


ADDIS ABABA UNIVERSITY
COLLEGE OF NATURAL SCIENCE
SCHOOL OF INFORMATION SCIENCE

# Constructing a predictive model for Real-Time ATM CARD Fraud Detection

---

BY

Tsegaye Nire

JUN 2017

ADDIS ABABA UNIVERSITY
COLLEGE OF NATURAL SCIENCE
SCHOOL OF INFORMATION SCIENCE

# Constructing a predictive model for Real-Time ATM CARD Fraud Detection

**BY**
**Tsegaye Nire**

A Thesis Submitted to the School of Information Science of Addis Ababa University in Partial Fulfillment of the Requirements for the Degree of Master of Science in Information Science

**Advisors:**
**MILLION MESHESHA (PhD)**

JUN 2017

ADDIS ABABA UNIVERSITY
COLLEGE OF NATURAL SCIENCE
SCHOOL OF INFORMATION SCIENCE

# Constructing a predictive model for Real-Time ATM CARD Fraud Detection

BY

Tsegaye Nire

JUN 2017

Name and Signature of Members of the Examining Board

Advisor: ----------Signature: ---------- Date: ----------
Examiner: ---------- Signature: ---------- Date: ----------
Examiner: ---------- Signature: ---------- Date: ----------

# DECLARATION

I declare that CONSTRUCTING A PREDICTIVE MODEL FOR REAL-TIME ATM CARD FRAUD DETECTION is my own work and that all the sources that I have used or quoted have been indicated and acknowledged by means of complete references.

---

TSEGAYE NIRE

JUNE 2017

The thesis has been submitted for examination with my approval as university Advisors

---

MILLION MESHESHA

JUNE 2017

## DEDICATION

This thesis is dedicated to my parents and my best friend Abebe Mamo for the inspiration they gave me during my studies.

Acknowledgment

First and foremost my special thanks go to the almighty God for his forgiveness with the courage and endurance to successfully complete this study. Next to this; I would like to express my gratitude and heartfelt thanks to my advisor Dr. Million Meshesha for his constructive comments. I would like to extend my appreciation to my best friends Zelalem Firisa and Mesfin Hunde for their constructive comment and ideas. Finally, I would like to thank my best friends Iyasu, Lilly and Temesgen for the constant assistance and encouragement they rendered to me.

# Contents

List of Figures ..... V
List of Tables ..... VI
List of Abbreviation ..... VII
Abstract ..... VIII
CHAPTER ONE ..... 1
INTRODUCTION ..... 1
1.1. BACKGROUND ..... 1
1.2. Motivation ..... 3
1.3. Statement of the problem ..... 4
1.4. Objective of the study ..... 6
1.4.1. General objective ..... 6
1.4.2. Specific objectives ..... 6
1.5. Methodology of the study ..... 7
1.5.1. Research design ..... 7
1.5.2. Data Sources ..... 9
1.5.3. Techniques ..... 10
1.6. Scope and limitation of the study ..... 10
1.7. Significance of the study ..... 11
1.8. Ethical Concern ..... 11
1.9. Organization of the thesis ..... 11
CHAPTER TWO ..... 12
LITERATURE REVIEW ..... 12
2.1. Overview ..... 12
2.2. Data Mining and Knowledge Discovery ..... 13
2.2.1. Why Data Mining ..... 13
2.3. Data Mining Task ..... 13
2.3.1. Descriptive Modeling ..... 14
2.3.2. Predictive ..... 14
2.4. Data mining techniques ..... 14
2.4.1. Classification ..... 14
2.4.2. Clustering ..... 15

2.4.3. Association ..... 16
2.5. Data mining process models ..... 16
2.5.1. KDD process ..... 16
2.5.2. CRISP-DM process ..... 17
2.5.3. The SEMMA process ..... 19
2.6. Application of Data Mining ..... 20
2.6.1. Data Mining for Financial Sectors ..... 20
2.6.2. Data Mining for Intrusion and Fraud Detection ..... 22
2.7. Card Related Frauds ..... 24
2.7.1. Application Fraud ..... 24
2.7.2. Lost / Stolen cards ..... 24
2.7.3. Account Take Over ..... 24
2.7.4. Fake and counter feit cards ..... 25
2.8. Related works ..... 26
CHAPTER THREE ..... 29
METHODS AND APPROACHES ..... 29
3.1. Overview ..... 29
3.2. Architecture ..... 30
3.3. K-Means ..... 31
3.4. Artificial Neural Network ..... 32
3.5. Decision Tree ..... 34
3.6. Naïve Bayes ..... 35
3.7. Preprocessing Techniques ..... 37
3.8. Evaluation methods ..... 37
CHAPTER FOUR ..... 38
BUSINESS AND DATA UNDERESTANDING ..... 38
4.1. Overview ..... 38
4.2.Data Mining process model ..... 38
4.2.1. Developing and understanding of the application ..... 40
4.2.2. Creating a target data set ..... 41
4.2.3. Data Cleaning and Preprocessing ..... 43
4.3. Format conversion ..... 47

CHAPTER Five ..... 48
EXPERMENTATION ..... 48
OVERVIEW ..... 48
5.1. Experimental design ..... 49
5.2. Weka interface ..... 50
5.3. Modeling using clustering technique ..... 50
5.4. Modeling using classification techniques ..... 52
5.4.1.J48 Decision tree ..... 53
5.4.2. Artificial Neural Network Modeling (Using MLP ) ..... 57
5.4.3. Experimentation using Naïve Bayes ..... 58
5.5. Evaluation of the Discovered Knowledge ..... 60
5.6. User Interface Design ..... 63
5.7. User acceptance testing Result ..... 64
CHAPTER SIX ..... 65
CONCLUSION AND RECOMENDATION ..... 65
6.1. CONCLUSION ..... 65
6.2. RECOMMENDATION ..... 66
Annex ..... 71
Annex 1. ISO 8583 Financial transaction code ..... 75
Annex 2. Sample ARRF data for Clustering ..... 76
Annex 3. Sample ARRF data for Classification ..... 77
Annex 6. The original collected sample data set ..... 78
Annex 7. Partial View of Decision tree generated with percentage splitting ..... 79
Annex 8. The Partial View of the Decision Tree J48 Model ..... 81
Annex 9. Supporting letter from AAU to the Bank ..... 82

## List of Figures

Figure 1.1. Knowledge discovery in data base process model ----------9

Figure 2.1. The Five Stages of KDD----------17

Figure 2.2. The CRISP data mining process model----------18

Figure 3.1. Real time Fraud detection system architecture----------31

Figure 3.2. Artificial neural network----------33

Figure 3.3. Feed Forward network ----------34

Figure 3.4. Decision tree----------35

Figure 5.1. Weka Interface after data set fetched into it ----------50

Figure 5.2. The Default parameter setting of J48 Algorithm in Weka----------55

Figure 5.3. User Interface of Fraud detection System ----------63

# List of Tables

Table 4.1. Data description and data type----------42

Table 4.2. Terminal Name mapped to Terminal Location----------45

Table 4.3. Mapped status Description----------46

Table 5.1. Result of Experiment----------51

Table 5.2. Experimentation using J48 and cross validation using 10 folds----------55

Table 5.3. Experimentation using MLP and with 70% data set to training and 30% for testing-56

Table 5.4.Experimentation using MLP and cross validation using 10 folds----------57

Table 5.6. Experimentation NaiveBayes and with 70% data set to training and 30% for testing58

Table 5.8. Comparison of Experimented classifiers ----------60

Table 5.9 UAT summary report----------64

# List of Abbreviation

**ARFF:** Attribute Relation File Format

**CSV:** Comma Separated Values

**FP:** False Positive

**TP:** False Positive

**KD:** Knowledge Discovery

**KDD:** Knowledge Discovery in Database

**ATM:** Automatic Teller Machine

**PSS:** Premium switch solution

**WEKA:** Waikato Environment for Knowledge Analysis

**POS:** Point of sales

**KDP:** knowledge discovery process

**Eth switch:** Ethiopian Switch

**UCI**: unique client identifier

**NN:** Neural Network

**ANN:** Artificial Neural Network

**SOMNN: S**elf-organizing map neural network

**SVM:** Support Vector Machines

**NRI:** Not-received items

**PIN:** personal identification number

**CBS:** Core Banking Solution

**ISO:** International Organization for Standardization standard

**MLP:** Multi-layer perceptron

**STAN:** System trace audit number

**GNU:** General public license

**BDS:** Band connector switch

## Abstract

Nowadays banks are getting closer to users through many channels. Due to the advancement of technology ATM and POS channels are the most popular mode of payments; with the existence of this technology the implication of financial crime against banks is accelerating rapidly. Fraud prevention now represents one of the biggest areas of concern in financial institution. To detect these types of attack in Ethiopia there is no precondition taken, though we are adopting many Banking technologies every time.

Recognizing this problem, in this study an attempt has been made to create a predictive model that helps to detect real-time ATM frauds using data mining techniques. Empirical research methodology was used, and specifically the experiments were conducted following the Knowledge Discovery in Database process model. The dataset used in this study has been extracted from Bank ATM application server. A total of 51,500 records selected. The study was conducted using WEKA software version 3.7.9 and four data mining algorithm namely, Simple k-means to cluster the data set and J48, MLP, Naive Bayes algorithm used for classification. Since the data set has poor data quality, data preprocessing was done on the original data set. The major preprocessing activities include fill in missing values, remove outliers; resolve inconsistencies, dimensionality reduction, size reduction and data transformation.

In this study the experimental result shows that ATM card fraud detection using K-means clustering followed by J48 decision tree register the highest score of 90.8% of accuracy and user acceptance testing shows that 91 % domain expert are satisfied with the developed model. Finally the result of this study reveals that applying data mining for detecting novel type of attack on ATM card related fraud generate interesting rules with the following attributes (channel type, transaction type, transaction amount, location, Branch-location, fraud status, terminal type). One major challenge was getting labeled data. Future research directions are forwarded to come up with an applicable and reliable system in the area of this research by having labeled training dataset and enable automatic synchronization from data mining software.

# CHAPTER ONE

# INTRODUCTION

## 1.1. BACKGROUND

Banking industry is one of the most important service sectors which touch the livelihood of millions of people. Banking can be traced back to the year 1694 with the establishment of the Bank of England (Wadsworth, 1944). The bank was started by a few individuals who were money lenders with the aim of lending money at interest (Ombati, 2010).The Banking sector consists of the public sector, private sector and foreign banks apart from smaller regional and cooperative banks.

In Ethiopia, the Banking sectors are growing nowadays, notably the number of private banks in the country is increasing from year to year (Gebremeskel, 2013). The presence of many banks in the country has created fierce competition in the traditional brick and mortar based market. All private and governmental banks have partially or entirely automated their Banking operations as per the national bank directive to implement standard core Banking solution (Yirdaw, 2013), and this makes easy for the banks to quickly plug new modules like Card Banking, Mobile Banking, SMS Banking, Internet Banking services easily.

