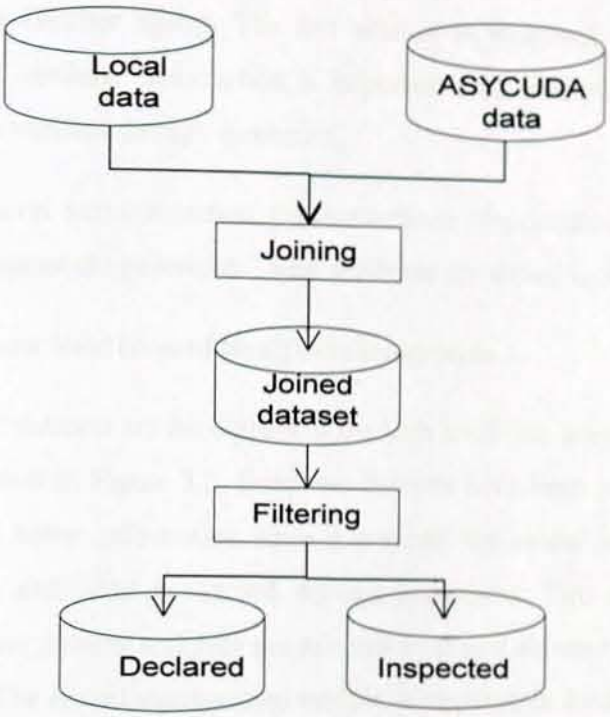


Figure 3. 2: Data source level data integration

As depicted in Figure 3.2, the research conducted on two data sources: local data of AAL branch and ASYCUDA data of ERCA, also called the central database data. The local data is a data which was collected from different department of AAL. The data is recorded for the purpose of report in AAL branch. Whereas ASYCUDA data is, a centrally administered database which is manipulated through ASYCUDA system. In the high level data integration therefore, natural joining of data integration performs inspected cargos information of the local data with the year 2011 and partial year of 2012 entire data set of selected tables. The joining operation is made using attributes *declaration number* and *year*. The filtering phase is responsible to split the joined dataset into three sub dataset. These are:

1. Those whose version information is the first
2. Those whose version information is the latest
3. Others

The filtering process generates only the first and the second category of sub dataset and termed as declared dataset and inspected dataset respectively. The first version is records which are

declared by importers (his/her agent). The last version is supposed to be the record after inspection. These two versions' information is important as it present what was declared by importer and what was verified through inspection.

After the data source level data integration, the two datasets (the declared and inspected) having 48 attributes of each dataset are generated. These attributes are shown in Annex A.

b. Low level (record level) data integration

Declared and inspected datasets are the outputs of the high level data integration. In the low level data integration as shown in Figure 3.3, these two datasets have been joined. Integrating these two versions provides better information since it presents the inconsistency between what is declared by importers and what is verified through inspection. Two records with the same *declaration number*, *item number* and *year* are selected by record selector and provided to record management module. The record management module is responsible for deriving new attributes, selecting values of attribute from either of the two versions and producing the study dataset.

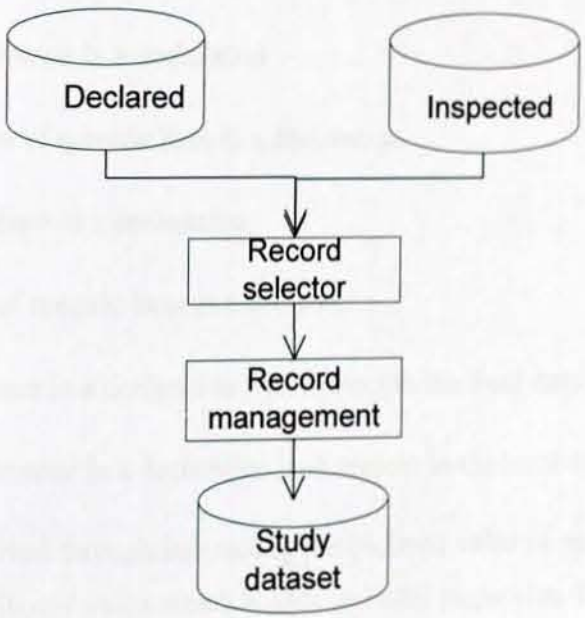


Figure 3.3: Record level data integration

During this level of data integration, values of some attributes which are declared by importer or declarant in the declaration process have been taken from first version. This is because, we are

going to build models which predict fraud behaviors of new incoming cargo with a list of known input values of independent variables where the value of dependent or a target attribute (i.e. class label) is unknown. So that, what is declared by importer/declarant is important for the prediction rather than what is edited (updated) after examination. Some of the attributes which have been taken from first version are *declarant identification number*, *consignee*, *customs office at border*, *HS-code*, *custom procedure code*, *country of origin*, *country of consignment* and etc. Attributes needed to derive attributes which show the inconsistency between the declaration and the actual cargo have been captured from both tables (declared and inspected versions) because these attributes such as revenue lose is calculated through subtracting first version value (value declared by importer/declarant) from last version value (what is verified through inspection) of total tax in a declaration. Attributes which are used to calculate revenue lose are "*total tax in a declaration*" *and "total tax for specific item in a declaration"*. In addition to this, attributes like "*total value of goods*" is taken from both versions and attributes which hold the infraction between the declared amount and what is obtained during inspection are produced. These attributes are:

- Total value difference in a declaration
- Value difference of specific item in a declaration
- Total tax difference in a declaration
- Tax difference of specific item in a declaration
- Total tax difference in a declaration with respect to the local data and
- Total value difference in a declaration with respect to the local data

These attributes are derived through subtracting the declared value of particular attributes' value from corresponding attributes' value which is obtained after inspection. The attributes provide an opportunity to construct class labels and to resolve inconsistency problem by correcting the class label or by removing noisy instances through cross checking the class label with these attribute values.

Attributes which contain fraud information of certain cargo and other attributes like examiner information have been taken from the last version. These attributes are not expected to be declared by importer during declaration. Fraud information of certain cargo is obtained after inspection and document checking is takes place.

After the completion of the two level data integration, the study dataset has 74315 instances and 63 attributes, which has been processed through the rest data preparation process. The attributes are presented at Annex B.

3.3.3 Data understanding

As we have mentioned in Section 3.3.1, the source of the data for this research is customs data of ERCA while the source of customs data is the declaration in which imported or exported goods are declared. A declaration is a document which contains different information about a cargo that is imported or exported. In our case, it deals about imported cargos. It is identified by a declaration number (registration number) with in specific year and facilitated by a declarant. Declarants are persons or companies who work on behalf of importers in performing importing process. An importer may delegate different declarant to process different declaration. A declarant may process declaration of different importing companies. A declaration must be presented with different attached supportive documents like invoice, certificate of origin, insurance and fright document. Based on the information which is filled in the declaration, the structure of the data might be highlighted as follows:

There can be up to 999 kinds of items pertinent to a single declaration (i.e. an importer may import up to 999 distinct items using a single declaration) [44].

Every item type in a declaration will have an HS-code and an item number which has been identified in a particular declaration. Due to the type of item (which is identified by HS-code) and the customs procedure code (CPC), up to 5 types of taxes (including duty) might be levied. A declaration may have various versions which will be created during changes (updates) are made on the declaration. When the declaration is created, its version will be 0. Then if change is made zero will be given for the newer version and the old one is replaced by 1. In the same fashion, the old one is replaced by the next consecutive number (k+1) whenever change is made in a particular declaration. Where, k is the old version number.

In this study 23807 physically inspected cargos' declarations have been taken for analysis. Out of these amounts of declarations, 5659 declarations are identified as they are declared with fraud. The rest 18548 declarations have been inspected and proved that they are free from fraud. The analysis of the data set during data understanding phase of the study looks as presented in Figure 3.4.

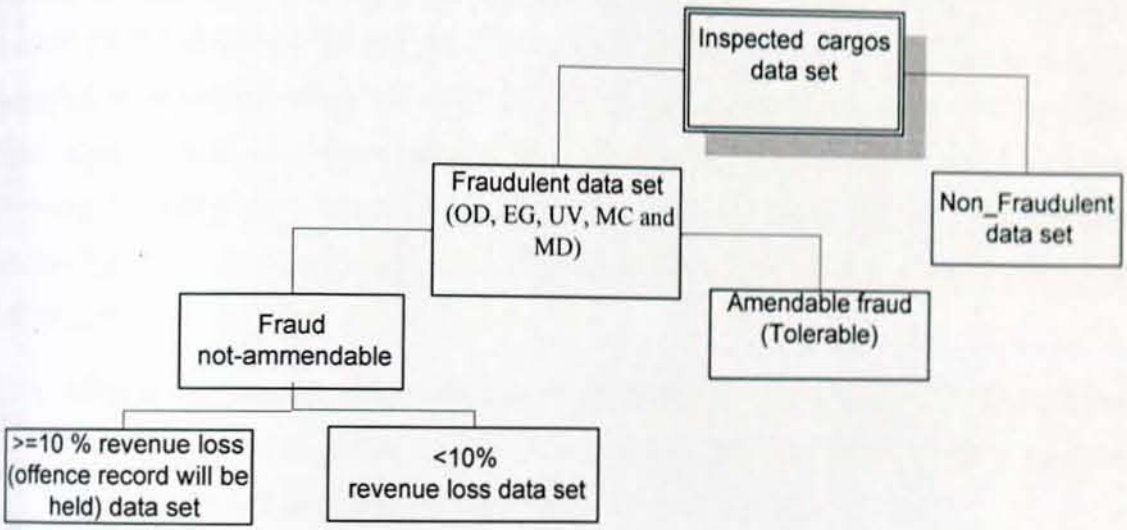


Figure 3. 4: The collected dataset analysis in data understanding

Those declarations identified as fraud also categorized into different fraud categories such as Origin Difference (OD), Extra Goods (EG), Undervaluation (UV), Misclassification (MC) and Mis-Description (MD). Though frauds differ in their type, the usual final goal is the same which is minimizing the amount of tax and duty due.

OD is a type of fraud which may made when importers tries to pay the smaller amount of tax and duty by changing the actual country in which the item is made into the incorrect one. EG is made when the amounts of goods are exceeded from the declared one or extra undeclared item/s is/are obtained during inspection. UV is a fraud code when the declared value (i.e. the summation of Cost, Insurance and Freight (CIF)) value of incoming goods is less than the actual one. MC is occurred when an item is misclassified into incorrect HS-code during declaration. HS-code is used to identify goods internationally. MD is made due to providing/declaring incorrect description of goods.

In addition to categorizing fraud into different fraud types (i.e. OD, EG, UV, MC and MD), customs fraud is further categorized as fraud which is amendable and not-amendable (what they call it "difference"). Fraud which is done intentionally will not be allowed amendment of the declaration. An importer who made this type of fraud will be penalized to pay duplicate amount of the duty and tax that must be paid or loss of revenue (i.e. the difference between the amount due and the paid one). In this case, information associated with revenue loss is not recorded in the ASYCUDA database. On the other hand, there are cases in which revenue loss might be happened on situations which are occurred beyond the capacity of importers. In other words, fraud might be occurred without the intentional activity of importers though it may include who are doing it intentionally. In such cases, amendment of the declaration will be allowed; and the revenue lost with this regard will be paid without any penalty. Whether an importer made fraud deliberately or not can be determined in two ways.

1. Whether the amount of tax and duty of the declared one will mismatch with the actual value or not. It is considered as it is done deliberately if the declared value has lesser amount of tax or duty than the actual one.

2. Whether the mistake which has been made, is tolerable by the authority or not. For instance, MC fraud is taken as tolerable fraud since classifying goods into correct HS-code is not an easy task even for customs officers because of its complexity. In addition to this, an importer may purchase goods in lower price than the reference value. Due to this, the importer may declare undervalue of goods and fallen on "undervaluation" (UV) fraud. Since it is tolerable type of fraud, the fraud will be treated under amending the value with the reference value and the revenue lost due to declaring smaller amount of items' value will be paid without any penalty.

According to the amount of revenue lost, declarations which are categorized under ***Not-amendable*** (see Figure 3.4) are further categorized as greater than or equal to 10% and less than 10% revenue loss, as compared to the total amount of duty and tax that must be paid in the declaration. Frauds associated with greater than or equal to 10% revenue lose is assumed to be the highest fraud and importers who did this fraud will be recorded as offence in their risk profile.

Though the analysis can be taken in the declaration level (i.e. a group of goods which are imported in a declaration can be considered as a record), there might be information which will be lost since a declaration can contain a number of items (i.e. up to 999 item can be imported in a declaration). So, our analysis has been held on item level (i.e. each item in a declaration will have a record). The study has been performed on 23807 declarations having 74315 records. Therefore, we will be advantageous through incorporating information of each item.

CHAPTER FOUR
DATA PREPARATION

4.1 Introduction

Today's real world databases are highly vulnerable to noisy, incomplete, and inconsistent data due to the hugeness of the data. Noise is a random error or variance in measured variable. Therefore, data preparation task is needed to handle such problems. Data preparation is the process of making the data suitable for model building task. It includes activities like data cleaning, data transformation and data reduction. Data cleaning task attempts to handle incomplete (missing) values, smoothing the noisy data and resolving the inconsistencies problem in the data. Data transformation is a process of transforming the data into a form which is appropriate for model building task. Data reduction is concerned with obtaining a reduced representation of the original data set [31].

The data preparation task improves the accuracy, efficiency and scalability of the classification model. Classification efficiency and scalability will be achieved when the time spent to prepare the data plus the time taken to make classification task using the prepared data is less than the time required to make a classification task using original unprepared data [31].

Data preparation in this research includes activities such as selecting relevant attributes from the available list of attributes, transforming the data into the format which is suitable for analysis, resolving data inconsistency problems and handling missing values.

4.2 Attribute Selection

Machine learning is an effective approach to discover hidden and previously unknown knowledge from data and to automate complex decision making processes. Memory optimization, complexity reduction and noise removing are major problems to be handled by systems that adopt such approach. One of the common approaches to overcome these problems is to remove irrelevant attributes and noisy examples from data before the model building task is started. Relevancy is usually measured in terms of the relationship between the target class and the other independent variables on the dataset. Attributes which are relevant to the target class

are used to form a concept description of pattern existing in the data. Removing irrelevant attributes will improve the predictive accuracy of the model [45].