Nowadays since most card banking related services become available, customers prefer using ATM channel rather than being served at branches. Customers preferring cash-fewer environments so to align with customer need all banks are distributing massive amounts of ATM cards; this is an indication, that in the future going to branches to get services might not be needed, rather customers will have self- serve using the available technologies, therefore, expanding branches alone might be inadequate.

The good news is banks are on the way of using modern technologies to increase efficiency and effectiveness in service delivery. However, these benefits do not come without risks whether it is for information being misused, fraud, service disruption or any other attacks interrupting the normal operation of computer-based information systems. The threats and

security breaches are highly increasing in recent years globally (Rajpreetl, 2013), and are no exceptions for Ethiopia.

In the present scenario, when the term fraud comes into a discussion, Banking fraud clicks to mind. In recent years banking fraud has increased immensely. Every year fraud in bank-related fraud is rising. Fraud presents a significant cost to our economy as well, in line with adopting modern Banking Technology; as a result the security issues need to be studied well.

Fraud management is a knowledge-Intensive task (Mamo, 2013). The main Artificial Intelligence techniques used for fraud management include data mining to classify, cluster, and segment the data and automatically to find relations and rules in the data that may signify unusual patterns, including those related to fraud (Falaki, 2014),Data mining is popularly used to combat frauds because of its effectiveness, and it is a well-defined procedure that takes data as input and produces models or patterns as output.

Data mining combines techniques from machine learning, pattern recognition, statistics, database theory, and visualization to generate concepts, concept interrelations, and unusual patterns automatically from large corporate databases (Adane T. , 2011)Its primary goal is to extract knowledge from data to support the decision-making, planning, and problem-solving process. Two primary functions of data mining are prediction and description (Hand, et al, 2006). Prediction involves finding hidden values/relationships/patterns from known values, and description provides interpretation of a large database. Classification is useful for prediction, whereas clustering, pattern discovery and deviation detection are for description of patterns in the data. Data mining enables data exploration and analysis without any specific hypothesis in mind, as opposed to traditional statistical analysis, in which experiments are designed around a particular theory. While this openness adds active exploratory aspect to data mining projects, it also requires that organizations use a systematic approach in order to achieve practical results.

The primary industries to use data analysis techniques to avoid fraud were the telephony companies, the insurance companies and the banks (Fawcett, 1997). One early example of successful implementation of data analysis techniques in the Banking industry is the Falcon fraud assessment system, which is based on a neural network shell (Mamo, 2013)

Since the Banking industry in Ethiopia is expanding, the researcher believes that data mining Technology has a lot to do with it. Currently, many Ethiopian organizations are making significant business decisions based on experts' judgment, which could miss hidden but critical knowledge that is crucial for the decision being made.

Modern Banking technologies are being adopted in Ethiopia Banking industry In spite of this adoption and operation, the security issue is still in its infant stage Therefore in this study, we propose a predictive model for Ethiopian banks which helps to detect real-time ATM frauds by applying the appropriate data mining tools and techniques, as literature shows that data mining is popularly used to combat fraud because of its effectiveness.

## 1.2. Motivation

The Banking sector has been a vital institution that contributes immensely to the sustainability and maintenance of the economy in any country. There are around 20 million customers who use the commercial banks in Ethiopia, and most of the banks found in our nation are trying to get a competitive advantage by applying the existing Technology to fit with the current market and to satisfy their customer needs. Most banks implement the technologies from other standardized software firms and adopt it according to their needs, but the risks associated with those technologies are not studied in detail. Thus, the security issue is the main thing which most of the banks are worrying nowadays.

A critical task to help businesses and financial institutions including banks is to take steps to prevent fraud and to deal with it efficiently and more, when it does happen (Anderson, 2007).These days, ATM card Banking is the preferable channel of usage by customers especially in connecting with PSS and Eth-switch which helps customers to use any available ATM without being restricted only to use their own bank's ATM.

In connection with the increasing number of ATM card banking customers, lots of frauds are happening daily. A study shows that automated teller machines, ATMs, are subject to various attacks (Oguweleka, 2011). The primary reason for this is the amount of cash inside the ATM safe (up to 500 000 ETB in a high volume ATM), but also the access to customer data that in turn gives access to cash. When we see the fraudulent nature they are completely

different to the existing frauds so that detecting such frauds tough. That shows that lack of study to propose the appropriate solutions before adopting the Technology.

Currently, banks are relying only on current security Technology on the hardware. However, the current security Technology is not able to identify unique types of attacks such as lost card, Account take over and Fake card attacks that are unknown at design time. Although own security device and special expert rules can be developed after a single attack has become known, such counteractions can only be initiated after the attack has happened, i.e., after loss or damage has already been incurred, since upgrading all ATMs in service is a time- and cost-intensive process. (Santhosh Baboo, 2015)

Therefore, research has to be done in the area of detecting and preventing real-time and unique ATM frauds, and that is what motivated us to propose a solution that can analyze the fraudulent behavior from log history and database using data mining techniques to develop a predictive model. Since upgrading all ATMs in service is a time- and cost-intensive process (Maik, 2014) this will help to detect behavior of novel or unknown frauds and real-time fraud and prevent those frauds before they happen.

## 1.3. Statement of the problem

Since modern Banking technologies have made a significant stride in recent years, the vulnerabilities and security issues associated were not considered carefully. Rather most of the banks have adopted the technologies without having detail assessments just not to lose the competitive advantage associated with having the modern and relatively latest technologies (Ibrahim, Amit, & Barron, 2010).

Automated teller machine (ATM) Card Banking is one of the modern Banking technologies (Linda, 2009). ATM works in such a way that a user inserts the card into ATM machine and the ATM asks the user to enter PIN. Once the PIN is confirmed, the middle-layer will request access to the core-Banking server and at this moment (like with any transaction, the user's bank account would be locked, until such time the transaction is completed, so double-spend will not occur (Raghavendra Patidar, 2011).

If all the validation checks for the customer being able to withdraw money from his/her account is good, then the money will be dispensed by the ATM. In a review of the existing ATM security solution level, the following short comings were found (Maik, 2014):

- The current security technologies are not able to detect a unique type of attacks such as Account take over, Fake card, Application Fraud, Lost or stolen card
- The current security technologies are generic which are common in all environments.
- Counter action in the current security solution is initiated after loss or damage has already incurred.

Timely information on fraudulent activities is strategic and critical to the banking industry (Oguweleka, 2011). Several studies were conducted throughout the world and contributed significantly in detecting the risk that comes with the application of technologies like Card banking fraud. (Maik, 2014) tried to propose an approach that is able to detect known as well as unknown attacks on ATMs and that does not require additional security hardware, , (Azeem, 2014) used Artificial neural network tuned by stimulated annealing algorithm to detect real time Credit card fraud detection, Local research endeavors have been checked from AAU (Addis Ababa university) space site and Bibliographic library. Few studies has been conducted in related to Fraud detection, Study conducted by (Adane T. , 2011) mine insurance data to detect fraud by following six step cios process model using decision tree algorithm, study by (Daniel, 2013) used data mining to support fraud detection using six step cios process model and j48 classification algorithm. As the researcher understands currently no study was done in concerning ATM card fraud detection in Ethiopian context so rather than just directly applying the Technology developing real time fraud detecting system in local flavor is an important urgent issues (Worku, 2009).

Therefore, this study attempt to explore the behavior of ATM frauds using data mining techniques. This is done with the assumption that significant deviation from the expected normal behavior as a reliable indicator of an attack, which can further help to predict future ATM frauds.

To this ends the following research questions are formulated:

- Which data mining techniques can be more suitable to solve Real time ATM card fraud detection?
- What is the important attribute to analyze the fraudulent behavior of a given transaction?
- Is it possible to use the result as middle ware application to detect Real time frauds?

## 1.4. Objective of the study

### 1.4.1. General objective

The general objective of this research is to develop a predictive model that helps to detect real-time ATM frauds using data mining techniques to provide secure and reliable services to the customers.

### 1.4.2. Specific objectives

The following specific objectives are outlined to achieve the general aim stated above:

- To review relevant works done in ATM fraud detection.
- To Analyze the security gap in the ATM card service.
- To collect and prepare the data for model building by selecting and cleaning the data.
- To construct a descriptive model to identify the natural grouping of ATM frauds using clustering algorithm.
- To create a predictive model for predicting ATM fraud.
- To apply the predictive model to develop middleware application that will help to identify ATM frauds in real time.
- To evaluate the performance of the model using effectiveness measure and user acceptance testing.

## 1.5. Methodology of the study

### 1.5.1. Research design

At the existing state of the world, financial organizations increase the availability of commercial facilities by employing Core Banking solution and innovative services such as ATM cards, Internet Banking and mobile Banking services. Besides, along with the rapid advances of e-commerce, the use of ATM card has become a convenience and necessary part of financial life, though this technologies are not free from fraud, Thus, this study is designed to develop a predictive model which help to identify real time ATM frauds, depending on the log data captured from ATM Application server.

The study follows Empirical research methods because, it's a research based on experiment, observation and testing to align it with the existing environment for detecting Real-time ATM card fraud detection. The experiments were conducted following the Knowledge Discovery in Database process model (Qaiser, Shafique, & Haseeb, 2014).

The **knowledge discovery process** (KDP), also called knowledge discovery in databases, seeks new knowledge in some application domain. It is defined as the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data. The process generalizes to non-database sources of data, although it emphasizes databases as a primary source of data (Cios, Swiniarsiki, & Kuargan, 2007).

KDD is selected because of the following facts:

- The KDD process is interactive and iterative (Cios, Swiniarsiki, & Kuargan, 2007)
- KDD is the process of extracting the hidden knowledge according to from the databases and it requires relevant prior knowledge and brief understanding of application domain and goals (Qaiser U. S., 2014)
- It supports both the academics and industrial data mining projects.

**The** (Fayyad, 1996)**. KDP model consists of nine steps, which are outlined as follows:**

**1.5.2.1. Developing and understanding the application domain: -** This step includes learning the relevant prior knowledge by analyzing the business process and the goals of the end user of the discovered knowledge. In this study the relevant prior knowledge were gained by asking the domain expert having the supporting letter provided by Addis Ababa University. In this step the nature of ATM frauds, the available log files the important attributes, the existing environment are gained.

**1.5.2.2. Creating a target data set:-**This helps to selects a subset of variables (attributes) and data points (examples) that will be used to perform discovery tasks. This step usually includes querying the existing data to select the desired subset. The data sources taken from the banks application server using .csv and text format.