When the attribute is irrelevant or redundant, it may degrade the performance of the model by leading into wrong decision; slow down or confuse the knowledge discovery process [31]. So that, irrelevant attributes that do not contribute for classification/prediction task have been identified and removed using attribute subset selection method; and manually through knowledge of data understanding. Annex C, D and E show the list of all attributes, derived attributes and discarded attributes respectively together with justification for the derived and discarded attributes.

4.2.1 Derived attributes/data transformation

When an attribute has many distinct values, it may result degrading the performance of the model by slowing down the time to build a model and by creating over fitting problem during prediction. Using higher concept level values will solve the problem which is occurred due to more distinct values. Moreover, attributes which are natural composite of two or more attributes are split to enhance information content of the dataset.

Since the information which is obtained from month of registration is better than the registration date, *registration month* is extracted from *registration date* and formed in a higher concept level. Attribute consignee (importer) contains three type of information such as importer's Taxpayer Identification Number (TIN), region of the importer and trade type in which the importer is working on. We can split substrings and construct attributes *consignee's region* and *consignee's trade type* that tell about the region of importer and the type of trade category respectively. Forming these attributes from attribute *consignee* provides higher concept level about consignee. More over constructing class label attributes is one of the major tasks which have been performed during data preparation. These class label attributes are: FRAUD, FRAUD-CATEGORY, FRAUD-LEVEL and FRAUD-RISK-LEVEL. These attributes are used interchangeably for four different models as class label attributes. These class labels of different models are described below.

- FRAUD is a class label attribute having two values: YES and No. The values state whether a cargo has fraud or not respectively.

- FRAUD CATEGORY is a class label attribute having ten distinct values which holds ten possible combinations of the fraud categories that might be associated with a cargo. The fraud categories are UV, MD, MC, EG and OD.

- FRAUD_LEVEL is a class label attribute having two values (HIGH and LOW) which state whether the infraction in the declaration is amendable or not (i.e. the fraud level is low or high).

- FRAUD_RISK_LEVEL is a class label attribute having three values which tell about a cargo's fraud risk level. The attribute has three distinct values: high, medium and low.

As we have discussed in Section 3.3.2, the number of attributes after the completion of two level data integration is 63. During data preparation the four class label attributes are derived and other attributes such as month of registration, importer region, importer license type have been derived, which exceeded the total number of attributes to be 70 which will be further processed by discarding the irrelevant attributes. These complete lists of attributes are presented in Annex C.

Another important aspect in attribute selection is discarding irrelevant attributes which are useless for the task under consideration. Accordingly, 24 attributes which are stated in Annex F are left and the rest which are described in Annex E have been discarded. How the derived attributes are derived and the reasons for removal of discarded attributes are described in Annex D and E respectively.

4.3 Data cleaning

In the initial data set for this research, there were records which didn't have class labels though there is information which indicates that the record is accompanied with certain fraud. Deleting records may result to loss interesting hidden patterns. Inconsistencies and missing values can be corrected and filled manually using external references [31]. Records which do not have class label are useless in model building task, unless we fill them with some mechanism.

Data cleaning in this research includes handling missing values and correcting inconsistencies. Therefore, the problem of missing and inconsistency class labels value, were resolved using the

derived attributes which were constructed by subtracting the declared value from inspected value during record level data integration. These attributes are *value difference of an item in a declaration, Total value difference in a declaration, total tax difference in a declaration, tax difference of an item in a declaration, total tax difference in a declaration with respect to the local data* and *total value difference in a declaration with respect to the local data.*

For example, a missing class label has been filled as undervaluation (UV) based on the value of derived attribute *"Total value difference in a declaration" for particular instances.* Since the attribute value is greater than 0, it shows the declared value was undervalued. So that, the class label value (i.e. fraud category) has been determined to be undervaluation (UV). In addition to this, the incorrect class label "not declared" has been corrected as Extra Goods ("EG") by checking the values of the attribute "total amount of items in a declaration" in two versions. The description of the attributes has been annexed in Annex C.

Moreover, out 74315 records which is the total amount of records collected for this study, 282 records have been removed because of the inconsistency which cannot be resolved using the above techniques.

4.4 Summary of the data set

Though data preparation is an iterative process, the final dataset to construct different models has 74033 instances and 24 attributes. Since this dataset will be used for different models, it needs some modification depending on the model requirement; the numbers of attribute and instances for different scenario/model have been changed simultaneously.

Attributes of the initial dataset including their descriptions and current status (i.e. whether they are removed or not) are annexed in Annex C. Attributes of the final (target) data set is presented in Annex F.

After the data preparation, our final dataset for analysis is analyzed and can be structured as presented in Figure 4.1.

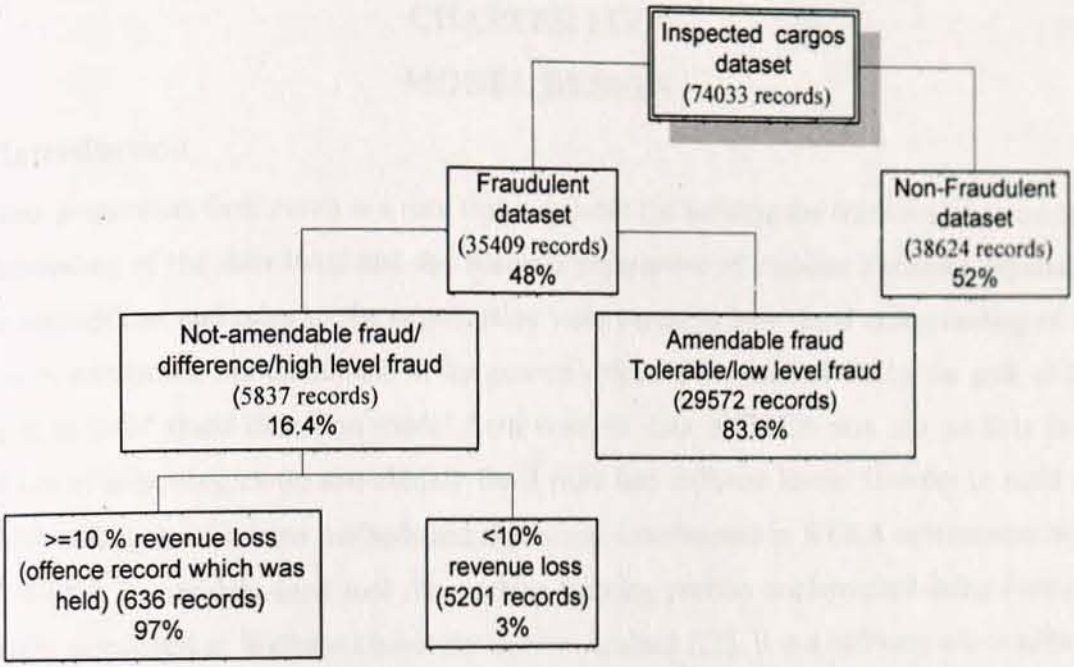


Figure 4. 1: Analysis on the target (final) dataset

As depicted in Figure 4.1, we have 74033 records for analysis with 24 attributes. Out of these amounts of records 35409 records (48% of the entire dataset) are associated with certain fraud and the rest 38624 records (52% of the entire dataset) are non-fraud. Out of 35409 fraud records 5837 (16.4% of the records associated with fraud) records are categorized into particular fraud type because of the fraudulent activities of importers intentionally whereas the rest are tolerable by the authority. On the other hand, based on the amount of revenue lost, these 5837 records further divided into two groups, in which whether the revenue lost is greater than or equal to 10%, or less than 10% according to the total amount of revenue which must be collected in the declaration. Here, the number of records concerned with greater than or equal to 10% revenue loss is 10.9% of the total difference records (records regarding frauds which are not tolerable). The rest 89.1% of these records are records of less than 10% revenue loss. Here, the issue of class imbalance is of great attention.

Based on the above categorization of the dataset, the data is prepared in the format which will be suitable for four scenarios which have been used for building the corresponding four fraud detection models. The scenarios have been described in Chapter five.

CHAPTER FIVE
MODEL DESIGN

5.1 Introduction

The data preparation final result is a data that is suitable for building the fraud detection models. Understanding of the data itself and the business perspective of customs clearance especially, those assumptions and rules of the organization were inputs to have good understanding of the task to be conducted. As mentioned in the general objective (in Section 1.6.1), the goal of this study is to build fraud detection model from customs data in ERCA that can predicts fraud behaviors of importing cargo and classify fraud risks into different levels. In order to build the required model, classification methods and algorithms incorporated in WEKA environment were used. WEKA is a widely used tool for machine learning process implemented using Java and originally developed at Waikato University in New Zealand [23]. It is a software which collects various machine learning algorithms, and includes tools for data preparation, classification, regression, clustering, association rules, and visualization [46]. In order to build the best classifier for the problem in our hand, different classification methods and algorithms were compared and evaluated. Consequently, the better techniques have been selected based on the resulted performance. So in this research, selection of best model has been made using accuracy as well as using other common evaluation metrics such as precision, recall, ROC, confusion matrix and time taken to construct the model [22].

In the design phase four relevant candidate scenarios were identified to best fit the organizational requirement. These scenarios were used to build in four different models. The four scenarios are presented in the following subsections.

5.1.1 First Scenario (Fraud prediction)

The first scenario is used to build a fraud detection model which has been built based on the whole data set and attempts to predict an importing cargo whether it has fraud or not. This model has two classes: a declaration that has at least one item which is fraudulent (i.e. the YES class) and a declaration that is non-fraud (i.e. the NO class). This model would predict a new importing cargo as "fraud" or "Non-fraud", represented by the class label values "YES" and "No" classes

respectively. This model is named as "*Fraud prediction Model (FPM)*"in the subsequent sections and chapters.

5.1.2 Second Scenario (Fraud category prediction)

The second scenario is used to build a prediction model which would predict fraud categories of importing cargos from the data which is identified as fraudulent. The fraud categories are UV (Undervaluation), MC (Mis-Classification), MD (Mis-Description), EG (Extra Goods) and OD (Origin Difference). The model has been built using the fraud dataset of the target dataset and should predict one or more of these fraud categories that associated with a cargo. A declaration which is labeled as fraudulent would have one or more of the above fraud categories.

For experimental setup of this scenario, codes of fraud categories are represented by binary digits (as stated in Table 5.1 fifth column) and converted into decimal digits (as shown on the sixth column of Table 5.1) which will be interpreted into fraud categories. The paradigm behind this idea is, there are five types of fraud and according to our study dataset, at most three types of fraud can be detected with in a cargo; so that, five binary digits are used to represent the fraud codes as presented in Table 5.1. If specific fraud is detected, the digit which is represented by that digit will be 1 or 0 if not.

Table 5.1: Description of fraud category

No.	Number of fraud types	Fraud codes	Number of records	Fraud codes in Binary	Aggregated code
1	1	OD	532	00001	1
2	1	EG	1927	00010	2
3	2	EG, OD	194	00011	3
4	1	MD	1568	00100	4
5	2	MD, OD	318	00101	5
6	2	MD, EG	781	00110	6
7	3	MD, EG, OD	522	00111	7
8	1	MC	2485	01000	8
9	1	UV	26603	10000	16
10	2	UV, MC	479	11000	24

The fraud category distribution of the dataset which has been used to construct the *fraud category prediction model* is also illustrated in Table 5.1 fourth column. Based on the observed data for this study, only valid combinations of fraud category/s per declaration are used to

identify possible class label values. They are ten in number and other possible combinations are not valid.

The table also illustrates how the class label attribute is constructed and how it will be interpreted. The data set for this scenario has ten valid and distinct class values (1, 2, 3, 4, 5, 6, 7, 8, 16 and 24). This model is named as "*Fraud Category Prediction Model (FCPM)*" in the remaining part of this thesis.

5.1.3 Third Scenario (fraud level prediction)

The third scenario is used to build a prediction model which would predict the cargos' fraud level (as high or low) that are classified as "fraud" in the first model. This model has been built based on the fraud dataset of target dataset and creates an opportunity to inform that whether the importer must be penalized a duplicate amount of the infraction or the infraction without penalty. This model is named as "*Fraud level Prediction Model (FLPM)*" starting from this part of the thesis.

5.1.4 Fourth Scenario (Fraud risk level prediction)

The last scenario is used to build a prediction model which would classify or predict new importing cargos in a degree of severity: into high (critical), medium or low risk of fraud. The model has been built using the whole dataset which has three class values: HIGH, MEDIUM and LOW. The high class value represents those which have difference. The medium class value represents amendment. Those cargos which have not any fraud are represented by low class value. The term difference and amendment have been discussed in Section 3.3.3. This prediction model can be named as "*Fraud Risk Level Prediction Model (FRLPM)*".

5.2 Architecture for fraud detection model

The general architecture of the fraud detection models looks like as depicted in Figure 5.1.

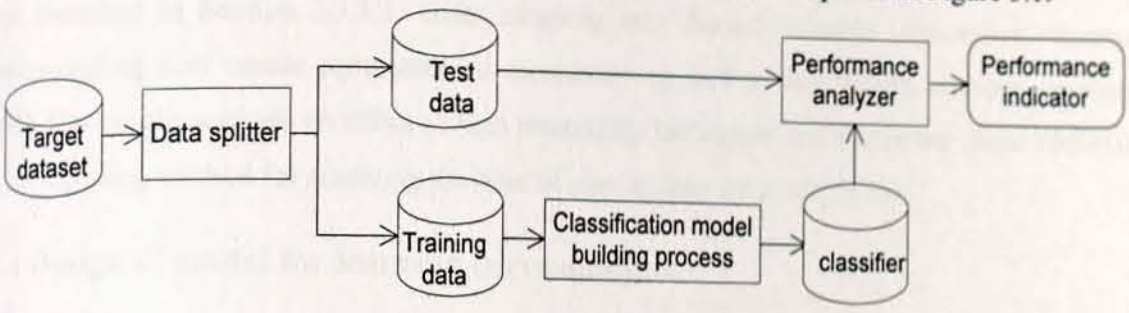


Figure 5.1: Architecture of fraud detection model building process

As mentioned in chapter four, the collected dataset has been prepared to obtain the target dataset. The target dataset should be prepared to be suitable for building each model which is described in the four different scenarios as stated in Section 5.1. As depicted in Figure 5.1, the target dataset is an input to model building and evaluation process. The data splitter divides the target dataset into two datasets: training dataset and test dataset. The training dataset is used to build the classifier model. Machine learning algorithms are used in the model construction process. The performance analyzer evaluates the performance of the built classifier by using the test dataset. Finally, the performance indicator will display the performance achieved.

5.3 Design to handle unbalanced data

One of the challenges in fraud detection application is the issue of class imbalance in which the fraud instances are highly under represented as compared with the normal class. In such cases the accuracy of the model cannot represent the real problem since it biases towards the normal (non-fraud) side and predicts the large number of fraud instances towards non-fraud [1].

Basically, the primary objective of fraud detection systems is detecting fraud; it must have a great attention. Therefore, in this research the issue of class imbalance has been considered as one of the basic aspect in design phase.

As we have seen in Figure 4.1, which represents the class distribution of the study dataset, if we consider the data distribution of the third scenarios; the minority (high fraud level) class instances are heavily under represented as compared with the majority (low fraud level) class

instances. This may create misclassification of minority (high fraud level) class into majority (low fraud level) class. Therefore possible balancing technique is required.

As described in Section 2.3.3.1, under sampling may discard valuable information whereas oversampling may create computational complexity as well as insufficient memory problem [36]. Having the analysis on effect of both resampling techniques, the researcher chose SMOTE over sampling method for resolving the issue of class imbalance in our dataset.

5.4 Design of model for learning curve analysis

In order to select the best testing mode for an algorithm, experiments have been done for each (i.e. the selected four) machine learning algorithms. The experiments have been done by increasing the size of the data by five percent, starting from five percent of the total data set till it reaches hundred percent of the total data set. Based on the result, the learning curves sketched and the best testing mode were determined. Learning curve is used to show whether the data set is sufficient or not.

Hence, if the performance's graph of the learning curve is constant (oscillate) while the data size is increased, the data is said to be sufficient. Whereas, if the performance is increasing while the data size is increased till it reaches 100 percent of the total dataset, the data is said to be insufficient.

K-fold cross validation is recommended for insufficient data because, in this testing mode the final reported result is the averaged of K experiments. Either of k-fold cross validation or percentage split testing option can be used if the data is sufficient. In our study, this testing mode selection approach has been applied.

CHAPTER SIX

EXPERIMENTATION AND EXPERIMENTAL ANALYSIS

6.1 Introduction

In this part of the research, four different models (fraud prediction model, fraud category prediction model, fraud level prediction model and fraud risk level prediction model) have been built based on the comparative study on the selected four machine learning algorithms (C4.5, CART, KNN and Naïve Bayes) and on the corresponding four scenarios that have been stated in Section 5.1. The models have been built using the recommended tool (WEKA) which implements all the above algorithms [23]. Systematic design of experimental process was followed to minimize the total number of experiments that might be required which in turn reduce to total model construction time and manageable experimental results for proper analysis.

During experimentation, first, the researchers conduct learning curves analysis for each machine learning algorithm to identify the appropriate testing mode for each of the four recommended machine learning algorithms. Once the testing mode is selected, the researchers identified the best algorithm among the four machine learning algorithms for each scenario. The selection was made based on the identified testing mode and the default parameters that each algorithm defines in the selected tool, which is WEKA. Finally, experiments were conducted using the selected machine learning algorithms for all the identified four scenarios and the possible parameter that can be tuned to improve the performance of the respective machine learning algorithm and scenarios.

6.2 Learning Curve analysis

In this research, the selection of testing mode has been performed based on the performance showed on the learning curves. The learning curves are constructed based on the experiments which were done by increasing the data size (as discussed in Section 5.5) and observing whether the model acquire knowledge from new instances or not. It is constructed using default parameters of the machine learning algorithms, the whole dataset for this study and the fraud non-fraud prediction scenario.

Figure 6.1, 6.2, 6.3 and 6.4 show the learning curves of NaiveBayes, C4.5, CART and KNN algorithms respectively. Figure 6.1(a), 6.2(a), 6.3(a) and 6.4(a) are the learning curves of NaiveBayes, C4.5, CART and KNN algorithms respectively using 10-fold cross validation whereas Figure 6.1(b), 6.2(b), 6.3(b) and 6.4(b) are the learning curves of NaiveBayes, C4.5, CART and KNN algorithms respectively using percentage split.

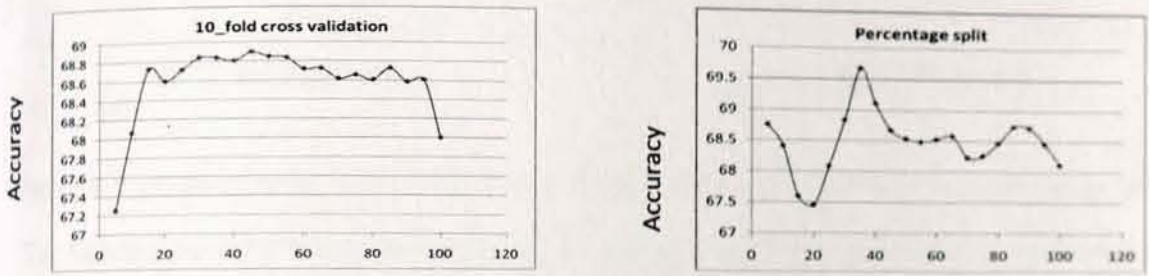


Figure 6. 1: Learning curves for NaiveBayes algorithm (a) using 10 fold cross validation (b) using percentage split

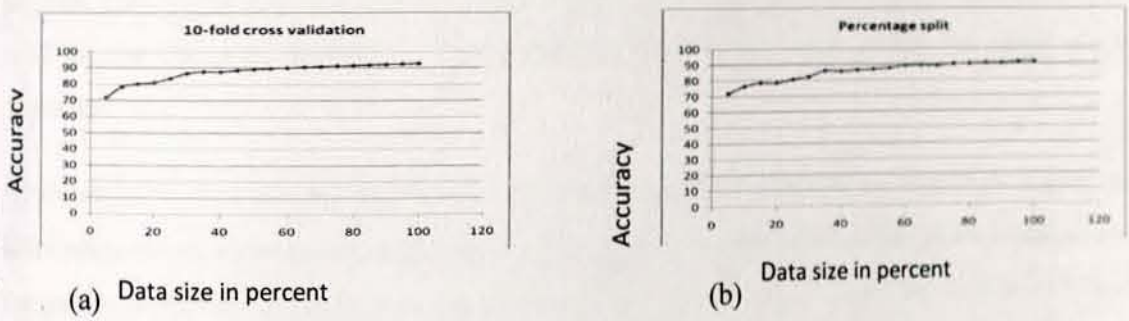


Figure 6. 2: Learning curves for C4.5 algorithm (a) using 10 fold cross validation (b) using percentage split

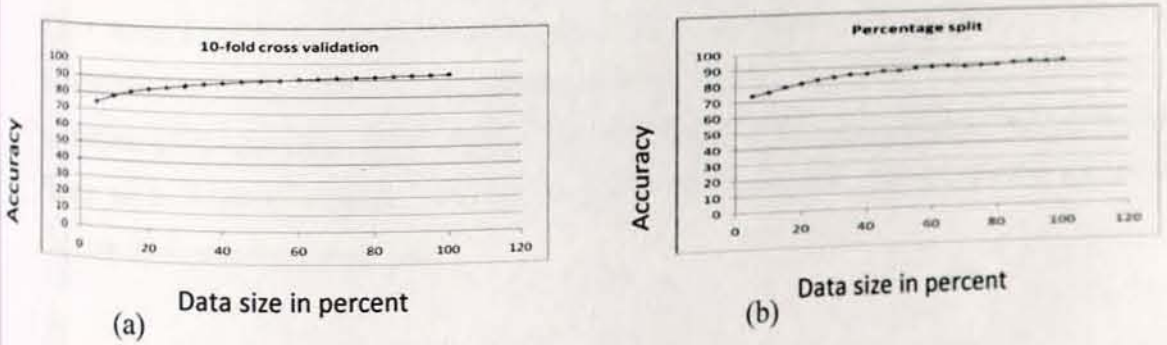


Figure 6. 3: Learning curves for CART algorithm (a) using10 fold cross validation (b) using percentage split

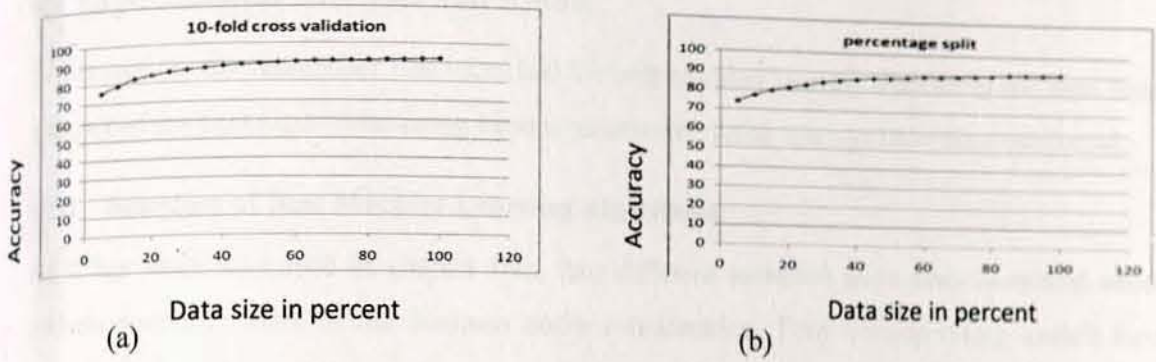


Figure 6. 4: Learning curves for KNN algorithm (a) using 10 fold cross validation (b) using percentage split

The learning curve of Naïve Bayes algorithm using percentage split shows that the performance is very poor as well as fluctuates while the data size is increasing. The fluctuation shown in Figure 6.1 indicates even if training data increases, performance might increase or decrease which is manifestation of poor capability of capturing knowledge for the intended classification purpose. This might indicate that the independence assumption of Naïve Bayes algorithm is not valid for our objective and the use of all attribute to determine class probability might not be desirable.

As shown in Figures 6.2, 6.3 and 6.4 (i.e. the learning curves for the algorithms C4.5, CART and KNN respectively) show, the accuracy increases while the data size is increased. At the end of the graphs, since the accuracy is increasing insignificantly, we can say that the data is sufficient. The same could be observed by looking the insignificant variation in performance between 10-fold cross validation and percentage split. This shows, one could use either of the approaches as a test mode for subsequent experiment using C4.5, CART, and KNN. However, 10-fold cross validation has been proved to be statistically good enough in evaluating the performance of classifiers [20]. Therefore, the researcher selected 10-fold cross validation as the appropriate test mode as it is more reliable in all possible scenario even if it took more time for model construction than percentage split.

6.3 Experimental Analysis and Result

In this section, the researcher first identified the best machine learning algorithms and then detail analysis of the best algorithms using various parameters tuning strategy have been conducted.

6.3.1 Selection of Best Machine Learning Algorithms

As it has been discussed in chapter five, four different scenarios have been identified which address specific issues of the business under consideration. Four corresponding models have been built based on the comparative study which was done in the four machine learning algorithm. These are:

- *Fraud prediction model (Scenario one)*
- *Fraud category prediction model (Scenario two)*
- *Fraud level prediction model (Scenario three)*
- *Fraud risk level prediction model (Scenario four)*

The first and the third scenarios are a two class problem whereas the second and the fourth scenarios have 10 and 3 distinct classes respectively as discussed on Chapter five. In this section, the selected test mode (10-fold cross validation) is applied to identify the best machine learning algorithm to each of the scenarios. Table 6.1 shows the performance of each machine learning algorithm for each scenario.

Table 6. 1: Comparison of four machine learning algorithms for the four scenarios

Algorithm	Performance (Accuracy in percent (%))			
	Fraud prediction model	**Fraud category prediction model**	**Fraud level prediction model**	**Fraud risk level prediction model**
C4.5	**92.8127**	**84.4333**	**89.1043**	**86.7451**
CART	**92.9072**	**80.0853**	**89.3834**	**85.1323**
KNN	92.0927	78.3549	87.1658	82.6672
NaiveBayes	68.3011	70.6347	69.6651	59.1951

As shown above in Table 6.1, C4.5 and CART are found to perform much better than KNN and Naïve Bayes methods. Hence, the researcher have selected these two machine learning algorithms to do further parameter tuning purpose. Moreover, as shown in the table, CART and C4.5 shows encouraging performance that would further improved via parameter tuning.

6.3.2 Analysis of Parameter Tuning on CART and C4.5

In this section, the researcher have tried to identify the most appropriate parameters that can be tuned to improve the performance of CART and C4.5 algorithms on the given four different scenarios. As a result, the researcher identified two parameters appropriate for CART machine learning algorithms. These are the number of folds in the internal cross validation (numFoldsPruning) and the minimum number of observation at the terminal node (MinNumObj). Table 6.2 shows the possible combinations of the parameters' values taken in CART analysis. It shows nine different experiments have been conducted to each scenario.

Table 6. 2: Tuned parameters for CART algorithms

Experiment #	Parameter tuned	
	numFoldsPruning	MinNumObj
1	5	2
2	5	5
3	5	10
4	7	2
5	7	5
6	7	10
7	10	2
8	10	5
9	10	10

Similarly, the researcher identified two parameters for C4.5 machine learning algorithm to improve the performance of the algorithm. These are the confidence factor (confidenceFactor), which is used for pruning the decision tree, and the minimum number of objects at the leaf node (MinNumObj). Table 6.3 shows the possible combinations of the parameters' value used in C4.5 for analysis purpose. The table shows, nine different experiments will be conducted to each scenario.

Table 6. 3: Tuned parameters for C4.5 algorithms

Experiment	Parameter tuned	
	confidenceFactor	MinNumObj
1	0.25	2
2	0.25	5
3	0.25	10
4	0.35	2
5	0.35	5
6	0.35	10
7	0.50	2
8	0.50	5
9	0.50	10

Nnine experiments were conducted on each two selected machine learning algorithms (C4.5 and CART) of the four different scenarios. The results of the experiments of each of the scenarios have been discussed below.

A) Parameter tuning for fraud prediction model (Scenario one)

Table 6.4 and Table 6.5 show the result of the parameter tuning on the first scenario using CART and C4.5 machine learning algorithms respectively.