**1.5.2.3. Data cleaning and preprocessing: -** This step consists of removing outliers, dealing with noise and missing values in the data, and accounting for time sequence information and known changes it helps to get the explicit content of the data. The extracted data were full of missing values, inconsistent attributes that are fixed.

**1.5.2.4. Data reduction and projection: -** This step consists of finding useful attributes by applying dimension reduction and transformation methods, and finding an invariant representation of the data. In this study dimensionality reduction, size reduction is performed using Weka software.

**1.5.2.5. Choosing the data mining task: -**Here the data miner matches the goals defined in Step 1 with a particular DM method, such as classification and clustering used. Clustering method use to group the available data set into similar groups (Fraud and Non Fraud).Classification algorithm used to construct the predictive model to help predicting the outcome of the test data set using the training data set.

**1.5.2.6. Choosing the data mining algorithm: -**The data miner selects methods to search for patterns in the data and decides which models and parameters of the methods used may be appropriate. From clustering K-means algorithm selected because of its simplicity and effectiveness (Kanungo, 2002)

**1.5.2.7. Data mining: -** This step generates patterns in a particular representational form, such as classification rules, decision trees, regression models, trends, etc. In this study the model constructed using weka. K-means with MLP, NaiveBayes and J48 decision tree.

**1.5.2.8. Interpreting mined patterns: -** Here the analyst performs visualization of the extracted patterns and models, and visualization of the data based on the extracted models. The Extracted data model interpreted with the help of domain expert and the researcher.

**1.5.2.9. Consolidating discovered knowledge: -** The final step consists of incorporating the discovered knowledge into the performance system, and documenting and reporting it to the interested parties. This step may also include checking and resolving potential conflicts with previously believed knowledge. Finally the researcher developed User interface application to test the developed model and to test its applicability using user acceptance testing result.

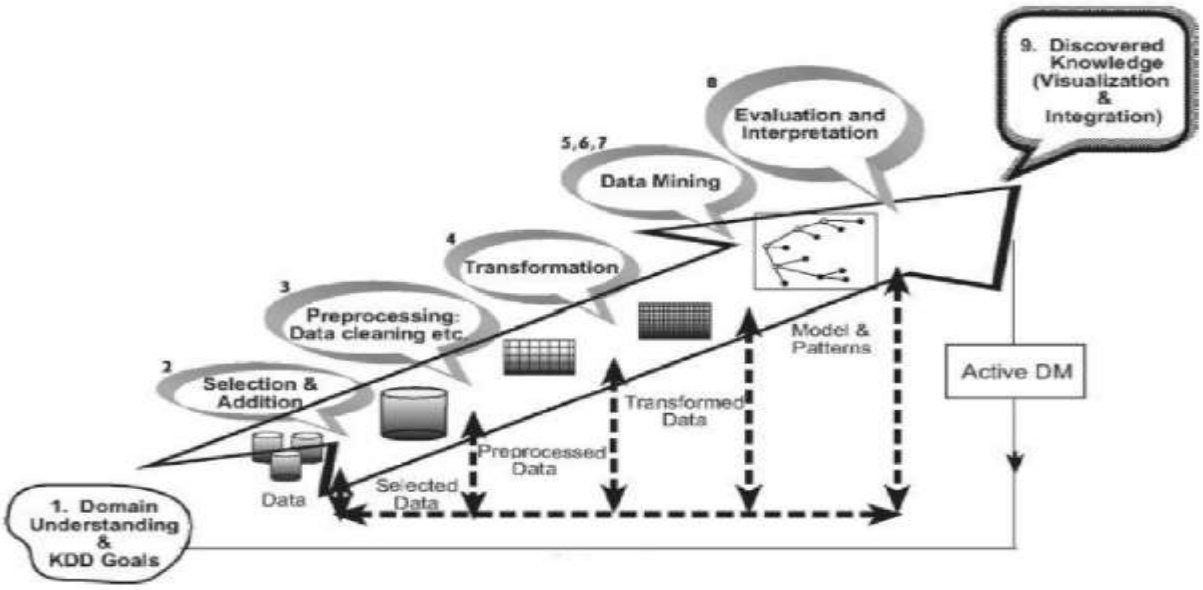



Figure 1.1. Knowledge discovery in data base process model (Fayyad, 1996)

### 1.5.2. Data Sources

Studies show that to detect real-time ATM Fraud by applying data mining Technology is the best solution. The quality of data mining depends on the quality and the volume of data. That is, to use data mining techniques, a huge number of data (Structured and unstructured) are

required for the training set. For this study, the data is collected from the log and core banking database regarding the available general facts.

### 1.5.3. Techniques

Data mining techniques are used to develop the predictive model; data mining is a process used to extract the implicit information and knowledge from the dataset. The data is extracted from the mass, incomplete, noisy, fuzzy and random data by which the data mining process is done. Particularly, in this study both clustering and classification techniques are used. Clustering helps to develop the descriptive model, it helps to find the classes with in the data set, classification techniques helps to classify the new data based on the training data set.

## 1.6. Scope and limitation of the study

The scope of this research is limited to providing predictive model by using data mining techniques to detect and prevent real-time ATM frauds for the case of banks in Ethiopia. To do these tasks clustering with classification techniques with artificial neural network and the decision tree J48 algorithm was used. This study adopted some basic techniques from the previously conducted study, since there is no such study conducted locally in related with ATM fraud detection and try to apply that to local context, the data set were taken from one of the private banks.

Due to the data set taken only from one bank this research did not include data from another bank in addition to this there are three log data captured from the server for the purpose of conducting this study the log data captured by .csv format is used the size of the data is around 200,000 records before preprocessing performed and 51,500 after the data cleaned, So further research need to be conducted by taking all banks log data set, high volume data set and those three log files since information they contain is different.

## 1.7. Significance of the study

When successfully implemented, the proposed predictive model can be used by many beneficiaries. Some of the beneficiaries are listed as follows but not limited to:

- Banks can use it as a middle ware Application to detect Frauds while transacting with customers
- Other organization like Eth-switch and PSS can use it to integrate with the existing switch solution to enable detecting fraudulent activity centrally.
- Researchers can use the model for conducting further research to develop more accurate Fraud detection system.

## 1.8. Ethical Concern

The data for this study is obtained from one of the private banks located in Ethiopia by presenting a cooperation letter from Addis Ababa University. The letter was directed to the concerned department from chief information officer (CIO) of the company.

## 1.9. Organization of the thesis

The thesis is organized into six chapters. The first chapter covers about background of the study, statement of the problem, objective of the study, scope and limitation of the study and Ethical concern. The second chapter covers literature review section which discus various concepts in related with this study particularly about data mining, about Application of data mining in financial organizations like banks and application of data mining for fraud detection and related works reviewed. In the third chapter the general research methodology discussed. In chapter four business and data understanding performed using KDD process model. In the fifth chapter the experimentation performed using data mining algorithms with WEKA software. Finally in chapter six conclusion and recommendation discussed based on the analysis and result of the experiment.

# CHAPTER TWO

# LITERATURE REVIEW

## 2.1. Overview

In the 21st century human beings used several technologies to adequate in the Society (Worku, 2009), this technologies increase the data volume captured. Each and every day people are using big data and these data are in different field. It may be in the form of documents, may be graphical formats, may be the video, may be records (varying array) .As the data are available in different formats, proper action has to be taken. Not only to analyze these data but also take a good judgment and maintain the data. As and when the customer will be required the data should be retrieved from the database and make better decision .This technique is actually is called data mining or Knowledge Hub or simply KDD (Knowledge Discovery Process).The important reason that attracted a great deal of concern in information Technology the discovery of useful information from large collections of data industry towards the field of "Data mining" is due to the observation of "we are data rich but information poor". There is a huge volume of data but we hardly able to turn them in to useful information and knowledge for top level decision making in business.

Data mining tools forecast future trends and behaviors, helps organizations to make proactive Knowledge driven decisions. The computerized, prospective analyses offered by data mining move beyond the analyses of past events provided by potential tools typical of decision support systems. Data mining tools can answer the questions that traditionally were time taking to resolve.

## 2.2. Data Mining and Knowledge Discovery

### 2.2.1. Why Data Mining

Nowadays the improvement of device and technologies has amplified the frequency of electronic data creation. Technologies are readily available in hand held devices; we are full of data but poor information. Even though the data are increasing in alarming rate, processing this data and getting benefit out of the data is a challenging duty's. Having those data only is just information/data overload so to get the hidden pattern from the data, building organization memory is needed. In order to get the most out of the available big data, it's better to apply some processing techniques, called data mining. As mentioned above development of Information Technology has generated a lot of databases and big data in various areas. In this study we focus on banking data which are very crucial, so that the research in databases and IT has given rise to an approach to store and execute this precious data for advanced decision making.

Data mining is a process of extraction of useful information and patterns from massive data. It is also called as knowledge discovery process, knowledge mining from data, knowledge extraction or data /pattern analysis. It is significant to point out that there is some vagueness about the term "data mining", which is in large part purposeful. This term initially referred to the algorithmic step in the data mining process, which initially was known as the Knowledge Discovery in Databases (KDD) process. However, over time this difference has been dropped and data mining, depending on the context, may refer to the entire process or just the algorithmic step. The entire process of KDD, envisioned by (Fayyad, 1996) there are nine steps under KDD which we follow in this research.

## 2.3. Data Mining Task

The tasks of data mining are very diverse and different because there are many patterns in an extensive database. Different kinds of techniques and methods are needed to find a different kind of pattern. Based on the kind of patterns we are looking for the goal of any data mining effort can be divided into one of the following two types (Yongjian, 2012) as descriptive and predictive modeling.

### 2.3.1. Descriptive Modeling

The descriptive data mining tasks characterize the general properties of the data in the database, Descriptive data mining focus on finding patterns describing the data that can be interpreted by humans, and produces new, non-trivial information based on the available data set. The goal of descriptive data mining is to gain an understanding of the analyzed system by uncovering patterns and relationships in large data sets. Therefore, the goal of a descriptive data mining model is to discover patterns in the data and to understand the relationships between attributes represented by the data. (Bramer, 2007).

### 2.3.2. Predictive

The Predictive data mining tasks perform inference using the current data to make prediction Predictive data mining involves using some variables or fields in the data set to predict unknown or future values of other variables of interest, and produces the model of the system described by the given data set. The goal of predictive data mining is to produce a model that can be used to perform tasks such as classification, prediction or estimation, and it is a supervised learning because the classes are predefined before the examination of the target data (Bramer, 2007).