Table 6. 4: Experiments of CART for fraud prediction

Experiment	#1	#2	#3	#4	#5	#6	#7	#8	#9
Accuracy (%)	92.9072	91.3025	89.0198	92.9032	91.3039	89.036	**92.9464**	91.3052	88.9711
Mean absolute error	0.0846	0.1132	0.1493	0.0846	0.1131	0.1497	**0.0838**	0.1139	0.149
Time taken to build model (sec)	6165.11	2549.34	2485	17114.2	3619.3	3498.2	**12217.82**	5152.97	5046.83
Number of Leaf	1974	1251	671	2338	1260	622	**2005**	1236	666
Size of the Tree	3947	2501	1341	4675	2519	1243	**4009**	2471	1331
Avg. TP Rate	0.929	0.913	0.89	0.929	0.913	0.89	**0.929**	0.913	0.89
Avg. FP Rate	0.072	0.089	0.112	0.072	0.089	0.112	**0.072**	0.089	0.113
Avg. Precision	0.929	0.914	0.891	0.929	0.914	0.891	**0.93**	0.914	0.89
Avg. Recall	0.929	0.913	0.89	0.929	0.913	0.89	**0.929**	0.913	0.89
Avg. F-Measure	0.929	0.913	0.89	0.929	0.913	0.89	**0.929**	0.913	0.89
Avg.ROC	0.963	0.961	0.947	0.964	0.961	0.947	**0.963**	0.961	0.947

According to the results of experiments in Table 6.4, experiment #7 with tuned parameter (numFoldsPruning=10 and MinNumObj=2) is obtained better performance (accuracy) as compared with the other eight experiments performed in CART algorithm for the first scenario.

Moreover, a closer look into Table 6.4 shows the following:

- The smaller the minimum number of objects the better the performance as can be seen in experiment #7, #4, #1
- The smaller the minimum number of objects the higher the construction time it takes

Table 6. 5: Experiments of C4.5 for fraud prediction

Experiment	#1	#2	#3	#4	#5	#6	#7	#8	#9
Accuracy (%)	92.813	90.737	87.262	92.994	90.753	87.75	**93.3746**	90.953	87.8284
Mean absolute error	0.1016	0.1268	0.1772	0.0984	0.8141	0.163	**0.0898**	0.1198	0.161
Time taken to build model(sec)	7.22	5.85	4.81	6.41	5.6	4.96	**6.52**	5.76	5.1
Number of Leaves	33768	25585	16598	34638	26345	19072	**39403**	27175	19765
Size of the tree	35227	26549	17090	36151	27370	19665	**41094**	28230	20374
Avg. TP Rate	0.928	0.907	0.873	0.93	0.908	0.877	**0.934**	0.91	0.878
Avg. FP Rate	0.075	0.096	0.133	0.073	0.096	0.126	**0.069**	0.094	0.125
Avg. Precision	0.93	0.909	0.877	0.932	0.909	0.879	**0.935**	0.911	0.88
Avg. Recall	0.928	0.907	0.873	0.93	0.908	0.877	**0.934**	0.91	0.878
Avg. F-Measure	0.928	0.907	0.872	0.93	0.907	0.877	**0.934**	0.909	0.878
Avg.ROC	0.966	0.96	0.934	0.967	0.961	0.946	**0.968**	0.962	0.947

In Table 6.5, comparison among the results of experiments which have been done on different tuned parameters of C4.5 algorithm for fraud prediction model, shows the accuracy obtained in experiment #7 (i.e. MinNumObj=2, confidence factors=.50) is better than that of the other experiments performed in C4.5 algorithm for the first scenario (fraud prediction).

Moreover, the closer observation into Table 6.5 indicates the following:

- The smaller the minimum number of object the better the performance as shown in experiment #7, #4, #1

- The smaller the minimum number of objects the higher the construction time it takes

After experiments were done by tuning parameters, the best tuned parameters which maximize the performance of the algorithm in the two algorithms (i.e. C4.5 and CART) have been selected and analyzed. Accordingly, the best algorithm which has been used to build the fraud prediction model has been chosen through comparison of these algorithms' accuracy as well as the other performance evaluation metrics such as precision, recall, ROC and time taken to build the model, by considering the best tuned parameters.

Table 6.6 shows the comparison of C4.5 and CART algorithms using the best tuned parameters. As a result, C4.5 has been selected as a best algorithm for building *Fraud Prediction Model*, achieving the highest result (i.e. 93.4% accuracy), as compared with CART machine learning algorithm.

Table 6. 6: Comparison of CART and C4.5 algorithms for fraud prediction

Algorithm	Accuracy (%)	Time taken to build model(sec)	Precision	Recall	Avg.ROC
C4.5	**93.3746**	**7.16**	**0.935**	**0.934**	**0.968**
CART	92.9464	12217.82	0.93	0.929	0.963

The confusion matrix result of experiment #7 of C4.5 (for fraud prediction scenario) as indicated in WEKA is presented in Table 6.7.

Table 6. 7: Confusion matrix of C4.5 experiment #7 for fraud prediction

		Predicted output		
		Yes	**No**	Total
Actual output	**Yes**	**31852**	3557	35409
	No	1348	**37276**	38624
	Total	33200	40833	**74033**

As shown in Table 6.7, the confusion matrix describes the number of instances which are classified correctly and incorrectly. Out of the total number of instances (74033) that were used for testing, 31,852 fraudulent instances are correctly classified as fraudulent, whereas 1,348 non-fraudulent records are misclassified as fraudulent. In the second row of the confusion matrix, 3,557 fraudulent records are incorrectly classified as non-fraudulent, whereas 37,276 instances are correctly classified as non-fraudulent. Since our study focuses on detecting fraud, model which has lesser false negative and higher true positive value in confusion matrix should be selected. Hence, according to our experimental results, C4.5 algorithm with parameters tuned in experiment #7 is chosen as best result by performing better accuracy, high true positive and less false negative result as compared with the other experiments performed in this scenario (fraud prediction). False negative instances are cargos with fraud which are classified as non-fraud. If this value is minimized the number of fraud instances which will be predicted as non-fraud by the model will be minimized. Accordingly, the number of cargos with fraud which will leave customs zone without inspection will be minimized.

B) Parameter tuning for fraud category prediction model (Scenario two)

Table 6.8 and Table 6.9 show the result of the parameter tuning on the second scenario using CART and C4.5 algorithms respectively.

Table 6. 8: Experiments of CART for fraud category prediction

Experiment	#1	#2	#3	#4	#5	#6	**#7**	#8	#9
Accuracy (%)	80.0853	79.7802	79.1701	80.2689	79.8057	79.1108	**80.3282**	79.817	79.0966
Mean absolute error	0.0579	0.0609	0.0642	0.0552	0.06	0.0636	**0.0546**	0.0592	0.0638
Time taken to build model (sec)	42315.97	24838.83	42153	42659.12	45849.14	43138	**62983.17**	80833.3	49852.6
Number of Leaf	355	255	155	337	275	151	**387**	282	151
Size of the Tree	709	509	309	673	549	301	**773**	563	301
Avg. TP Rate	0.801	0.798	0.792	0.803	0.798	0.791	**0.803**	0.798	0.791
Avg. FP Rate	0.486	0.516	0.556	0.463	0.507	0.55	**0.456**	0.499	0.551
Avg. Precision	0.77	0.767	0.76	0.773	0.765	0.756	**0.774**	0.765	0.756
Avg. Recall	0.801	0.798	0.792	0.803	0.798	0.791	**0.803**	0.798	0.791
Avg. F-Measure	0.765	0.756	0.741	0.772	0.758	0.742	**0.774**	0.76	0.742
Avg.ROC	0.8	0.78	0.765	0.812	0.789	0.771	**0.816**	0.794	0.77

According to the results of experiments in Table 6.8, Experiment #7 with tuned parameter (numFoldsPruning=10 and MinNumObj=2) is obtained with better performance (accuracy) as compared with the other eight experiments performed in CART algorithm for the second scenario.

Moreover, a closer look into Table 6.8 shows the following:

- The smaller the minimum number of objects the better the performance as can be seen in experiment #7, #4, #1
- The higher the number of fold the higher the construction time it takes as can be seen in experiment #7, #8, #9

Table 6. 9: Experiments of C4.5 for fraud category prediction

Experiment	#1	#2	#3	#4	#5	#6	#7	#8	#9
Accuracy (%)	**84.4333**	83.298	81.7871	84.1961	83.1427	81.5753	84.1396	82.8518	81.259
Mean absolute error	**0.0366**	0.0405	0.0471	0.0349	0.0387	0.0458	0.0332	0.0369	0.0431
Time taken to build model (sec)	**1.53**	0.56	0.39	0.95	0.53	0.36	0.91	0.52	0.37
Number of Leaves	**14828**	9883	6254	15827	10352	6702	22738	11993	7745
Size of the tree	**15268**	10197	6417	16299	10681	6882	23260	12360	7952
Avg. TP Rate	**0.844**	0.833	0.818	0.842	0.831	0.816	0.841	0.829	0.813
Avg. FP Rate	**0.324**	0.354	0.409	0.306	0.336	0.395	0.288	0.32	0.368
Avg. Precision	**0.829**	0.814	0.793	0.829	0.814	0.792	0.83	0.814	0.791
Avg. Recall	**0.844**	0.833	0.818	0.842	0.831	0.816	0.841	0.829	0.813
Avg. F-Measure	**0.831**	0.817	0.794	0.832	0.818	0.795	0.834	0.818	0.797
Avg.ROC	**0.928**	0.916	0.892	0.934	0.923	0.899	0.938	0.93	0.911

In Table 6.9, comparison among the results of experiments which have been done on the selected tuned parameters of C4.5 algorithm for fraud category prediction, shows the accuracy achieved in experiment #1 (i.e. MinNumObj=2, confidence factors=.25) is better than that of the other experiments performed in C4.5 algorithm for the second scenario (fraud category prediction).

Moreover, a closer observation into Table 6.9 shows the following

- The smaller the minimum number of objects the better the performance as seen in experiment #1,#4,#7
- The smaller the minimum number of objects the higher the construction time it takes as can be seen in experiment #1,#4,#7

After experiments were done by tuning parameters, the best tuned parameters which maximize the performance of the algorithm in each algorithm (i.e. in C4.5 and CART) are selected and analyzed. Accordingly, the best algorithm has been chosen through comparison of these algorithms' accuracy by considering the best tuned parameters. ROC, Precision, recall and time taken to build the model also considered in the comparison. Table 6.10 shows the comparison of C4.5 and CART algorithms using the best tuned parameters. As a result, C4.5 is selected as a best algorithm for building *Fraud Category Prediction Model* by achieving the highest performance (i.e. 84.43% accuracy), as compared with the performance obtained on CART.

Table 6. 10: Comparison of CART and C4.5 algorithms for fraud category prediction

Algorithm	Accuracy (%)	Time taken to build model(sec)	Precision	Recall	Avg.ROC
C4.5	**84.4333**	**1.53**	**0.829**	**0.844**	**0.928**
CART	80.3282	62983.17	0.774	0.803	0.816

The confusion matrix result of experiment #1 of Table 6.9 (which performed better as compared with the other experiments performed for the second scenario) as indicated in WEKA is presented in Table 6.11.

According to our experiments, C4.5 algorithm (which achieves 84.43% accuracy) is the best algorithm to build fraud category prediction model, followed by CART algorithm.

As we have described on Chapter five, this model is built to identify the type/s of fraud associated with certain cargo which is identified as fraud by *fraud prediction model*. According to the pattern of fraud type observed on the past data, 10 different type of fraud category have

been identified as stated in Table 5.1. Consequently, the results shown in Table 6.11 can be interpreted as follows.

As shown in the confusion matrix presented in Table 6.11, 269 instances are correctly classified as they are associated with origin difference fraud category; whereas 88 undervalued instances and 1 instance which is associated with undervaluation and misclassification fraud are misclassified as they have origin difference fraud. In the second column, 696 instances are correctly classified as Extra Goods category of fraud, whereas, 10,1,56, 374 and 17 instances in the column are incorrectly classified as Extra goods however their actual classes are Mis-Description of goods (i.e. class 4), they have both Mis-Description, extra goods and origin difference (i.e. class 7), Misclassification of goods (class 8), Undervaluation of goods (class 16) and both undervaluation and misclassification of goods (class 24) respectively. As in Table 6.11, the diagonal elements of the table shows the number of instances which are correctly classified by the classifier, whereas the numbers above and below the diagonal elements are interpreted by taking the row heading as actual class and the column heading as predicted class.

Table 6. 11: Confusion matrix of C4.5 experiment #1 for Fraud Category Prediction

		Predicted output										
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**16**	**24**	Total
Actual output	**1**	**269**	0	0	0	0	0	0	1	262	0	532
	2	0	**696**	1	11	0	1	1	64	1140	13	1927
	3	0	0	**44**	0	0	0	0	0	150	0	194
	4	0	10	0	**634**	0	0	3	22	883	16	1568
	5	0	0	0	0	**65**	0	0	5	248	0	318
	6	0	0	0	0	0	**334**	0	13	434	0	781
	7	0	1	0	1	0	0	**238**	28	254	0	522
	8	0	56	0	34	7	15	44	**2006**	323	0	2485
	16	88	374	46	280	118	208	144	65	**25272**	8	26603
	24	1	17	0	37	0	0	0	0	85	**339**	479
	Total	358	1154	91	997	190	558	430	2204	29051	376	35409

C) Parameter tuning for Fraud level prediction (Scenario three)

Scenario three has been used to build a classifier which predicts whether the declaration of a cargo associated with certain fraud needs amendment or not (it deals about fraud level prediction into high or low). The dataset which is used to construct this model has highly imbalanced class. Class imbalance degrades the predictive capability of a model by misclassifying minority class instances as majority class [37]. To come up with this problem, we have resampled the data set using SMOTE analysis. Therefore, the experimental analyses of *Fraud level prediction Model* construction are based on oversampling the minority class by 400%.

Table 6.12 and Table 6.13 show the result of the parameter tuning on the third scenario using CART and C4.5 machine learning algorithm respectively.

Table 6. 12: Experiments of CART for fraud level prediction

Experiment	#1	#2	#3	#4	#5	#6	#7	#8	#9
Accuracy (%)	89.3834	88.849	87.8891	89.4072	88.9341	87.964	**89.4089**	88.9205	87.9248
Mean absolute error	0.1356	0.1512	0.1706	0.1334	0.1483	0.1693	**0.1335**	0.1482	0.1675
Time taken to build model(sec)	1323.58	1474.8	1263.35	1838.56	1804.89	2126.49	**2546.39**	2553.36	2593.53
Number of Leaves	874	594	341	729	639	373	**862**	598	395
Size of the tree	1747	1187	681	1457	1277	745	**1723**	1195	789
Avg. TP Rate	0.894	0.888	0.879	0.894	0.889	0.88	**0.894**	0.889	0.879
Avg. FP Rate	0.106	0.111	0.121	0.106	0.111	0.12	**0.106**	0.111	0.121
Avg. Precision	0.894	0.889	0.879	0.894	0.89	0.88	**0.894**	0.889	0.88
Avg. Recall	0.894	0.888	0.879	0.894	0.889	0.88	**0.894**	0.889	0.879
Avg. F-Measure	0.894	0.888	0.879	0.894	0.889	0.88	**0.894**	0.889	0.879
Avg.ROC	0.957	0.95	0.942	0.957	0.952	0.943	**0.957**	0.953	0.944

According to the results of experiments in Table 6.12, Experiment #7 with tuned parameter (numFoldsPruning=10 and MinNumObj=2) is obtained with better performance (i.e. accuracy) as compared with the other eight experiments performed in CART algorithm for the third scenario (fraud level prediction).

Moreover, a closer observation into Table 6.12 shows the following

- The smaller the minimum number of objects the better the performance as seen in experiment 7,#4,#1
- The higher the number of fold the higher the construction time it takes as can be seen in experiment #7, #8, #9

Table 6. 13: Experiments of C4.5 for fraud level prediction

Experiment	#1	#2	#3	#4	#5	#6	#7	#8	#9
Accuracy (%)	89.1043	88.4235	87.5691	89.1179	88.4609	87.644	**89.4004**	88.8081	87.964
Mean absolute error	0.1417	0.1529	0.1671	0.7824	0.1469	0.1633	**0.1262**	0.1376	0.1541
Time taken to build model(sec)	5.12	4.13	3.53	4.59	4.13	3.71	**4.54**	3.92	3.73
Number of Leaves	14617	11982	9834	15564	12783	10148	**16878**	14099	11731
Size of the tree	15357	12549	10265	16358	13397	10610	**17730**	14754	12220
Avg. TP Rate	0.891	0.884	0.876	0.891	0.885	0.876	**0.894**	0.888	0.88
Avg. FP Rate	0.109	0.116	0.124	0.109	0.115	0.124	**0.106**	0.112	0.12
Avg. Precision	0.891	0.884	0.876	0.891	0.885	0.876	**0.894**	0.888	0.88
Avg. Recall	0.891	0.884	0.876	0.891	0.885	0.876	**0.894**	0.888	0.88
Avg.ROC	0.953	0.95	0.944	0.955	0.952	0.946	**0.959**	0.956	0.951

In Table 6.13 , comparison among the results of experiments which have done on different tuned parameters of C4.5 algorithm for fraud level prediction model, shows the accuracy achieved on experiment #7 (i.e. MinNumObj=2, confidence factors=.50) is better than that of the other experiments performed in C4.5 algorithm for the third scenario.

More over a closer look into Table 6.13 shows the following

- The smaller the minimum number of object the better the performance as seen in experiment #1, #4, #7

- The smaller the minimum number of object the higher the construction time it takes as can be seen in experiment #1, #4, #7

After experiments (by tuning parameter) were done on the two better algorithms for *Fraud level prediction Model*, we have identified the results of the experiments of tuned parameters in which the performance of these algorithms is maximized (i.e. Experiment #7 of Table 6.12 for CART and experiment #7 of Table 6.13 for C4.5). Accordingly, the best algorithm has been chosen through comparison of them. In Table 6.14, the comparison of C4.5 and CART algorithms for construction of *Fraud level prediction Model,* the performance of the models built in both (C4.5 and CART) algorithms is approximately equal. Rather, C4.5 took minimum amount of time to build the model as compared with the time required to build the model using CART and also C4.5 shows better value of ROC than CART. So, we have chosen C4.5 as a best algorithm for building this model by obtaining 89.4% accuracy.

Table 6. 14: Comparison of CART and C4.5 algorithms for fraud level prediction

Algorithm	Accuracy (%)	Time taken to build model (sec)	Precision	Recall	Avg.ROC
C4.5	**89.4004**	**4.54**	**0.894**	**0.894**	**0.959**
CART	89.4089	2546.39	0.894	0.894	0.957

The confusion matrix result for experiment #7 of C4.5 (i.e. for fraud level prediction) as indicated in WEKA is presented in Table 6.15.

Table 6. 15: Confusion matrix of C4.5 on experiment #7 for fraud level prediction

		Predicted output		
		YES	NO	Total
Actual output	YES	**26214**	2971	29185
	NO	3257	**26315**	29572
	Total	29471	29286	58757

In the confusion matrix on Table 6.15, out of 58,757 instances, 26,214 instances are correctly classified by the model as high level fraud, whereas 3,257 low level fraud instances are

incorrectly classified as high level fraud by the classifier. In the second column of the confusion matrix, 26,315 instances are correctly classified by the model to low level fraud, whereas 2,971 high level fraud instances are incorrectly classified as low level fraud.

Here, the lesser the value in misclassification of high level fraud into low level fraud will increase the revenue, due to providing an opportunity in detecting intolerable (high level) fraudulent activities. The researcher considered this during model selection of fraud level prediction model.

D) Parameter tuning for fraud risk level prediction Model (Scenario four)

The fourth scenario is slightly similar with the interpretation of risk levels in ASYCUDA selectivity method though we are using a scientific method (machine learning techniques) to predict the risk level of a cargo. Selectivity method has been discussed in Section 2.2.

Table 6.16 and Table 6.17 show the result of the parameter tuning on the fourth scenario using CART and C4.5 machine learning algorithm respectively.

Table 6. 16: Experiments of CART for fraud risk level prediction

Experiment	#1	#2	#3	#4	#5	#6	#7	#8	#9
Accuracy (%)	85.1323	83.375	80.66	85.1904	83.3547	80.66	**85.2579**	83.3358	80.6586
Mean absolute error	0.1256	0.1502	0.1797	0.1241	0.1498	0.1797	**0.1232**	0.1492	0.1794
Time taken to build model (sec)	5979.75	4872.58	4363.19	8688.84	7286.4	6424.25	**12271.17**	11159.47	9225.49
Number of Leaves	2159	1314	747	2159	1291	753	**2317**	1314	778
Size of the tree	4317	2627	1493	4317	2581	1505	**4633**	2627	1555
Avg. TP Rate	0.851	0.834	0.807	0.852	0.834	0.807	**0.853**	0.833	0.807
Avg. FP Rate	0.112	0.133	0.163	0.11	0.133	0.162	**0.11**	0.132	0.162
Avg. Precision	0.842	0.823	0.794	0.843	0.823	0.794	**0.844**	0.822	0.794
Avg. Recall	0.851	0.834	0.807	0.852	0.834	0.807	**0.853**	0.833	0.807
Avg. F-Measure	0.845	0.824	0.793	0.846	0.824	0.793	**0.846**	0.824	0.793
Avg.ROC	0.943	0.932	0.91	0.943	0.932	0.911	**0.943**	0.933	0.911

According to the results of experiments in Table 6.16, Experiment #7 with tuned parameter (numFoldsPruning=10 and MinNumObj=2) is obtained with better performance (accuracy) as compared with the other eight experiments performed in CART algorithm for the fourth scenario (risk level prediction).

Moreover, a closer look into Table 6.16 shows the following:

- The smaller the minimum number of objects the better the performance as can be seen in experiment #7, #4, #1
- The higher the number of fold the higher the construction time it takes as can be seen in experiment #7,#8,#9

Table 6. 17: Experiments of C4.5 for fraud risk level prediction

Experiment	#1	#2	#3	#4	#5	#6	#7	#8	#9
Accuracy (%)	86.7451	84.3664	81.0747	86.6965	84.3029	81.5474	**86.818**	84.3232	81.4569
Mean absolute error	0.1058	0.1259	0.1636	0.1027	0.1233	0.1506	**0.099**	0.1209	0.1489
Time taken to build model(sec)	13.64	10.47	5.63	8.38	5.87	4.93	**8.13**	6.26	6.07
Number of Leaves	36875	26882	17058	38332	28092	20371	**40487**	29192	21509
Size of the tree	38565	27955	17606	40103	29249	21042	**42365**	30405	22207
Avg. TP Rate	0.867	0.844	0.811	0.867	0.843	0.815	**0.868**	0.843	0.815
Avg. FP Rate	0.095	0.116	0.154	0.092	0.113	0.142	**0.089**	0.111	0.141
Avg. Precision	0.86	0.835	0.802	0.861	0.836	0.806	**0.863**	0.837	0.805
Avg. Recall	0.867	0.844	0.811	0.867	0.843	0.815	**0.868**	0.843	0.815
Avg. F-Measure	0.862	0.838	0.802	0.863	0.838	0.808	**0.865**	0.839	0.808
Avg.ROC	0.958	0.951	0.917	0.96	0.952	0.935	**0.961**	0.953	0.937

In Table 6.17, comparison among the results of experiments which have been done on different tuned parameters of C4.5 algorithm for fraud risk level prediction model, shows the accuracy achieved in experiment #7 (i.e. MinNumObj=2, ConfidenceFactors=.50) is better than that of the other experiments performed to improve the performance of C4.5 algorithm for the specified scenario.

Moreover, a closer look into Table 6.17 shows the following:

- The smaller the minimum number of objects the better the performance as can be seen in experiment #7, #4, #1

Experiments have been done through parameter tuning in order to maximize the performance of the algorithms which were selected as best algorithms to build *fraud risk level prediction model*. Consequently, the best algorithm for constructing this *model* has been selected by comparing the maximized result. Experiments #7 of Table 6.16 and Table 6.17 are the results of the best tuned parameter (obtained better accuracy) in CART and C4.5 algorithms respectively. Table 6.18 shows how C4.5 algorithm performs better predictive performance than CART. As a result, C4.5 algorithm has been chosen as best machine learning algorithm to build *fraud risk level prediction model*, by achieving 86.8% of accuracy.

Table 6.18: Comparison of CART and C4.5 algorithms for fraud risk level prediction

Algorithm	Accuracy (%)	Time taken to build model(sec)	Precision	Recall	Avg.ROC
C4.5	**86.818**	**8.13**	**0.863**	**0.868**	**0.961**
CART	85.2579	12271.17	0.844	0.853	0.943

The confusion matrix result of experiment # 7 of C4.5 for *Fraud risk Level Prediction* model is presented in Table 6.19.

Table 6. 19: Confusion matrix of C4.5 experiment #7 for fraud risk level prediction

		Predicted Outputs			
Actual output		HIGH	MEDIUM	LOW	Total
	HIGH	**2552**	2704	581	5837
	MEDIUM	2301	**24255**	3016	29572
	LOW	176	981	**37467**	38624
	Total	5029	27940	41046	74033

The confusion matrix in Table 6.19 shows, out of 74,033 testing instances 64,274 instances are correctly classified into high, medium and low classes whereas, the rest are misclassified

incorrectly. In particular, 2,552 instances are correctly classified as high risk level (difference or not-amendable fraud), whereas 2301 medium risk level and 176 low risk level instances are incorrectly classified as high risk level. 24255 and 37467 instances are correctly classified as medium risk level and low risk level respectively. Moreover, 2704 high risk level instances and 981 low risk level instances are incorrectly classified as medium risk level. 581 high risk level and 3016 medium risk level instances are incorrectly classified as low risk level (i.e. non-fraudulent goods).

After the detailed analysis is performed by comparing the algorithms and tuning the parameters' of C4.5 and CART machine learning algorithms, the best machine learning algorithm for constructing the four prediction models has been selected. Accordingly, C4.5 is the best machine learning algorithm to construct the four prediction models followed by CART and KNN algorithms. Moreover, C4.5 is a fast machine learning algorithm to classify instances with large data set. The performance of the prediction models obtained through the four different scenario and the two different machine learning algorithms are described in Table 6.20.

Table 6. 20: Comparison of algorithms C4.5 and CART for the four prediction models

	Performance of predictive accuracy			
Algorithm	FPM	FCPM	FLPM	FRLPM
C4.5	**93.4**	**84.4**	**89.4**	**86.8**
CART	92.9	80.1	89.4	85.2

As shown in Table 6.20, C4.5 outperforms CART in all the possible scenarios by considering the construction time and ROC result of fraud level prediction model in Table 6.14.

6.4 Discriminant attributes selected for the study

In ASYCUDA selectivity method, five variables are considered to be risk factors. These are Commodity code, country of origin, Customs Procedure Code (CPC), importer and declarant. However during our analysis, the researcher found that there are variables which are not identified as risk factors in ASYCUDA, during discriminant attribute selection process. The relevancy of the variables (attributes) is determined based on the concept contribution to decide

the class label (target attribute) (i.e. whether the cargo is fraudulent or not) on the given dataset. Independent variables which are used in this research can be listed in the level of relevancy in the sequence listed on Table 6.21. Variables which provide more concept in predicting the correct class written first and those provide lesser concept are listed latter (i.e. in descending order based on their relevancy) together with their gain ratio value of gain ratio feature evaluator. Table 6.21 shows the list of ranked attribute using gain ratio attribute selection measure.

Table 6. 21: Ranked attributes using gain ratio feature evaluator

Rank	Attribute	Gain Ratio Value
1	Extended procedure code	0.07032
2	Withholding rate	0.06214
3	National procedure code	0.05149
4	Value Added Tax (VAT) rate	0.04812
5	Total item	0.04513
6	Sure rate	0.03392
7	Terms of delivery	0.03156
8	Currency	0.02176
9	Country of consignment	0.02161
10	HS-code(commodity code)	0.02
11	EXC_RATE	0.0186
12	Country of origin	0.018
13	Customs border	0.01794
14	Declarant	0.01429
15	Package	0.0142
16	Year	0.01317
17	Duty rate	0.01235
18	Trade type	0.01002
19	Importer region	0.00819
20	Month of registration	0.00447

As can be seen from Table 6.21, the top ten ranked attributes are selected as basic attributes for this study as they have gain ratio of greater than or equal to 0.02. These ten basic attributes can be categorized into three as follows.