## 2.4. Data mining techniques

### 2.4.1. Classification

Classification is a derivation of a function or model which determines the class of an object based on its attributes. A set of objects is given as the training set in which every object is represented by a vector of attributes along with its class. A classification function or model is constructed by analyzing the relationship between the attributes and the classes of the object in the training set. Such a classification function or model can be used to classify future object and develop a better understanding of the classes of the object in the database (Jiawei, 2011).
For example from a set of ATM card users who serve as a training set a classification model can be built which concludes suspected fraudulent transaction from the ATM card and ATM. The classification techniques can be used to predict fraudulent transaction based on the customers data such as location, branches, transaction history, etc.

**Types of classification models** (Jiawei, 2011)**:**

- Classification by decision tree induction
- Bayesian Classification
- Back propagation
- Support Vector Machines (SVM)
- Classification Based on Associations
- K-Nearest Neighbor Classifiers
- Case Based Reasoning
- Genetic Algorithm
- Rough Set Approach
- Fuzzy Set Approach

### 2.4.2. Clustering

Clustering is the identification of classes, also called cluster or groups, for a set of objects whose classes are unknown. The objects are clustered that the intra class similarities are maximized and the interclass similarities are minimized based on some criteria defined on the attributes of objects. Once the clusters are known, the objects are labeled with their corresponding clusters and communal feature of the objects in a cluster are summarized to form the class description (Jiawei, 2011).

For example, a bank may cluster customer into several groups based on the similarities of their age, income, residence, etc, and the common characteristics of the customers in a group can describe that group of customer. The cluster helps the bank to understand the customers better and thus provide more relevant products and customized services.

**Types of clustering methods** (Jiawei, 2011)**:**

- Hierarchical (divisive) Methods
- Partitioning Methods
- Density Based Methods
- Grid-Based Methods
- Model Based Algorithms

### 2.4.3. Association

Association and correlation are usually to find frequent item set findings among big data sets. This type of finding helps businesses to make certain decisions, such as catalog design, cross marketing and customer shopping behavior analysis. Association Rule Algorithms need to be able to produce rules with confidence values less than one. However the number of likely Association Rules for a given dataset is very large and a high proportion of the rules are usually of little (if any) value. (Jiawei, 2011)

**Types of Association rule:**

- Multilevel association rule
- Multidimensional association rule
- Quantitative association rule

## 2.5. Data mining process models

### 2.5.1. KDD process

The KDD process, as presented in (Fayyad, 1996) is the process of using DM methods to extract what is deemed knowledge according to the specification of measures and thresholds, using a database along with any required preprocessing, sub sampling, and transformation of the database. There are considered five stages, presented in figure 2:

1. Selection: - This stage consists of generating a target data set, or concentrating on a subset of variables or data samples, on which discovery is to be done.
2. Preprocessing: - This stage consists of the target data cleaning and preprocessing in order to get consistent data.
3. Transformation: - This stage consists of the transformation of the data using dimensionality reduction or transformation methods.
4. Data Mining: - This stage consists of the searching for patterns of interest in a particular Representational form, depending on the data mining objective (usually, prediction)
5. Interpretation/Evaluation: - This stage consists of the interpretation and evaluation of the mined patterns.

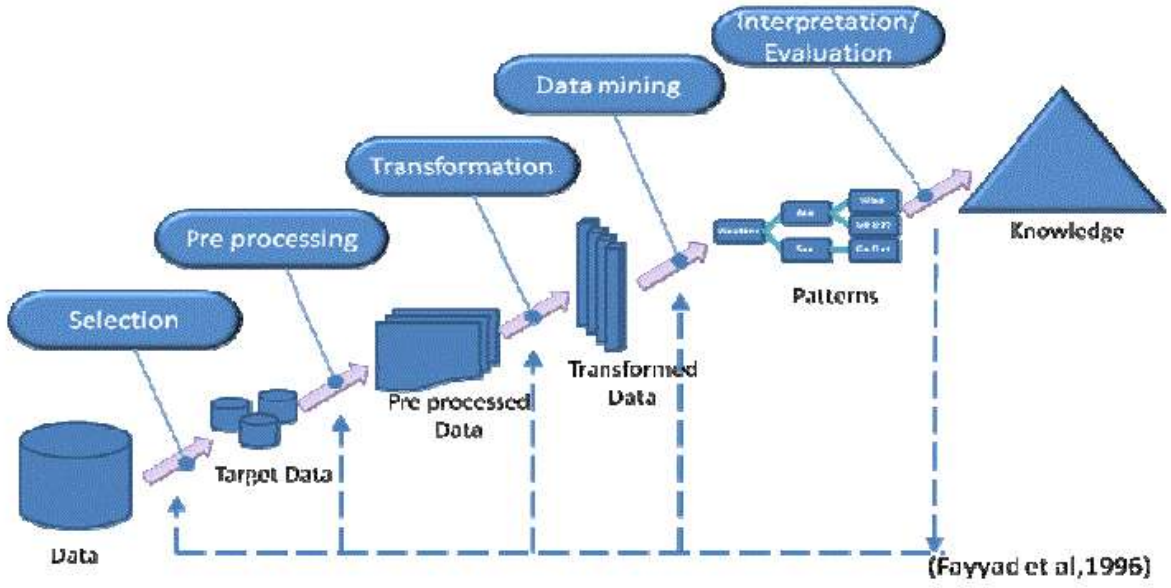



Figure 2.1. The Five Stages of KDD (Fayyad, 1996)

The KDD process is interactive and iterative, involving numerous steps with many decisions being made by the user. (Cios, Swiniarsiki, & Kuargan, 2007).Additionally, the KDD process must be preceded by the development of an understanding of the application domain, the relevant prior knowledge and the goals of the end-user. It also must be continued by the knowledge consolidation by incorporating this knowledge into the system (Fayyad, 1996).

### 2.5.2. CRISP-DM process

The CRISP-DM process was developed by using the effort of a consortium initially composed with DaimlerChrysler, SPSS and NCR. CRISP-DM stands for Cross-Industry Standard Process for Data Mining. It consists of a cycle that comprises six stages (figure 3):

1. **Business understanding: -**This initial phase focuses on understanding the project objectives and requirements from a business perspective, then converting this knowledge into a data mining problem definition and a preliminary plan designed to achieve the objectives.
2. **Data understanding: -**The data understanding phase starts with an initial data collection and proceeds with activities to get familiar with the data, to identify data quality problems, to discover first insights into the data or to detect interesting subsets to form hypotheses for hidden information.

3. **Data preparation: -**The data preparation phase covers all activities to construct the final dataset from the initial raw data.
4. **Modeling: -** In this phase, various modeling techniques are selected and applied and their parameters are calibrated to optimal values.
5. **Evaluation: -**At this stage the model (or models) obtained are more thoroughly evaluated and the steps executed to construct the model are reviewed to be certain it accurately achieves the business objectives.
6. **Deployment: -**Creation of the model is not the end of the project. Even if the purpose of the model is to increase knowledge of the data, the knowledge gained will need to be organized and presented in a way that the customer can use it.

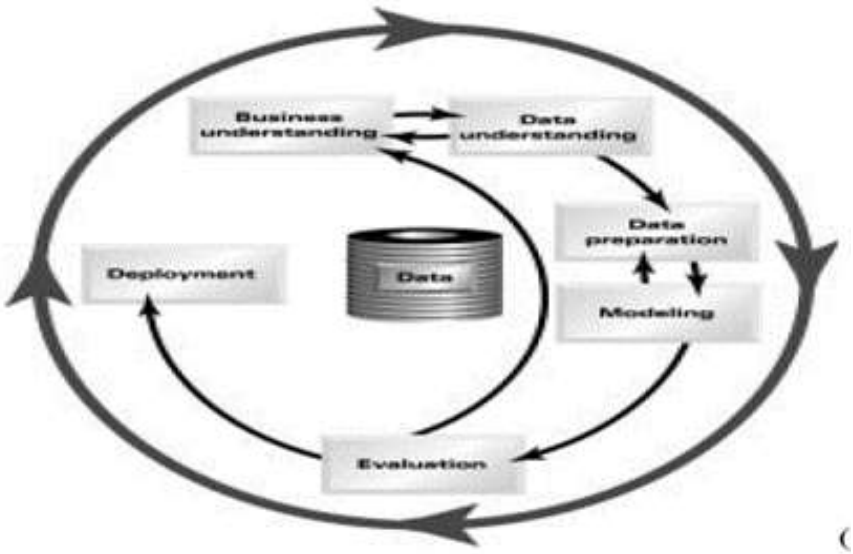



Figure 2.2. The CRISP data mining process model(Chapman et al, 2000)

The sequence of the six stages is not rigid, as is schematize in figure 2. CRISP-DM is extremely complete and documented. All his stages are duly organized, structured and defined, allowing that a project could be easily understood or revised (Santos &Azevedo, 2005). Although the CRISP-DM process is independent of DM chosen tool, it is linked to the SPSS Clementine software

### 2.5.3. The SEMMA process

The SEMMA process was developed by the SAS Institute. The acronym SEMMA stands for Sample, Explore, Modify, Model, Assess, and refers to the process of conducting a data mining project. The SAS Institute considers a cycle with five stages for the process (Jiawei, 2011):

1. **Sample: -** This stage consists on sampling the data by extracting a portion of a large data set big enough to contain the significant information, yet small enough to manipulate quickly. This stage is pointed out as being optional.
2. **Explore: -**This stage consists of the exploration of the data by searching for unanticipated trends and anomalies to gain understanding and ideas.
3. **Modify: -**This stage consists of the modification of the data by creating, selecting, and transforming the variables to focus the model selection process.
4. **Model: -** This stage consists of modeling the data by allowing the software to search automatically for a combination of data that reliably predicts the desired outcome.
5. **Assess:-**This stage consists of assessing the data by evaluating the usefulness and reliability of the findings from the data mining process and estimate how well it performs.

Although the SEMMA process is independent of DM chosen tool, it is linked to the SAS Enterprise Miner software and pretends to guide the user on the implementations of DM applications. SEMMA offers an easy to understand process, allowing an organized and adequate development and maintenance of DM projects. It thus confers a structure for its conception, creation and evolution, helping to present solutions to business problems as well as to find DM business goals.

## 2.6. Application of Data Mining

Because data mining is a relatively young discipline with wide and diverse applications, there is still a non-trivial gap between general principles of data mining and application specific, effective data mining tools. In this section, we study a few application domains and discuss how customized data mining tools should be developed for such applications..

### 2.6.1. Data Mining for Financial Sectors

Most banks and financial institutions offer a broad range of Banking services (such as checking and savings accounts for business or individual customers), credit (such as business, mortgage, and automobile loans), and investment facilities (such as mutual funds). Some also offer insurance services and stock investment services. Financial data collected in the Banking and financial industry is often comparatively complete, reliable, and of high quality, which facilitates systematic data analysis and data mining. (kamber, 2006)

**Here we present a few typical cases:**

**Design and construction of data warehouses for multidimensional data analysis and**

**Data mining:** Like many other applications, data warehouses need to be constructed for banking and financial data. Multidimensional data analysis methods should be used to analyze the general properties of such data. For example, one may like to view the debt and revenue changes by month, by region, by sector, and by other factors, along with maximum, minimum, total, average, trend, and other statistical information. Data warehouses, data cubes, multi-feature and discovery-driven data cubes, characterization and class comparisons, and outlier analysis all play important roles in financial data analysis and mining.