1. Attributes which are basic in both ASYCUDA and this study
2. Attributes basic in this study but not in ASYCUDA
3. Attributes basic in ASYCUDA but not identified as top ten attributes in this study

Attributes in first category are CPC (Extended procedure code, National procedure code) and tariff (Withholding rate, Value Added Tax (VAT) rate and sure rate). Attributes in the second category are total number of item in a declaration, terms of delivery, currency, country of consignment and HS-code (commodity code). Attributes in the last category are country of origin, declarant, importer and duty rate.

Here, the second category attributes which are basic in fraud prediction but they are not in ASYCUDA having been justified.

During analysis of this study, the researcher visualized that *"Total number" of items* is an attribute which have more information to classify a given cargo as fraudulent or not. These leads the number of item in a given cargo matters the possibility of fraudulent activities. *Terms of delivery* is the other important variables which must be considered on risk parameter selection in customs of ERCA. This variable might be exposed to risk since the attribute has an impact on the CIF value in which the duty and tax payable are calculated. For instance, if "*terms of delivery*" is Free On Board (FOB), fright and insurance are covered by seller. Therefore, tax and duties are calculated only from the items cost in which the item is purchased. Importers might use this hole to evade duties and tax payable by declaring incorrect terms of delivery. *"Country of consignment"* must be considered as one of risk metrics because customer fraudulent behaviors are dependent on the country in which the goods are purchased in addition to the country in which made. The trade policy of the country in which the trade is made, has a great impact for frauds like invoice falsification, in addition to the country in which the item is made. *"Currency"* also might have information that leads about the country in which the trade is made. The other important attribute which is not considered as basic attribute in the selectivity is HS-code (commodity code). The analysis in this study indicates that in addition to analyzing the tariff and value of the commodity, HS_ code by it self must be considered as risk element.

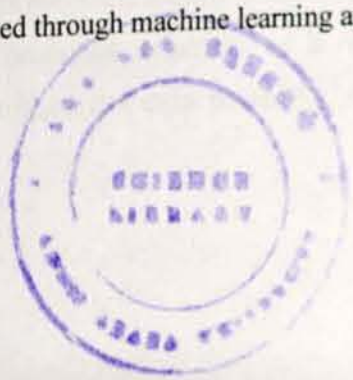
CHAPTER SEVEN
CONCLUSION AND RECOMMENDATION

7.1 Conclusion

In this research, various conflicting issues which face customs have been focused and assessed to improve the quality of service and minimize fraudulent activities. There are huge amounts of transactions but limited amount of resources to examine all these transactions. Genuine customers would like declaration clearance within short period of time but examination process might take too much time which is hardly acceptable. Custom office has the responsibility to control fraudulent declarations and take legal actions but significant number of fraudulent declarations might pass due to lack of sufficient knowledge about the fraudulent declarations.

Generally, controlling and facilitation are the most important issues in customs declaration processing and assessing risk levels. Controlling refers to fraud detection whereas facilitation deals with providing efficient services to the importers. Fraud detection is one part of the overall fraud controlling techniques which can be achieved through automation and helps to reduce the manual parts of screening/checking process. In ASYCUDA, these conflicting issues were proposed to be solved by risk level classification technique using selectivity method which uses five parameters from the details of the declarations. These parameters are: tariff/value of the declared commodities, the customs procedure code associated to the specific declaration, the importer identity, the declarant who is representing the importer for the declaration and origin of the commodity. The fundamental problem to ASYCUDA risk leveling is, restricting the specified five variables to assign risk level which may lead to direct the declaration into incorrect channel. As a result, those problems stated above become a long standing concern of the custom authority.

The goal of this research is, to build fraud detection models which provide solution for the above conflicting issues based on the customs data of ERCA and improves the quality of service in customs. Here, fraud detection refers to a technique which is used to control fraud with some classification/ prediction technique which is implemented through machine learning approach.



This study follows scientific research process that involves six phases from problem understanding to model analysis and evaluation. These phases are problem understanding, data collection and understanding, data preparation, model selection, model parameter estimation (model construction or training), and model analysis and evaluation. During problem understanding phase, details of the customs authority business process were reviewed and evaluated from the view of their core business objectives and goal that they would achieve. The existing system is analyzed and gaps were identified. As a result problems like lack of customer satisfaction, loss of government revenue due to fraudulent customer behavior were identified.

During data collection and understanding phase, the data collection task was carried out through integration of the local data of AAL with ASYCUDA data using two level data integration: data source level integration and record level integration based on the physically inspected cargos of 15 months records. The researcher analyzed and interpreted the collected data for its relevance to bridge the gap identified in problem understanding phase.

In the third phase, the collected data has been prepared to be suitable for model construction of the intended fraud detection purpose. In this phase, the researcher undertake a number of activities such as identification of relevant attribute and discarding attributes with is not relevant for the intended purpose, deriving new attribute which is intended to carry more information than the existing attributes, handling missing values and resolving inconsistency. Moreover, the data is structured into four scenarios each differs on their dependent variable selection. These scenarios are used as a base for building corresponding fraud detection models. The first scenario concerned with identifying importing cargo whether it is fraudulent or not (fraud prediction). The second scenario deals about identifying the fraud category associated with the declaration which is identified as fraudulent (fraud category prediction). The third scenario attempts to classify the fraud associated with cargo as high or low fraud level (fraud level prediction). Finally, the last scenario concerned with classifying the risk level of importing cargo as high, medium or low level (fraud risk level prediction).

During the model selection phase, The researcher adopt the IEEE recommendation to maximize the performance of the models and achieve better prediction capability as it is practically impossible to do the experiment in all prediction algorithms that exist nowadays. Accordingly,

C4.5, CART, KNN and Naïve Baye are selected which are recommendations of best classification models by IEEE.

In the fifth phase, model construction (model parameter estimation); the data prepared in phase three is split into training data set and testing data set. At this stage, the training data prepared for each scenario is given to the selected model. As a result, the model parameters were estimated and the required classifier for each scenario has been generated. For each model parameter estimation, the researcher attempt to improve the performance through parameter tuning.

During this phase, Based on the four scenarios generated in the third phase four corresponding models have been built. These are *fraud prediction model, fraud category prediction mode, fraud level prediction model* and *fraud risk level prediction model.*

The *fraud prediction model* should classify the importing cargos' declaration as fraud or non-fraud. The *fraud category prediction model* should predict the fraud category of the declaration which is identified as fraudulent by the first model. The *fraud level prediction model* should predict the cargo which is identified as fraudulent as high or low level of fraud. Lastly, *fraud risk level prediction model* predicts the cargos risk level as high, medium or low.

The last phase, which is analysis and evaluation, was carried out on the target test dataset on the constructed model. The result achieved in the four different scenarios showed that C4.5 is the best machine learning algorithm to build prediction models for all the four scenarios. The accuracy achieved for the first, second, third and fourth scenarios are **93.4%, 84.4%, 89.4% and 86.8%** respectively.

As a conclusion, the study puts forward that automatic fraud detection model can be built through machine learning techniques. According to our experimental analysis results, we can conclude that C4.5 is the best algorithm for

- Fraud prediction
- Fraud type identification
- Fraud level prediction

- Fraud risk level prediction

The researcher also observed that, CART and KNN are the next best classification model in all the scenarios next to C4.5.

Moreover, parameter tuning is a good technique to obtain better performance of a model in C4.5, whereas it is almost worthless in KNN and CART which is to mean that parameter tuning doesn't improve performance as compared to the default parameters. In C4.5 algorithm, the accuracy of the model increases when the values of the parameters "The confidence factor used for pruning" increased and" The minimum number of instances per leaf" decreased.

Generally, the researcher concluded the following based on the finding stated in this research. Firstly, it is shown that prediction models could address the problem of the custom Authority as the model shows excellent performance in predicting fraudulent activities and the type of fraud based on the initially provided declaration data as the models obtained ROC of above 0.9 . Moreover, the models give more flexibility to the authority as it can be trained incrementally that would consider situational changes without problem. As a result, it minimizes human intervention to the system.

7.2 Recommendation

The research identified the possibility of building fraud detection model from customs data in ERCA that can predicts fraud behavior of importing cargo and classify fraud risks into different levels. In this process, the research identified important recommendation for further investigation and to ERCA.

First of all, the researcher observed that ERCA assign fraud code after examining the declaration of a cargo. The fraud code is computed from the examination of all items in a declaration. The fraud might happen in one or more items of a declaration and fraud code will be assigned to the declaration. However, the code is assigned at the declaration level not at the specific item level. This is valid and sound to identify fraud types involved in the declaration. But according to the researcher, assigning the specific type of problem investigated at item level and do the analysis at item level will help ERCA to identify the type of items and the possible fraud that might arise at

item level. Having such information at item level in a declaration might provide better understanding about importer intension towards fraudulent behavior at item level.

Moreover, the researcher recommends ERCA to use the fraud detection model proposed and built to mitigate the drawbacks of ASYCUDA. The researcher already contacted the responsible personnel and they are willing to test the system after the system (i.e. user interface and functional requirements) gets completed and permission is given from AAU, Department of Computer Science.

Reference

[1] H. Shao, H. Zhao and G. chang. "Applying data mining to detect fraud behavior in customs declaration." in *Proceedings of the First International Conference on Machine learning and Cybernetics*, 2002, PP.1241-124.

[2] W. Yaqin and S. Yuming, "Classification Model Based on Association Rules in Customs Risk Management Application." in *International Conference on Intelligent System Design and Engineering Application,* 2010, pp.436-439.

[3] "Introduction about ERCA." Internet: http://www.erca.gov.et/index.jsp?id=aboutus, 2012 [Jun. 6, 2012]

[4] "A Proclamation to provide for the Establishment of the Ethiopian Revenues and Customs Authority: Proclamation No. 587/2008 or 587/2000 EC." Internet: http://www.2merkato.com/2008043034/customs-regulation-in-ethiopia [Oct. 5, 2011].

[5] "UNCTAD Trust Fund on Trade Facilitation Negotiations Technical Note 12." Internet: http://www.slideshare.net/Micheal22/risk-management-in-customs-procedures , Nov, 2008[Jul. 1, 2012].

[6] B. Laporte. "Risk management systems: using data mining in developing countries' customs administrations." *World Customs Journal*, Vol.5, no.1, PP. 17-28, 2011.

[7] Risk Management Department technical staffs, *Risk management policy and strategy document*, Risk Management Department of ERCA, 2011.

[8] R.L. Skinner. "Survey of DHS Data Mining Activities." Internet: http://www.oig.dhs.gov/assets/Mgmt/OIG_06-56_Aug06.pdf, Aug. 15, 2006 [Oct. 1, 2011].

[9] B. B aştabak and T. Medeni. "Application of informatics technologies into customs: Origin and tariff code diversion, impacts and identification problem." In *International journal of ebusiness and egovernment studies*, Vol. 3, no.1, PP.17-27, 2011.

[10] "ASYCUDA User Countries, Territories and Regions." http://www.asycuda.org/countrydb.asp. [Feb 02, 2012.]

[11] L. Yan-hai and S. lin-yan. "Study and Applications of Data Mining to the Structure Risk Analysis of Customs Declaration Cargo." in *Proceedings of the 2005 IEEE International Conference on e-Business Engineering (ICEBE'05)*, 2005, PP. 761 - 764.

[12] C. Phua1, V. Lee, K. Smith and R. Gayler. "A Comprehensive Survey of Data Mining-based Fraud Detection Research." http://arxiv.org/ftp/arxiv/papers/1009/1009.6119.pdf, Mar. 2007.

[13] N.T. Roman, R. Rezende, C.D. Ferreira, L.A. Digiampietri, L.A.A. Meira and J. Filho. "Attribute Value Specification in Customs Fraud Detection." in *The Proceedings of the 10th International Digital Government Research Conference*, 2009, PP. 264-271.

[14] F. Bonchi, F. Giannotti, G. Mainetto, D. Pedreschi, "A Classification-Based Methodology for Planning Audit Strategies in Fraud Detection", *ACM*, pp.175-184, 1999.

[15] A. Kumar and V. Nagadevara. "Development of Hybrid Classification Methodology for Mining Skewed Data Sets – A Case Study of Indian Customs Data." in *IEEE*, 2006, pp. 584-591.

[16] R. Jaime and L. Amorim. "KDD, SEMMA AND CRISP-DM: A PARALLEL OVERVIEW." in *IADIS European Conference Data Mining, 2008*, PP. 182-185.

[17] Habtom Gebregzaber. "Application of data mining technology in predicting the seroprevalence of HIV, HCV, HBV: in the case of the national blood bank service." Master's Thesis of Addis Ababa University, Addis Ababa, Ethiopia, 2011.

[18] P. Chapman et al "CRISP-DM 1.0 Step-by-step data mining guide." 2000.

[19] R.S.Michalski and K. Kaufman. "Learning patterns in noisy data: the AQ approach." *Springer-Verlag Berlin Heidelberg*, Vol. 2049, PP. 22–38, 2001.

[20] R.D. Lakshmi and N.Radha. "Machine Learning Approach for Taxation Analysis using Classification Techniques." *International Journal of Computer Applications (0975 – 8887)*, Vol. 12, no. 10, Jan. 2011, PP.1-6.

[21] X. Wu et al. "Top ten algorithms in data mining." *Springer-Verlag London Limited*, 2007, PP. 1-37.

[22] J. Davis and M. Goadrich. "The Relationship between Precision-Recall and ROC Curves." Appearing in Proceedings of the 23rd International Conference on Machine Learning, Pittsburgh, PA. 2006.

[23] A. H. Wahbeh, Q. A. Al-Radaideh, M. N. Al-Kabi and E. M. Al-Shawakfa. "A Comparison Study between Data Mining Tools over some Classification Methods." *International Journal of Advanced Computer Science and Applications, Special Issue on Artificial Intelligence*, Vol.2, PP. 18-26, 2011.

[24] "UNCTAD Technical Assistance in Trade Facilitation." Internet: http://www.asycuda.org/pdf%20docs/UNCTADTechnAssi.pdf [Jan.20, 2012].

[25] K. Mikuriya. "Risk management: a critical Customs tool." WCO news (Jun, 2010), Sec. 62, PP.24-25.

[26] H. Verrelst, E. Lerouge, Y. Moreau, J. Vandewalle, C. Störmann and P. Burge. "A rule based and neural network system for fraud detection in mobile communications." unpublished.

[27] J. Hollmén. "Probabilistic Approaches to Fraud Detection." Licentiate's Thesis, Helsinki University of Technology, Helsinki, Finland, 1999.