**Loan payment prediction and customer credit policy analysis:** Loan payment prediction and customer credit analysis are critical to the business of a bank. Many factors can strongly or weakly influence loan payment performance and customer credit rating. Data mining methods, such as attribute selection and attribute relevance ranking, may help identify important factors and eliminate irrelevant ones. For example, factors related to the risk of loan payments include loan-to-value ratio, term of the loan, debt ratio (total amount of monthly debt versus the total monthly income), payment to-income ratio, customer income level, education level, residence

region, and credit history. Analysis of the customer payment history may find that, say, the payment-to-income ratio is a dominant factor, while education level and debt ratio are not. The bank may then decide to adjust its loan-granting policy so as to grant loans to those customers whose applications were previously denied but whose profiles show relatively small risks according to the critical factor analysis.

**Classification and clustering of customers for targeted marketing:** Classification and Clustering methods can be used for customer group identification and targeted marketing.

For example, we can use classification to identify the most crucial factors that may influence a customer's decision regarding Banking. Multidimensional clustering techniques may identify customers with similar behaviors regarding loan payments. These can help identify customer groups, associate a new customer with an suitable customer group, and facilitate targeted marketing.

**Detection of money laundering and other financial crimes:** Detecting money laundering and other financial crimes, is significant to integrate information from multiple databases (like bank transaction databases, and federal or state crime history databases), as long as they are potentially related to the study. Multiple data analysis tools can then be used to detect unusual patterns, such as large amounts of cash flow during certain periods, by particular groups of customers. Useful tools include data visualization tools (to display transaction activities using graphs by time and by groups of customers), linkage analysis tools (to identify links among different customers and activities), classification tools (to filter unrelated attributes and rank the highly related ones), clustering tools (to group different cases), outlier analysis tools (to detect unusual amounts of fund transfers or other activities), and sequential pattern analysis tools (to characterize unusual access sequences). These tools may identify meaningful relationships and patterns of activities and help investigators focus on suspicious cases for further detailed examination (Kovalerchuk, 2014).

### 2.6.2. Data Mining for Intrusion and Fraud Detection

The security of our computer systems and data is at incessant risk. The extensive growth of the Internet and increasing availability of tools and tricks for intruding and attacking networks have prompted intrusion detection to become a critical component of network administration. An intrusion can be defined as any set of actions that threaten the integrity, confidentiality, or availability of a network resource (such as user accounts, file systems, system kernels, and so on).

Most commercial intrusion detection systems are limiting and do not provide a complete solution. Such systems typically employ a misuse detection strategy. Misuse detection searches for forms of program or user behavior that match known intrusion scenarios, which are stored as signatures. These hand-coded signatures are laboriously provided by human experts based on their extensive knowledge of intrusion techniques. If a pattern match is found, this signals an event for which an alarm is raised. Human security analysts evaluate the alarms to decide what action to take, whether it be shutting down part of the system, alerting the relevant Internet service provider of suspicious traffic, or just noting unusual traffic for future reference. An intrusion detection system for a large complex network can typically generate thousands or millions of alarms per day, representing an overwhelming task for the security analysts. Because systems are not static, the signatures need to be updated whenever new software versions arrive or changes in network configuration occur. An additional, major drawback is that misuse detection can only identify cases that match the signatures. That is, it is unable to detect new or previously unknown intrusion techniques. Novel intrusions may be found by anomaly detection strategies. Anomaly detection builds models of normal network behavior (called profiles), which it uses to detect new patterns that significantly deviate from the profiles. Such deviations may represent actual intrusions or simply be new behaviors that need to be added to the profiles. The main advantage of anomaly detection is that it may detect novel intrusions that have not yet been observed. Typically, a human analyst must sort through the deviations to ascertain which represent real intrusions. A limiting factor of anomaly detection is the high percentage of false positives. New patterns of intrusion can be added to the set of signatures for misuse detection.

As we can see from this discussion, current traditional intrusion detection systems face many limitations. That has led to an increased interest in data mining for intrusion detection.

**The following are areas in which data mining Technology may be applied or Further developed for intrusion and fraud detection:**

**Fraudulent pattern analysis and the identification of unusual patterns: -**Fraudulent activity costs the telecommunication industry millions of dollars per year. It's essential to (1) identify potentially fraudulent users and their atypical usage patterns; (2) detect attempts to gain fraudulent entry to customer accounts; and(3) discover unusual patterns that may need special attention, such as busy-hour frustrated call attempts, switch and route congestion patterns, and periodic calls from automatic dial-out equipment (like fax machines) that have been improperly programmed. Many of these patterns can be discovered by multidimensional analysis, cluster analysis, and outlier analysis (Sowjanya & Jysotsna, 2013).

**Development of data mining algorithms for intrusion detection: -**Data mining algorithms can be used for misuse detection and anomaly detection. In misuse detection, training data are labeled as either "normal" or "intrusion." A classifier can then be derived to detect known intrusions. Research in this area has included the application of classification algorithms, association rule mining, and cost-sensitive modeling. Anomaly detection builds models of normal behavior and automatically detects significant deviations from it. Supervised or unsupervised learning can be used. In a supervised approach, the model is developed based on training data that are known to be "normal." In an unsupervised approach, no information is given about the training data. Anomaly detection research has included the application of classification algorithms, statistical approaches, clustering, and outlier analysis. The techniques used must be efficient and scalable, and capable of handling network data of high volume, dimensionality, and heterogeneity (Golmohammadi & Zaine, 2015).

**Association and correlation analysis, and aggregation to help select and build discriminating Attributes: -** Association and correlation mining can be applied to find relationships between system attributes describing the network data. Such information can provide insight regarding the selection of useful attributes for intrusion detection. New attributes derived from aggregated data may also be helpful, such as summary counts of traffic matching a particular pattern.

**Analysis of stream data: -**Due to the transient and dynamic nature of intrusions and malicious attacks, it is crucial to perform intrusion detection in the data stream environment. Moreover, an

event may be normal on its own, but considered malicious if viewed as part of a sequence of events. Thus it is necessary to study what sequences of events are frequently encountered together, find sequential patterns, and identify outliers. Other data mining methods for finding evolving clusters and building dynamic classification models in data streams are also necessary for real-time intrusion detection (Anita & Ravindra, 2013).

## 2.7. Card Related Frauds

### 2.7.1. Application Fraud

This type of fraud occurs when a person falsifies an application to acquire a credit card. Application fraud can be committed in three ways (Tej paul, 2003):

- Assumed identity, where an individual illegally obtains personal information of another individual and opens accounts in his or her name, using partially legitimate Information.
- Financial fraud, where an individual provides false information about his or her financial status to acquire credit.
- Not-received items (NRIs) also called postal intercepts occur when a card is stolen from the postal service before it reaches its owner's destination.

### 2.7.2. Lost / Stolen cards

A card is lost/stolen when a legitimate account holder receives a card and loses it or someone steals the card for criminal purposes. This type of fraud is in essence the easiest way for a fraudster to get hold of other individual's credit cards without investment in Technology. It is also perhaps the hardest form of traditional credit card fraud to tackle (Tej paul, 2003).

### 2.7.3. Account Take Over

This type of fraud occurs when a fraudster illegally obtains a valid customers' personal information. The fraudster takes control of (takeover) a legitimate account by either providing the customer's account number or the card number. The fraudster then contacts the card issuer, masquerading as the genuine cardholder, to ask that mail is redirected to a new address. The fraudster reports card lost and asks for a replacement to be sent (Tej paul, 2003).

### 2.7.4. Fake and counter feit cards

The creation of counterfeit cards, together with lost / stolen cards poses the highest threat in credit card frauds. Fraudsters are always finding new and more innovative ways to create counterfeit cards. Some of the techniques used for creating false and counterfeit cards are listed below (Tej paul, 2003):

1. **Erasing Magnetic Stripe:** A fraudster can tamper an existing card that has been acquired illegally by erasing the metallic strip with a powerful electro-magnet. The fraudster then tampers with the details on the card so that they match the details of a valid card, which they may have attained, e.g., from a stolen till roll. When the fraudster begins to use the card, the cashier will swipe the card through the terminal several times, before realizing that the metallic strip does not work. The cashier will then proceed to input the card details into the terminal manually. This form of fraud has high risk because the cashier will be looking at the card carefully to read the numbers. Doctored cards are, as with many of the traditional methods of credit card fraud, becoming an outdated method of illicit accumulation of either funds or goods.

2. **Creating a Fake Card:** A fraudster can create a fake card from scratch using sophisticated machines. That is the most common type of fraud though fake cards require a lot of effort and skill to produce. Modern cards have many security features all designed to make it difficult for fraudsters to make good quality forgeries. Holograms have been introduced in almost all credit cards and are tough to forge effectively. Embossing holograms onto the card itself is another problem for card forgers.

3. **Altering card details:** A fraudster can alter cards by either re-embossing them by applying heat and pressure to the information originally embossed on the card by a legitimate card manufacturer or by re-encoding them using computer software that encodes the magnetic stripe data on the card.

4. **Skimming:** Most cases of counterfeit fraud involve skimming, a process where actual data on a card's magnetic stripe is electronically copied onto another. Skimming is fast emerging as the most popular form of credit card fraud. Employees/cashiers of business establishments have been found to carry pocket skimming devices, a battery-operated electronic magnetic stripe reader, with which they swipe customer's cards to get hold of

customer's card details. The fraudster does this while the customer is waiting for the transaction to be validated through the card terminal. Skimming takes place unknown to the cardholder and is thus tough, if not impossible to trace. In other cases, the details obtained by skimming are used to carry out fraudulent card-not-present transactions by fraudsters. Often, the cardholder is unaware of the fraud until a statement arrives showing purchases they did not make.

5. **White plastic:** A white plastic is a card-size piece of plastic of any color that a fraudster creates and encodes with legitimate magnetic stripe data for illegal transactions. This card looks like a hotel room key but contains legitimate magnetic stripe data that fraudsters can use at POS terminals that do not require card validation or verification (for example, petrol pumps and ATMs).

## 2.8. Related works

Several authors have tried to study card Banking fraud detection system especially in developed countries. There are several research combined to develop real time credit card fraud detection system by combining data mining methods. Some of them are listed as follows:

Study by (Maik, 2014) attempt to propose an approach that can detect known as well as unknown attacks on ATMs and that does not require additional security hardware. The idea is to utilize automatic model generation techniques to learn patterns of normal behavior from the status information of standard devices comprised in an ATM; a significant deviation from the learned behavior is an indicator of a fraud attempt. Finally empirical evaluation using a real-world data set that has been recorded on a public ATM within a time period of nine weeks and it shows promising results and underlines.

(Azeem, 2014)Used artificial neural network tuned by stimulated annealing algorithm to detect real time Credit card fraud detection. As the author stated there will be a lot of work to be done for fraud detection because the activity of user is different in each transaction which causes the training of any ANN to be difficult and future works need to be done by combining Simulated Annealing and Genetic Algorithm to create a best model, it will gives better result than any other.

(Santhosh Baboo, 2015) Investigated how the hidden Markov model can detect whether an incoming transaction is fraudulent or not depending on spending pattern of customers.
(Oguwelka, 2011) Used one of data mining techniques (Neural Network) to detect credit card fraud. The CCF detection system was designed to run at the background of existing banking software and attempt to discover illegitimate transactions entering on real-time basis. The study reinforces the validity and efficiency of ANNs as a research tool and laid a solid groundwork for intelligent detection methodologies to be used in an operational fraud detection system.

(Raghavendra Patidar, 2011)Study fraudulent transaction through the neural network along with the genetic algorithm this paper provides an overview of the concept of Data Mining and different frauds in Banking. The article by (Linda Delamaire (UK), 2009)has reviewed recent findings in the credit card field and defined different types of frauds and review alternative way of detecting frauds by comparing and analyzing the recent study conducted. The study made in this paper is likely to have beneficial attributes in terms of cost savings and time efficiency. The significance of the application of the techniques reviewed here is in the decreasing of credit card fraud and assures that still ethical issues are there when genuine credit card customer misclassified.

There are Local researches that attempt to apply data mining for fraud detection. In the study conducted by (Daniel, 2013) an attempt has been made to apply the DM technology in support of detecting and predicting fraudulent tax claims in the tax authorities using Six step Cios data mining process model, the model is then evaluated with a separate test dataset and scored an accuracy of 98.65% of classifying new tax datasets as fraud and non-fraud suspicious claims. The study has shown that it is possible to identify those fraud suspicious tax claims and suggest concrete solutions for detecting them, using DM techniques.

Study conducted by (Tesfaye, 2013) attempt to construct predictive model for subscription fraud detection using data mining techniques random forest algorithms which gives 99% of accuracy to achieve the objective the researcher used call Detail Record (CDR) data and recommended to use additional data source.

However, no studies are done in local flavor concerning ATM card Fraud detection since fraud natures are changing from time to time and the behavior of frauds are different from the developed countries (Zeineba, 2005)..

# CHAPTER THREE

# METHODS AND APPROACHES

## 3.1. Overview

The Experiments were conducted following the Knowledge Discovery in Database process model. KDD is an iterative and interactive model (Qaiser, Shafique, & Haseeb, 2014).KDD is the process of extracting the hidden knowledge according from the databases, KDD requires relevant prior knowledge and brief understanding of application domain and goals (Qaiser, Shafique, Umair, & Haseeb, 2014), In this study, we have incorporated KDD process. Both classifications (supervised) and clustering (unsupervised) data mining tasks are used in this study. From clustering Artificial neural network algorithm is used and for classification decision tree algorithm is employed. As various study shows neural network based fraud detection is the best way to detect a real time and new type of fraud because neural network is supervised in nature (S. O. Falaki1, 2012).

The Neural network based fraud detection is based entirely on the human brain working principal. Neural network Technology has made a computer capable of “thinking”. As human brain learn through experience and use its knowledge or experience in making the decision in daily life problem the same technique is applied with the ATM card fraud detection Technology. When a particular consumer uses its ATM card, there is a fixed pattern of credit card use, made by the way consumer uses its ATM card.

In spite of pattern of ATM card use, a neural network is also trained about the various credit card fraud faced by a particular bank previously. Based on the pattern of uses of ATM card, the neural network make use of prediction algorithm on these pattern data to classify that whether a particular transaction is fraudulent or genuine. When an unauthorized user is using a credit card the neural network based fraud detection system check for the pattern used by the fraudster and matches with the pattern of the original card holder on which the neural network has been trained, if the pattern matches the neural network declare the transaction ok.

## 3.2. Architecture

Conceptual frameworks provide generalizations about processes and the interaction of the process. The framework suggested for fraud detection in transactions connected to ATM card is shown in Figure 3.1. First, we state some suppositions about a proper transaction performance by an ATM card owner work so that training data need to be supplied to the detection engine then the detection engine will apply artificial neural network algorithm. User has ATM card for a specific account number. In other words, all transactions of the ATM card owner are performed manually or by the credit card in question. Conditions considered are mostly carried out by terminal either by using ATM stations or POS machines for example are checked if the account is active and has sufficient balance. Correct Pin to check this information the terminal will automatically read the card information. The design of the architecture is based on two components; one component is fraud detection engine which analyze the existing data by applying artificial neural network or J48 and the second component is an investigator this module will analyze the current Card behavior and classify the customer to the right Fraudulent transaction.

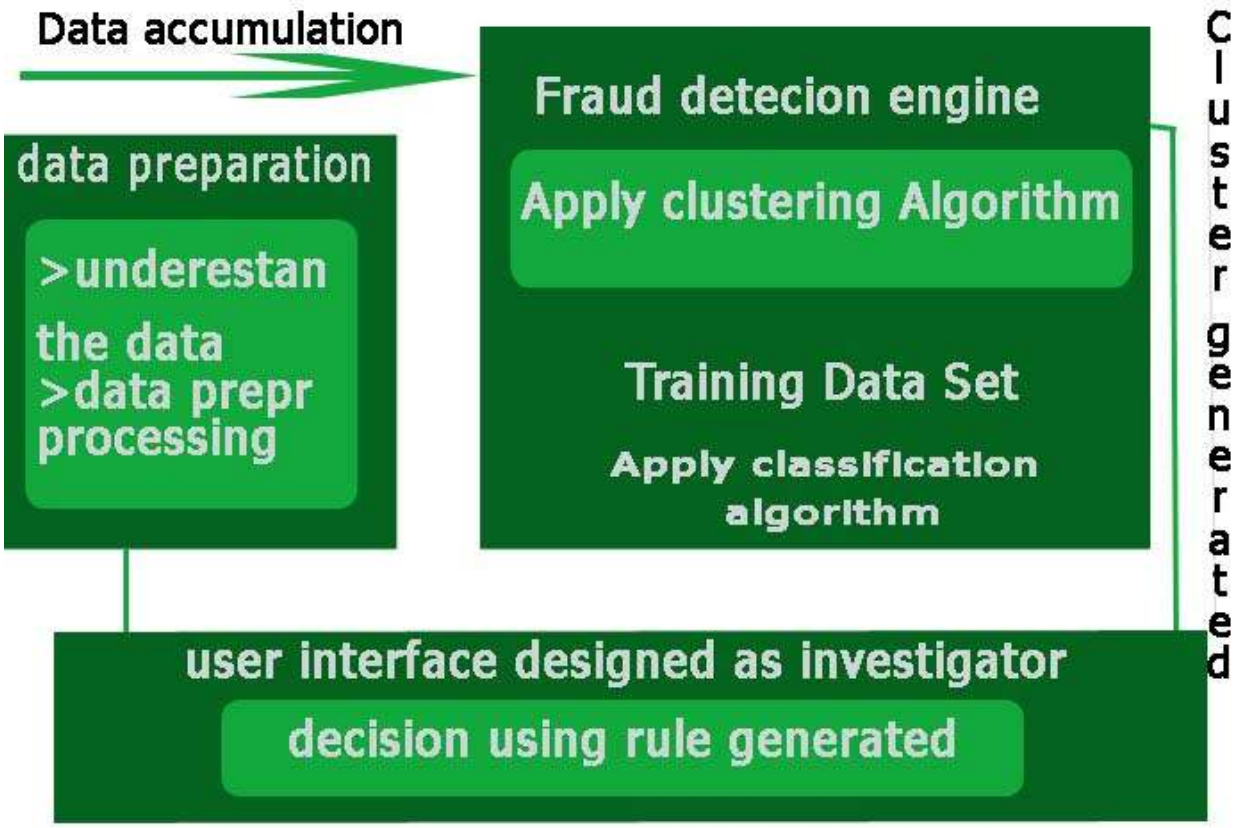



Figure 3.1. Real time Fraud detection system architecture

## 3.3. K-Means

K-means is one of the simplest unsupervised learning algorithms that solve the well-known clustering problem (Xie, Jiang, Xie, & Gao, 2011). K-means algorithm is used for discovering transaction weather it is fraud transaction or legitimate transaction. Cluster is developed to discover fraud in credit card fraud transaction which is divided into Fraud and Non Fraud Transaction. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters (assume k clusters) fixed a priori. The main idea is to define k centroid, one for each cluster.

This centroid should be placed in achievable way because of different location causes different result. So, the better choice is to place them as much as possible far away from each other. The next step is to take each point belonging to a given data set and associate it to the nearest centroid. When no point is pending, the first step is completed and an early group age is done. At

this point we need to re-calculate k new centroid as bary center of the clusters resulting from the previous step. After we have these k new centroid, a new binding has to be done between the same data set points and the nearest new centroid. A loop has been generated. As a result of this loop we may notice that the k centroid change their location step by step until no more changes is done. In other words centroids do not move any more.

Before conducting the experiments the researcher has tried to compare and select best descriptor of the data source and discus with the domain expert in respective to fraud detection then decide to cluster the data using two number of clusters, so accordingly the first experiment is conducted to obtain cluster using K-mean, it is an algorithm to classify or to group your objects based on attributes/features into K number of group. K is positive integer number. The grouping is done by minimizing the sum of squares of distances between data and the corresponding cluster centroid.

The algorithm has following steps (Raval & Jani, 2015):-

1. Select K points as the initial centroid.
2. Assign all points to the closest centroid.
3. Recompute the centroid of each cluster.
4. Repeat steps 2 and 3 until the centroid don't change.

## 3.4. Artificial Neural Network

Neural network more accurately called artificial neural network (ANN), are computational models that consist of many simple processing units that communicate by sending signals to one another over a large number of weighted connections (Karlik & Yuksek, 2007). Artificial Neural Networks were inspired by biological findings relating to the behavior of the brain as a network of units called neurons. The human brain is estimated to have around 10 billion neurons each connected on average to 10,000 other neurons. Each neuron receives signals through synapses that control the effects of the signal on the neuron. These synaptic connections are believed to play a key role in the behavior of the brain. The fundamental building block in an Artificial Neural Network is the mathematical model of a neuron as shown in Figure 3.2.