[28] M. Sternberg and R. Reynolds. "Using cultural algorithms to support re-engineering of the rule-based expert systems in dynamic performance environments: a case study in fraud detection." *IEEE Transactions on Evolutionary Computation*, Vol.1, no.4, 225–243, Nov. 1997.

[29] A. Lucia, J. Jambeiro, N. Trevisan, D. Cristiano, A. Luis and A. Andreia. "Uses of Artificial Intelligence in the Brazilian Customs Fraud Detection System." in *The Proceedings of the 9th Annual International Digital Government Research Conference*, 2008, pp.181-187.

[30] I. H. Witten, E. Frank and M. A. Hall. *Data Mining Practical Machine Learning Tools and Techniques Third Edition*. USA: Morgan Kaufmann Publishers, 2011, PP.3-215.

[31] J. Han and M. Kamber. *Data Mining: concepts and Techniques.* San Fransisco: Morgan kufman Publishers, 2001, PP.105-325.

[32] M. A. Hall. " Correlation-based Feature Selection for Machine Learning". PhD thesis, University of Waikato, NewZealand, 1999.

[33]L. Rokach and O. Maimon. *Data mining with decision trees theory and applications*. Singapore: World Scientific Publishing Co. Pte. Ltd, 2007, PP. 71-73.

[34] S. Wu and P. Flach."A scored AUC Metric for Classifier Evaluation and Selection". Proceedings of the ICML 2005 workshop on ROC Analysis in Machine Learning, 2005.

[35] M. Fatourechi, R. K. Ward, S.G. Mason, J. Huggins, A. Schlögl and G. E. Birch. "Comparison of Evaluation Metrics in Classification Applications with Imbalanced Datasets." *Seventh International Conference on Machine Learning and Applications,* 2008, PP. 777-782.

[36] V. García J.S. Sánchez R.A. Mollineda R. Alejo J.M. Sotoca "The class imbalance problem in pattern classification and learning. " 2007. http://marmota.dlsi.uji.es/WebBIB/papers/2007/1_GarciaTamida2007.pdf.

[37] N.V. Chawla, K.W. Bowyer, L.O. Hall, and W. P. Kegelmeyer. "SMOTE: Synthetic Minority Over-sampling Technique." *Journal of Artificial Intelligence Research,* Vol. 16. PP. 321–357, 2002.

[38] T. Jo and N. Japkowicz. "Class Imbalances versus Small Disjuncts." *SIGKDD Explorations*, Vol. 6, PP. 40-49, 2004.

[39] Y. Peng and J. Yao. "AdaOUBoost: Adaptive Over-sampling and Under-sampling to Boost the Concept Learning in Large Scale Imbalanced Data Sets." *ACM*, 2010, PP. 111-118.

[40] M.A. Maloof. "Learning when data sets are imbalanced and when costs are unequal and unknown." Workshop on Learning from Imbalanced Data Sets II, ICML, Washington DC, 2003.

[41] Y. Li, C. Shu and Y. Wang. "Study on Management Mechanism and Risk Control of China E-port." *Advances in Information Sciences and Service Sciences*, Vol. 3, PP. 92-103, Aug. 2011.

[42] Helen Tefera. "Application of data mining technology to identify significant patterns in census or survey data." Master's Thesis of Addis Ababa University, Addis Ababa, Ethiopia, 2003.

[43] *ASYCUDA users manual*. Ethiopian Revenues and Customs Authority, 2008.

[44] *ASYCUDA++ Documentation*, UNCTAD/SITE *Rev 1.9*.

[45] F. Ibrahim. "An empirical study on the incompetence of attributes selection criteria." In *ISMIS '96 Proceedings of the 9th International Symposium on Foundations of Intelligent Systems*, 1996, PP. 458 – 467.

[46] "Weka 3: Data Mining Software in Java",

http://www.cs.waikato.ac.nz/ml/weka/ [Jun. 22, 2012].

Annex A: Description of attributes obtained from the ERCA database

No.	Attribute	Attributes name in the data	Category	Description
1	Year	KEY_YEAR	Period	A Year that the transaction was performed.
2	Declaration number	SAD_REG_NBER	General data	The declaration identification number which is given by the system(ASYCUDA)
3	Declarant	KEY_DEC	trade operators	Transitor/declarant identification code. Transitor is an agent who works on behalf of importer in conducting customs business.
4	Reference number	KEY_NBER	General data	Reference Number that a declaration will be identified. This number is given by declarant.
5	Importer	SAD_CONSIGNEE	trade operators	importer identification code
6	Registration date	SAD_REG_DATE	Period	Declaration registration Date
7	Total item	SAD_ITM_TOTAL	General data	Total number of Items in the declaration
8	Terms of delivery	SAD_TOD_COD	General data	Contract about the delivery place (cost) of goods
9	Declaration version number	SAD_NUM	General data	Declaration internal version number(i.e. since the number is changed whenever the declaration is updated)
10	Value of goods	SAD_STAT_VAL	Financial	Statistical value (total value of goods) in a declaration
11	Total tax	SAD_TOTAL_TAXES	Financial	total amount of all duties and tax in a declaration
12	Customs border	SAD_CUO_BORD	General data	Customs office code through which the goods crossed the border
13	Mode of transport	SAD_MOT_BORD	Transport	Mode of transport at border.
14	Terms of payment	SAD_TOP_COD	General data	terms of payment
15	Currency	SAD_CUR_COD	General data	The currency code which is used for importing the cargo
16	Examiner 1(Assessor)	USR_EX1	General data	Examiner for the declaration
17	Examiner 2 (inspector)	USR_EX2	General data	Chief examiner for declaration
18	Color	SAD_CLR	color code	The color code which indicates the risk level of cargo that is assigned to specific declaration by the system
19	Country of consignment	SAD_CTY_1DLP	General data	Indicates the country of last consignment of goods.
20	Item number	ITM_NBER		Item number in a declaration
21	HS-code	SADITM_HS_COD	Procedure	Commodity identification code
22	Package	SADITM_PACK_KNDCOD	Goods Description	Type of packages (code)
23	Extended procedure code	SADITM_EXTD_PROC	Procedure	Extended customs procedure, it is part of customs procedure code(4 digits of cpc which has 7 digits)
24	National procedure code	SADITM_NAT_PROC	Procedure	National procedure (additional code to customize in the local context).
25	Country of origin	SADITM_CTY_ORIGCOD	General data	Country of origin code

26	value of an item	SADITM_STAT_VAL	Financial	Statistical value of the item in a declaration
27	Total tax for an item	SADITM_ITM_TOTAMT	Financial	Total amount of duties and taxes for an item
28	Duty rate	DUTY_RATE	Duties and Taxes detail	Duty rate for the item
29	Duty amount	DUTY_AMT	Duties and Taxes detail	Duty amount due for the item
30	Value Added Tax (VAT) rate	VAT_RATE	Duties and Taxes detail	VAT rate for the item
31	Value Added Tax (VAT) amount	VAT_AMT	Duties and Taxes detail	VAT amount due for the item
32	Excise Rate	EXC_RATE	Duties and Taxes detail	Excise tax rate for the item
33	Excise amount	EXC_AMT	Duties and Taxes detail	Excise tax amount due for the item
34	Withholding rate	WHLD_RATE	Duties and Taxes detail	Withholding tax rate for the item
35	Withholding amount	WHLD_AMT	Duties and Taxes detail	withholding tax amount due for the item
36	Sure rate	SUR_RATE	Duties and Taxes detail	Sur tax rate for the item
37	Sure amount	SUR_AMT	Duties and Taxes detail	Sur tax amount due for the item
38	Fraud type	FRAUD_TYP	Fraud type	The type of fraud detected in the declaration(i.e. UV,MC,MD.EG,OD)
39	Fraud code 1	VERIF_INF_COD1	Fraud type	Fraud code1
40	Fraud code 2	VERIF_INF_COD2	Fraud type	Fraud code2
41	Fraud code 3	VERIF_INF_COD3	Fraud type	Fraud code3
42	Examiner	EXAMINER	General data	The examiner who detect the fraud
43	Approved by	APPROVED_BY	General data	The higher officer who approved the fraud
44	Difference in percent	DIFF_IN_PERCENT	Revenue loss	Percentage of revenue loss as compared the total tax payable
45	Is revenue loss greater than or equal to ten percent	IS_GT_10	Revenue loss	It indicates whether the revenue loss is greater or less than ten percent of total amount tax and duty payable per the declaration
46	Frequency of fraud	FREQUENCY_OF_FRAUD	Fraud type	The frequency of higher fraud (i.e. greater than or equal to10% revenue loss) by the importer
47	Difference in value of goods	DIFF_IN_STAT_VAL	Revenue loss	Difference in statistical value which is taken from the local data
48	Additional tax and duty	ADDITIONAL_TAX_DUTY	Revenue loss	Revenue loss which is taken from the local data

Annex B: Attributes after two level data integration

No.	Attribute	Category	Description	Remark	
				Derived from	Used to derive
1	Year	Period	A Year that the transaction was performed.		
2	Declaration number	General data	The declaration identification number which is given by the system(ASYCUDA)		
3	Declarant	trade operators	Transitor/declarant identification code. Transitor is an agent who works on behalf of importer in conducting customs business.		
4	Reference number	General data	Reference Number that a declaration will be identified. This number is given by declarant.		
5	Importer	trade operators	importer identification code		
6	Registration date	Period	Declaration registration Date		6,7
7	Total item	General data	Total number of Items in the declaration		9
8	Terms of delivery	General data	Contract about the delivery place (cost) of goods		
9	Declaration version number	General data	Declaration internal version number(i.e. since the number is changed whenever the declaration is updated)		
10	Value of goods verified through inspection	Financial	Statistical value (total value of goods) in the declaration in the last version		15
11	Value of goods declared by importer	Financial	Statistical value of the declaration on first version		15
12	Difference in total value of goods	Financial	Difference between statistical values of the two version(No.13-No.14)	13, 14	
13	Difference in total value of goods with respect to the local data	Financial	The difference in total value with respect to the local data		
14	Total tax based on importer declaration	Financial	total amount of all duties after inspection		19
15	Total tax after inspection	Financial	total amount of all duties before inspection		19
16	Revenue loss per a declaration	Financial	Tax difference between total amount of all duties before and after inspection	19	
17	Revenue loss with respect to the local data	Financial	Revenue loss(difference between total amount of all duties in 2 version) which considers the AAL office use data(local data)		
18	Customs border	General data	Customs office code through which the goods crossed the border		
19	Mode of transport	Transport	Mode of transport at border.		
20	Terms of payment	General data	terms of payment		
21	Currency	General data	The currency code which is used for importing the cargo		
22	Examiner 1(Assessor)	General data	Examiner for the declaration		
23	Examiner 2 (inspector)	General data	Chief examiner for declaration		
24	Color	color code	The color code which indicates the risk level of cargo that is assigned to specific declaration by the system		
25	Country of consignment	General data	Indicates the country of last consignment of goods.		
26	Item number		Item number in a declaration		
27	HS-code	Procedure	Commodity identification code		

28	Package	Goods Description	Type of packages (code)		
29	Extended procedure code	Procedure	Extended customs procedure, it is part of customs procedure code(4 digits of cpc which has 7 digits)		
30	National procedure code	Procedure	National procedure (additional code to customize in the local context).		
31	Country of origin	General data	Country of origin code		
32	Verified value of an item	Financial	Statistical value of the item in a declaration after verified through inspection		37
33	Declared value of an item	Financial	Statistical value of the item in a declaration before inspection		37
34	Difference in value of an item	Financial	Difference in statistical values of an item between first & last version(No.35-No.36)	35, 36	
35	Total tax for an item after inspection	Financial	Total amount of duties and taxes for an item in a declaration after verification through inspection		40
36	Total tax for an item based on importer declaration	Financial	Total amount of duties and taxes for an item before inspection		40
37	Revenue loss per an item	Financial	Difference in total tax owed per item (no. 38-no.39)	38, 39	
38	Duty rate	Duties and Taxes detail	Duty rate for the item		
39	Duty amount	Duties and Taxes detail	Duty amount due for the item		
40	Value Added Tax (VAT) rate	Duties & Taxes detail	VAT rate for the item		
41	Value Added Tax (VAT) amount	Duties & Taxes detail	VAT amount due for the item		
42	Excise Rate	Duties & Taxes detail	Excise tax rate for the item		
43	Excise amount	Duties & Taxes detail	Excise tax amount due for the item		
44	Withholding rate	Duties & Taxes detail	Withholding tax rate for the item		
45	Withholding amount	Duties & Taxes detail	withholding tax amount due for the item		
46	Sure rate	Duties & Taxes detail	Sur tax rate for the item		
47	Sure amount	Duties & Taxes detail	Sur tax amount due for the item		
48	Fraud type	Fraud type	The type of fraud detected in the declaration(i.e. UV,MC,MD.EG,OD)		
49	Undervaluation	fraud type	Undervaluation of the statistical value	51, 57-59	
50	Mis-classification	fraud type	misclassification of commodities into incorrect HS-code	51, 57-59	
51	Mis_Discription	fraud type	Incorrect description of the commodities	51, 57-59	
52	Extra Goods	fraud type	Obtaining Extra goods that was not declared	51, 57-59	
53	Origin Difference	fraud type	Origin difference (declaring incorrect origin of goods)	51, 57-59	
54	Fraud code 1	Fraud type	Fraud code1		52-56
55	Fraud code 2	Fraud type	Fraud code2		52-56
56	Fraud code 3	Fraud type	Fraud code3		52-56
57	Examiner	General data	The examiner who detect the fraud		
58	Approved by	General data	The higher officer who approved the fraud		
59	Difference in percent	Revenue loss	Percentage of revenue loss as compared the total tax payable		
60	Is revenue loss greater than or equal to ten percent	Revenue loss	It indicates whether the revenue loss is greater or less than ten percent of total amount tax and duty payable per the declaration		
61	Frequency of fraud	Fraud type	The frequency of higher fraud (i.e. greater than or equal to10% revenue loss) by the importer		
62	Difference in value of goods	Revenue loss	Difference in statistical value which is taken from the local data		
63	Additional tax and duty	Revenue loss	Revenue loss which is taken from the local data		

Annex C: Description of the complete list of attributes (existing, derived and removed)