**The three primary components of the (artificial) neuron are:**

1. The synapses or connecting links that provide weights, $w_j$ , to the input values, $x_j$ for j =1,...m;

2. An adder that sums the weighted input values to compute the input to the activation function v=

3. An activation function g (also called a squashing function) that maps v to g(v) the output value of the neuron. This function is a monotone function.

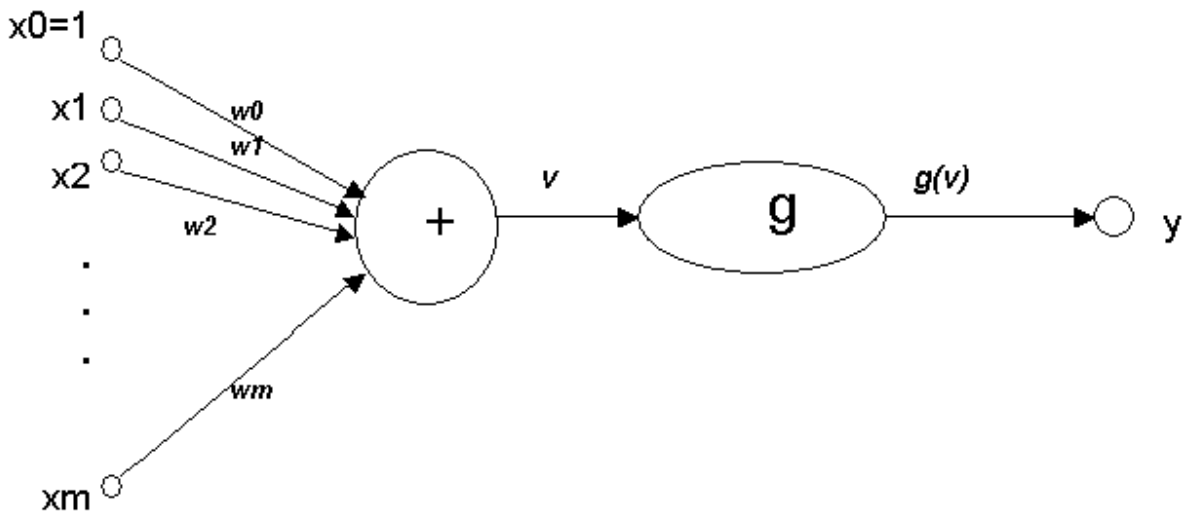



Figure 3.2. artificial neural network (Karlik & Yuksek, 2007)

While there are numerous different (artificial) neural network architectures that have been studied by researchers, the most successful applications in data mining of neural networks have been multilayer feed forward networks. These are networks in which there is an input layer consisting of nodes that simply accept the input values and successive layers of nodes that are neurons as depicted in Figure 5. The outputs of neurons in a layer are inputs to neurons in the next layer. The last layer is called the output layer. Layers between the input and output layers are known as hidden layers. Figure 6 is a diagram of this architecture.

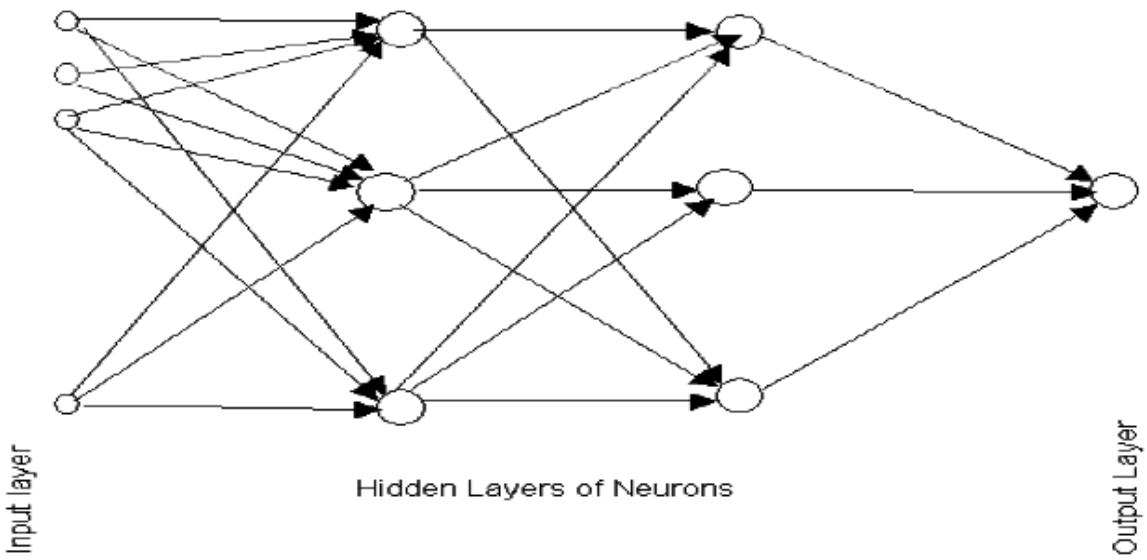



Figure 3.3. Feed Forward network (Karlik & Yuksek, 2007)

A crucial features of an Artificial neural network is their adaptive nature where "learning by example" replaces" programming" in solving problems (Mohamadi, Muda, & Yasin). This feature makes such computational models very attractive in application domain where one has a little or incomplete understanding of the problem to be solved but where training data is readily available.

To conclude ANN have powerful pattern classification and pattern recognition capability through learning and generalize from experience. ANN are non-linear data driven self-adaptive approach as opposed to traditional model based method (Kumar A. , 2011).

## 3.5. Decision Tree

Once the clustering model is developed, the next step of this study is developing the Predictive model using the classification techniques. A decision tree is a classifier expressed as a recursive partition of the instance space (Chauhan & Chauhan, 2013). The decision tree consists of nodes that form a rooted tree, meaning it is a directed tree with a node called "root" that has no incoming edges. All other nodes have exactly one incoming edge. A node with outgoing edges is called an internal or test node. All other nodes are called leaves (also known as terminal or decision nodes). In a decision tree, each internal node splits the instance space into two or more sub-spaces according to a certain discrete function of the input attributes values. In the simplest and most common case, each test considers a single attribute, such that the instance space is

partitioned according to the attribute's value. In regard to numeric attributes, the condition refers to a range.

Each leaf is assigned to one class representing the (Qaiser U. S., 2014) appropriate target value. Alternatively, the leaf may hold a probability vector indicating the probability of the target attribute having a certain value. Instances are classified by navigating them from the root of the tree down to a leaf, according to the outcome of the tests along the path. Figure 7 describes a decision tree that reasons whether or not a potential customer will respond to a direct mailing. Internal nodes are represented as circles, whereas leaves are denoted as triangles. Note that this decision tree incorporates both nominal and numeric attributes. Given this classifier, the analyst can predict the response of a potential customer (by sorting it down the tree), and understand the behavioral characteristics of the entire potential customers population regarding direct mailing. Each node is labeled with the attribute it tests, and its branches are labeled with its corresponding values.

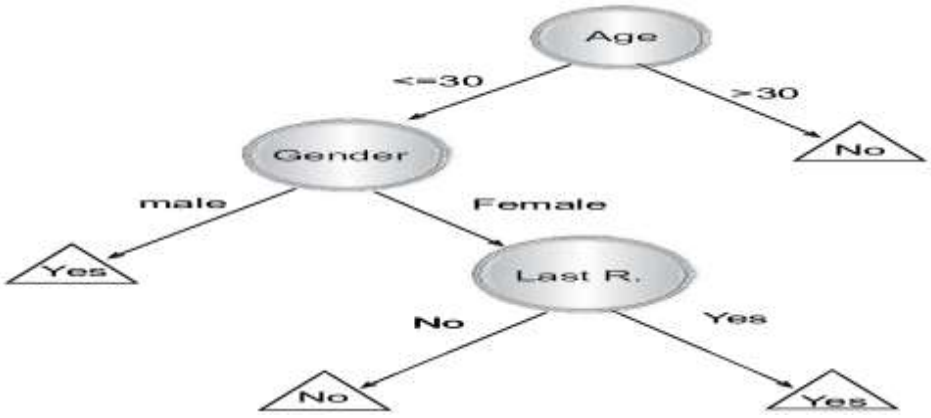



Figure 3.4. Decision Tree (Golmohammadi & Zaine, 2015)

## 3.6. Naïve Bayes

Naive Bayes algorithm is one of the algorithms contained on classification techniques. Naive Bayes classification method is a probability and statistics raised by the British scientist Thomas Bayes, which predict future opportunities based on the experience of earlier and became known as Bayes' Theorem. The naïve theorem is combined with an attribute condition where it is assumed to be independent. Naive Bayes classification is assumed that there is or is not a specific characteristic of a class has nothing to do with the characteristics of other classes. Equation of Bayes' theorem is (Leni, 2016)

$$P\ (H/X)=P(X/A).P(H)/P(X)$$

Remarks:
X=Unknown data class
H= Hypothesis data
P(H|X)=Hypothesis probability
P(H)=Prior probability
P(X|H)=Condition probability
P(X)=Probability

To explain Naive Bayes theorem, note that the classification process requires some clues to determine what classes are suitable for the samples analyzed. Therefore, the Bayes theorem above adjusted as

$$P(C/F1….Fn)= P(C)P(Ff1….Fn)/C/P(F1…Fn)$$

C represents the class, while the variable F1 ... Fn represents the characteristics of instructions needed to perform the classification. The formula is explained that the chances entry of samples of certain characteristics in the class C (Posterior). It is the chance appearance of class C (before the entry of the sample, often called priors), multiplied by the likelihood of the characteristics of the sample characteristics to the class C (also called likelihood), divided by the likelihood of global characteristics of the sample characteristics (also called evidence).
The flow of Naive Bayes method is as follows:

1. Read the training data
2. Calculate the amount and probability, but if the numerical data:
    a) Find the mean and standard deviation of each parameter is numeric data.
    b) Find probabilistic value by counting the number of the corresponding data from the same category divided by the amount of data in the category.
3. Getting the values in the table mean, standard deviation and probability.

## 3.7. Preprocessing Techniques

Preprocessing of the data in data set preparation for clustering and classification can involve data cleaning to reduce noise or handle missing values, relevance analysis to remove irrelevant or redundant attributes, and data transformation, such as generalizing the data to higher-level concepts or normalizing the data (Witten, & Frank, 2005).

The aim of data preprocessing is to clean the data set and increase the effectiveness of the data mining result. The selected data have different formats; missing values and inconsistencies. In order to clean the data, data mining techniques of filling the data with common attributes used.