No.	Attribute	Category	Description	Status	Remark	
					Derived from	Used to derive
1	Year	Period	A Year that the transaction was performed.	Existing		
2	Declaration number	General data	The declaration identification number which is given by the system(ASYCUDA)	Deleted		
3	Declarant	trade operators	Transitor/declarant identification code. Transitor is an agent who works on behalf of importer in conducting customs business.	Existing		
4	Reference number	General data	Reference Number that a declaration will be identified. This number is given by declarant.	Deleted		
5	Importer	trade operators	importer identification code	"Deleted"		6,7
6	Importer region	trade operators	Importer regional address	Derived	5	
7	Trade type	trade operators	The trade type of the importer license	Derived	5	
8	Registration date	Period	Declaration registration Date	"Deleted"		9
9	Month of registration	Period	The month in which the declaration is registered	Derived	8	
10	Total item	General data	Total number of Items in the declaration	Existing		
11	Terms of delivery	General data	Contract about the delivery place (cost) of goods	Existing		
12	Declaration version number	General data	Declaration internal version number(i.e. since the number is changed whenever the declaration is updated)	Deleted		
13	Value of goods verified through inspection	Financial	Statistical value (total value of goods) in the declaration in the last version	"Deleted"		15
14	Value of goods declared by importer	Financial	Statistical value of the declaration on first version	"Deleted"		15
15	Difference in total value of goods	Financial	Difference between statistical values of the two version(No.13-No.14)	"Deleted"	13, 14	68
16	Difference in total value of goods with respect to the local data	Financial	The difference in total value with respect to the local data	"Deleted"		67
17	Total tax based on importer declaration	Financial	total amount of all duties after inspection	Deleted		19
18	Total tax after inspection	Financial	total amount of all duties before inspection	Deleted		19
19	Revenue loss per a declaration	Financial	Tax difference between total amount of all duties before and after inspection	"Deleted"	19	67
20	Revenue loss with respect to the local data	Financial	Revenue loss(difference between total amount of all duties in 2 version) which considers the AAL office use data(local data)	"Deleted"		67, 69
21	Customs border	General data	Customs office code through which the goods crossed the border	Existing		
22	Mode of transport	Transport	Mode of transport at border.	Deleted		
23	Terms of payment	General data	terms of payment	Deleted		
24	Currency	General data	The currency code which is used for importing the cargo	Existing		
25	Examiner 1(Assessor)	General data	Examiner for the declaration	Deleted		

26	Examiner 2 (inspector)	General data	Chief examiner for declaration	Deleted		
27	Color	color code	The color code which indicates the risk level of cargo that is assigned to specific declaration by the system	Deleted		
28	Country of consignment	General data	Indicates the country of last consignment of goods.	Existing		
29	Item number		Item number in a declaration	Deleted		
30	HS-code	Procedure	Commodity identification code	Existing		
31	Package	Goods Description	Type of packages (code)	Existing		
32	Extended procedure code	Procedure	Extended customs procedure, it is part of customs procedure code(4 digits of cpc which has 7 digits)	Existing		
33	National procedure code	Procedure	National procedure (additional code to customize in the local context).	Existing		
34	Country of origin	General data	Country of origin code	Existing		
35	Verified value of an item	Financial	Statistical value of the item in a declaration after verified through inspection	Deleted		37
36	Declared value of an item	Financial	Statistical value of the item in a declaration before inspection	Deleted		37
37	Difference in value of an item	Financial	Difference in statistical values of an item between first & last version(No.35-No.36)	" Deleted"	35, 36	67,68
38	Total tax for an item after inspection	Financial	Total amount of duties and taxes for an item in a declaration after verification through inspection	"Deleted"		40
39	Total tax for an item based on importer declaration	Financial	Total amount of duties and taxes for an item before inspection	"Deleted"		40
40	Revenue loss per an item	Financial	Difference in total tax owed per item (no. 38-no.39)	" Deleted"	38, 39	67
41	Duty rate	Duties & Taxes detail	Duty rate for the item	Existing		
42	Duty amount	Duties & Taxes detail	Duty amount due for the item	Deleted		
43	Value Added Tax (VAT) rate	Duties & Taxes detail	VAT rate for the item	Existing		
44	Value Added Tax (VAT) amount	Duties & Taxes detail	VAT amount due for the item	Deleted		
45	Excise Rate	Duties & Taxes detail	Excise tax rate for the item	Existing		
46	Excise amount	Duties & Taxes detail	Excise tax amount due for the item	Deleted		
47	Withholding rate	Duties & Taxes detail	Withholding tax rate for the item	Existing		
48	Withholding amount	Duties & Taxes detail	withholding tax amount due for the item	Deleted		
49	Sure rate	Duties & Taxes detail	Sur tax rate for the item	Existing		
50	Sure amount	Duties & Taxes detail	Sur tax amount due for the item	Deleted		
51	Fraud type	Fraud type	The type of fraud detected in the declaration(i.e UV,MC,MD.EG,OD)	"Deleted"		67
52	Undervaluation	fraud type	Undervaluation of the statistical value	" Deleted"	51, 57-59	68
53	Mis-classification	fraud type	misclassification of commodities into incorrect HS-code	" Deleted"	51, 57-59	68
54	Mis_Discription	fraud type	Incorrect description of the commodities	" Deleted"	51, 57-59	68
55	Extra Goods	fraud type	Obtaining Extra goods that was not declared	" Deleted"	51, 57-59	68
56	Origin Difference	fraud type	Origin difference (declaring incorrect origin of goods)	" Deleted"	51, 57-59	68
57	Fraud code 1	Fraud type	Fraud code1	"Deleted"		67, 52-56
58	Fraud code 2	Fraud type	Fraud code2	"Deleted"		67, 52-56
59	Fraud code 3	Fraud type	Fraud code3	"Deleted"		67, 52-56
60	Examiner	General data	The examiner who detect the fraud	Deleted		
61	Approved by	General data	The higher officer who approved the fraud	Deleted		

62	Difference in percent	Revenue loss	Percentage of revenue loss as compared the total tax payable	Deleted		
63	Is revenue loss greater than or equal to ten percent	Revenue loss	It indicates whether the revenue loss is greater or less than ten percent of total amount tax and duty payable per the declaration	Deleted		
64	Frequency of fraud	Fraud type	The frequency of higher fraud (i.e. greater than or equal to10% revenue loss) by the importer	Deleted		
65	Difference in value of goods	Revenue loss	Difference in statistical value which is taken from the local data	Deleted		
66	Additional tax and duty	Revenue loss	Revenue loss which is taken from the local data	Deleted		
67	Fraud	Fraud type	A class label attribute which has two values that informs whether the declaration associated with certain fraud or not	Derived	16, 19, 20, 37, 40, 51, 57-59	
68	Fraud category	Fraud type	A class label attribute which has ten distinct values and informs that what type of fraud the declaration will have.	Derived	52-56	
69	Fraud risk level	Fraud type	A class label which has three distinct values. It indicates the severity or the level fraud associated with the declaration (i.e.high, medium or low).	Derived	20, 67	70
70	Fraud level	Fraud type	It is a class which has two distinct values. It denotes the level of fraud whether it is high or low.	Derived	69	

NB: Description of status field values of the above table (i.e. Annex C): in the attribute status column of the annex there are values: Existing, Derived, Deleted and "Deleted".

These words are interpreted as:

- Existing: the attribute is originally taken from the customs data and retain its presence after data preparation as it is.
- Derived: the attribute is constructed from one or more of other attributes.
- Deleted: the attribute is found to be irrelevant and removed
- "Deleted": the attribute is used to derive another attribute but the attribute itself is deleted after the new attribute is derived

The columns under Remark in Annex C give further information about attributes labeled with status "Deleted" and Derived. If the attribute status is "Deleted", the remark shows the attribute/s that is/are derived from the deleted attribute. If the attribute status is labeled with Derived, the Remark shows from which the attribute is derived. The numbers in the remark shows the number in which the attribute is specified in Annex C.

Annex D: List of derived attributes

NO.	Attribute Name	Transformed from	Derivation
1	Importer region	Consignee	splitting sub-string from consignee
2	Importer trade type	Consignee	splitting sub-string from consignee
3	Month of registration	Registration date	splitting sub-string from registration date
4	Difference in total value of goods	• Value of goods declared • Value of goods obtained through examination	Subtracting the total value of goods which is declared by importer from total value of goods which is verified through inspection
5	Difference in value of an item	• Declared value of an item • Verified value of an item	Subtracting value declared by importer from value verified through inspection
6	Revenue loss per a declaration	• Total tax based on importer declaration • Total tax after inspection	Subtracting Total tax declared by importer from Total tax after inspection
7	Revenue loss per an item	• Total tax for an item based on importer declaration • Amount of tax verified after examination	Subtracting the amount of tax declared by importer from tax which is investigated after inspection
8	Origin Difference	Fraud type/fraud code1-3	The attribute "fraud type" has five different values. Making these values as column heading provides these attributes which informs whether a specific cargo associated with this specified fraud category or not.
9	Undervaluation		
10	Extra Goods		
11	Mis-Classification		
12	Mis-Description		
13	Revenue loss with respect to the local data	Additional tax and duty and #6	Revenue loss (difference between total amount of all duties in 2 version (it considers the AAL local data))
14	Difference in total value of goods with respect to the local data	Difference in value of goods and #4	The difference between statistical values of the two version (it considers the AAL local data)
15	FRAUD_RISK-LEVEL	• Fraud code • Additional tax and duty	By grouping instances in the dataset into three groups: fraud-not-amendable, fraud-amendable, non-fraud. Attribute "additional tax and duty" is used to identify whether the declaration is amendable or not.
16	FRAUD_CATEGRY	Combination of #8-12	It is constructed from the fraud categories (UV,OD,EG, MC,MD)
17	FRAUD	Fraud type,#5,#6,#7,13,14 and fraud code 1-3	It is derived from attribute "fraud type" through converting NULL(i.e. non-fraud) values into NO and the rest into YES and the other attributes are used to construct missed and inconsistent in fraud type
18	FRAUD_LEVEL	Fraud risk level	By excluding the non-fraudulent instances

Annex E: List of deleted attributes

Deleted Attributes	Justification for deletion
Registration date	It creates over fitting problem, because this attribute is irrelevant for future (these dates will not come in the future). So it is not irrelevant in building predictive model.
Declaration number, reference number	Both have the same information and are irrelevant in the model building because they do not have any information except to identify a declaration.
Item number	It does not have any information except to identify a the sequence of items in a declaration.
Importer	Since it has many distinct values it has been changed to higher concept level.
Value of goods declared by importer, Value of goods verified through inspection, Total tax based on importer declaration, Total tax after inspection, Verified value of an item, Declared value of an item, Total tax for an item after inspection, Total tax for an item based on importer declaration	This attributes are used to derive attributes associated with revenue lose such as *Difference in total value of goods*, *Revenue loss per a declaration*, *Difference in value of an item* and *Revenue loss per an item*. They have been deleted since they don't have information in the existence of these derived attributes. Though these derived attributes will be deleted later after they are used for class labels construction and inconsistency removal.
Duty amount, VAT amount, excise amount, withholding amount, sure amount,	In the existence of duty rate, VAT rate, Excise rate, withholding rate and sure rate, using attributes which are the derivation of these attributes will create information redundancy. If the rate is given no need of the derived value using the rate.
Fraud code1, fraud code2, fraud code3, fraud type	These attributes is used to construct fraud category attributes, such as UV, MC, MD, EG, OD. They will be useless after construction of these attributes.
Difference in percent, Is revenue loss greater than or equal to ten percent	*Is revenue loss greater than or equal to ten percent* is constructed in higher concept level of *Difference in percent*. No need of this attribute, because these attributes are beyond the scope of the study.
Difference in value, additional tax and duty	These attributes were needed to construct *Difference in total value of goods with respect to the local data*, *Revenue loss with respect to the local data*. They are useless after construction of these attribute and creates

	redundancy. Though these derived attributes will be deleted later after they are used for class labels construction and inconsistency removal.
Declaration version	It is irrelevant because we are using attributes from two version (a record contains attribute values of different version). And the use of this attribute is to indicate the record version, here no need of knowing version of the record.
Examiner, Approved by	These attributes deleted because the attributes was needed to map this attributes into usr_ex1 and usr_ex2 in order to fill missing values of these attribute. But they couldn't be mapped because they are not as such related.
Difference in total value of goods, Revenue loss per a declaration, Difference in value of an item, Revenue loss per an item, Revenue loss with respect to the local data and Difference in total value of goods with respect to the local data	These attributes are appeared (known) after examination. Using attributes which will be known after examination is worthless for predictive model building. Rather this attributes were used to construct class label and to resolve inconsistency problem in class labels.
Frequency of fraud	The concept is beyond the scope of the study.
Mode of transport, terms of payment	Their irrelevancy is determined by attribute subset selection method using weka.
Examiner 1, examiner 2	Using customs officer's information as dependent attribute for the model construction is beyond the scope our work.
Color code	In the situation that we are dealing with the inefficiency of the current selectivity method, using the systems prediction as input will lead to poor performance.
UV,OD,EG, MC,MD	These attributes are used to construct a class label attribute *Fraud category*, no need of these attributes after the construction of these attributes.

Annex F: Attributes of the final dataset

No.	Attribute	Data type
1	Extended procedure code	Nominal
2	Withholding rate	Numeric
3	National procedure code	Nominal
4	Value Added Tax (VAT) rate	Numeric
5	Total item	Numeric
6	Sure rate	Numeric
7	Terms of delivery	Nominal
8	Currency	Nominal
9	Country of consignment	Nominal
10	HS-code(commodity code)	Nominal
11	EXC_RATE	Numeric
12	Country of origin	Nominal
13	Customs border	Nominal
14	Declarant	Nominal
15	Package	Nominal
16	Year	Nominal
17	Duty rate	Numeric
18	Trade type	Nominal
19	Importer region	Nominal
20	Month of registration	Nominal
21	Fraud	Nominal
22	Fraud category	Nominal
23	Fraud level	Nominal
24	Fraud risk level	Nominal

Declaration

I, the undersigned, declare that this thesis is my original work and has not been presented for degree in any other university, and that all sources of materials used for the project have been acknowledged.

Declared by:

Name: **Meriem Muhammed**

Signature: _____

Date: 01-03-2013

Confirmed by advisor:

Name: **Sebsibe Hailemariam(PhD)**

Signature: _____

Date: _____

Place and date of submission: Addis Ababa University, March, 2013.