## 3.8. Evaluation methods

To conduct these study both clustering and classification techniques are used. In order to cluster the data set K-means algorithm is used as it is a very good general-purpose clustering algorithm since K is easily determined. For classification three algorithm used namely J48 decision tree, MLP and NaiveBaye algorithm used in the system experimented using 10-Fold cross validation and percentage split. The model constructed is evaluated using accuracy, recall, precision and user acceptance testing feedback.

# CHAPTER FOUR

# BUSINESS AND DATA UNDERESTANDING

## 4.1. Overview

To conduct this study, the researcher selected KDD process model. KDD starts from selection of data. In this study the researcher has taken an available dataset from bank's Application log file. As discussed in literature review chapter about KDD, the researcher requires gaining relevant prior knowledge by analyzing the business process and the goals of the end user of the discovered knowledge.

The overall process of data mining for ATM eCard Fraud Detection:

- Gathering data of domain and related knowledge
- Selecting transactional log dataset based on date and time or quantity or a combination of both, customer based or customer group based.
- Preprocessing (Remove noise, handling missing value, Transformation into a suitable form for mining.
- Using Data mining technique which is usually searching patterns based on models such as classification, neural network
- Model evaluation to identify representing knowledge

## 4.2.Data Mining process model

The KDD or Knowledge Discovery Databases (Qaiser, Shafique, & Haseeb, 2014)is the process of extracting the hidden knowledge from databases. KDD requires relevant prior knowledge and brief understanding of application domain and goals. KDD process model is iterative and

interactive in nature. There are nine different steps or stages of this model and these are given below.

1. **Developing and understanding of the application:** This is the first stage of KDD process in which goals are defined from customer's view point and used to develop and understand about application domain and its prior knowledge.
2. **Creating a target database:** This is the second stage of KDD process which focuses creating on target data set and subset of data samples or variables. It is critical stage because knowledge discovery is performed on all these.
3. **Data cleaning and preprocessing:** This is the third stage of KDD process which focuses on target data cleaning and pre-processing to complete and consistent data without any noise and inconsistencies. In this stage strategies are developed for handling such noisy and inconsistent data.
4. **Data Transformation:** This is the fourth stage of KDD process which focuses on transformation of data from one form to another so that data mining algorithms can be implemented quickly. For this purpose different data reduction and transformation methods are implemented on target data.
5. **Choosing the suitable data mining task:** This is the fifth stage of KDD process in which appropriate data mining task is chosen based on particular goals that are defined in first stage. The examples of data mining method or tasks are classification, clustering, regression and summarization etc.
6. **Choosing suitable data mining Algorithm:** This is the sixth step of KDD process in which in which one or more appropriate data mining algorithms are selected for searching different patterns from data. There are many algorithms present today for data mining but appropriate algorithms are selected based on matching the overall criteria for data mining.
7. **Employing data mining algorithm:** This is the seventh step of KDD process in which selected algorithms are implemented

8. **Interpreting Mining pattern:** This is the eighth step of KDD process that focuses on interpretation and evaluate of mining patterns. This step may involve in extracted patterns visualization.
9. **Using discovered knowledge:** This is the last and final step of KDD process in which the discovered knowledge is used for different purposes. The discovered knowledge can also be used interested parties or can be integrated with another system for further action.

### 4.2.1. Developing and understanding of the application

Nowadays the number of ATM card users has increased from day to day, and ATM and POS terminals are coming to be the preferable transactional medium since ATM cards are much safer from theft than cash and since using ATM card makes life easy. Currently, in Ethiopia, the number of Bank customers is increasing, and banks are also changing the way of reaching out to customers by implementing core banking solution which helps to centralize bank's repository. At the moment, all banks have implemented CBS with the enforcement of central bank (National Bank of Ethiopia). Having CBS has made the implementation of value-added modules like Card Banking, Internet Banking, and Mobile Banking easily pluggable, although these technologies have their drawbacks. One of the risks that come with these technologies is fraud attacks. Nowadays, users of card banking are increasing from day to day. Especially, card customers are increasing because of the system implemented by Ethswitch. Ethswitch enabled a system where customers can be served by any Bank's ATMs regardless of where they are issued their card from. Since there is no way to detect fraud in the current system, the researcher has tried to review Card banking related frauds that will happen in developed countries, and to analyze the types of frauds happening since banks are unable to detect fraud transaction in the existing system. In this study, ATM application server data are going to be analyzed to develop real-time fraud detection predictive model.

As we have discussed above Data mining is a logical method which used to examine through an enormous amount of data in order to find useful information. The problem this study going to address is developing real-time ATM fraud detection model so that data mining will be applied to find patterns (that were previously unidentified) using supervised (classification) and

unsupervised (clustering) data mining tasks. Once these patterns are identified they can further be used to make definite decisions for the development of their businesses, this study, to develop a real-time predictive model to detect ATM frauds. The researcher followed the following procedure. Firstly the researcher will cluster the available data into the respective class of fraud level, and clustering helps to identify the hidden pattern existing in the data. At the end using this cluster the second step is to classify the new incoming data using the clustered data and current status of the card customer. Structured log data concerning user behavior was initially collected from one of the private banks to perform the tasks mentioned above. Since the quality of the data and the data structure is very crucial, the researcher has gone through intensive processes to ensure the availability and the quality of the data. Additionally, the researcher endeavors to come up with the best sort of data to relate the existing data with the existing problem.

### 4.2.2. Creating a target data set

This step helps to selects a subset of variables (attributes) and data points (examples) that will be used to perform discovery tasks. This step usually includes querying the existing data to select the desired subset. The application server maintains three types of log. From those logs, two of them are maintained in notepad as a text file, and the third one has CSV data format. Because of the content and the format the third type of log data is used for this study. The way log captured in application server follows ISO 8583 standards. Thus, for the purpose of conducting this study, the Bank's ATM server logs are used. Fraud detection system requires enormous amount of data source. Accordingly, to conduct the experiment, a one-year data was extracted from the bank's application server.

To understand and align the existing data with the problem at hand, analysis of the data and the data structure itself were conducted with the domain expert(Card banking team leader). Accordingly, the following data were found in the ATM application server logs. As the domain expert says the log data sets are presented following ISO 8583 standards. ISO 8583 is financial transaction card originated message standards. For instance in channel type column; 11 stands for POS, 12 stands for ATM, 14 stands for reversal of transaction. The detail of this code presented under Annex 1.

The data description and data types are illustrated in the table 4.1.

| No. | Attributes | Data Type | Descriptions |
|---|---|---|---|
| 1. | BDS | Text | Its host id(connector of DB and switch) |
| 2. | Date | Date Time | Transaction date |
| 3. | Date and Time | Date Time | Full information of date including timestamp |
| 4. | BIN | Number | Bank Identification Number |
| 5. | Terminal ID | Number | ATM or POS machine Id |
| 6. | Channel Type | Number | Chanel type ATM or POS machine |
| 7. | CIN | Number | It is Card Identification Number |
| 8. | AcctNumb | Number | Account Number of the customer |
| 9. | Transaction Type | Number | Transaction type (cash withdrawal, transfer, statement) |
| 10. | Transaction Amount | Number | Transaction amount |
| 11. | TID | Number | Transaction Identification number |
| 12. | STAN | Number | System trace audit number |
| 13. | Resp.Type | Number | Response time for a given transaction |
| 14. | Resp.cod | Number | Response code of the transaction |
| 15. | Stat Desc. | Text | Status description of a given transaction |
| 16. | Decl.Reason | Text | Decline reason of a given transaction |
| 17. | Terminal name & loctn | Text | Terminal name through with the customer served |
| 18 | Nearness to branch | Text | Near to branch (far from branch) |

Table 4.1. Data description and data type

### 4.2.3. Data Cleaning and Preprocessing

The log data generated from ATM Application server log file has some missing attributes, inconsistence and duplicate attributes. Concerning importance, all log attributes are not necessary, so the researcher tried to filter the relevant data from the vast data set. For instance, some attributes seem to duplicate like Date & Date with time stamp. Thus for this study, a date with full information selected. Under preprocessing, tasks like data cleaning, data integration, transformation and data reduction techniques are addressed.

Data cleaning and preprocessing is the third stage of KDD process which emphasizes on target data cleaning and pre-processing to complete and create consistent data. It helps to avoid any noise and inconsistencies. Data preprocessing is an often neglected but important step in data mining process., The phrase "Garbage in garbage out" is particularly applicable to data mining and machine learning, Thus, to get a remarkable result from data mining the researcher has drawn strategies to deal with such noisy and inconsistent data.

The core role of this stage is to ensure cleanness as much as possible and making available in a way suitable for a later stage of data mining. It deals with data cleaning, transformation and reduction techniques.

The big challenge of fraud detection system implementation is getting fraud transaction out of the data-set since it rarely happens with in a transaction., Thus to get a reliable result, getting a huge volume of data-sets is important. Additionally, the quality of the data is also major determinant of the result. In data mining the data cleaning task is a resource intensive task.

In this study since most of the data were not clean the researcher has applied data cleaning process which involves filling in missing values, smooth noisy data, identify and remove outliers, and resolving inconsistencies.

All columns of the data in the dataset are not relevant for this study. So for the purpose of conducting this experiment the researcher have eliminated some of the attribute like(BDS connector switch), Transaction ID, date, STAN, terminal id from the record. So, after eliminating irrelevant and unnecessary data only a total of 51,500 datasets are used from 209,803 for the purpose of conducting this study.

Data preprocessing method are divided into the following category namely, Data cleaning, Data integration, data transformation and data reduction.

#### *4.2.3.1. Data Cleaning*

The data collected from application servers are structured in CSV format. But, the information contained in the data lack some attributes like location, which need to be mapped from the terminal name. And some of the terminal names are empty, and thus it's not complete to fill the missed value of ATM terminal. The researcher used a global constant to fill in the missing value techniques which use constant value to fill the missed value which is (NN not known), because of the recommendation gained from subject matter expert, terminal name is not the only attribute for determining fraud level of a given transaction. As recommended by the domain expert the duplicates records around 100,000 records are deleted since they are duplicate and inconsistent, and the missing values of Terminal name (7,000 records are filled with common values of NN which stand for not known).

#### *4.2.3.2. Data Transformation*

In data transformation, the data are transferred into suitable forms for data mining process. In this study, the terminal name and mapped status attributes values are transformed. The original data contains some of values which hold irregular attributes like combinations of ATM location and name. And the location names are derived from the terminal name, attributes are mapped to numeric values to make them suitable for data mining and some generalization techniques are used to name the location of the terminal as shown in table 4.2.

| Terminal Name | Location |
|---|---|
| Adam Branch ATM,Hawas Branch ATM,Torbanobbo branch ATM,direint hotel ATM | Adama |
| Cinema Ethiopia ATM, Ambasador Cinema ATM,AbaNebso, Branch,Kersa Branch ATM,Finfine Branch ATM,Ambasador cinema ATM,cinema Ethiopia ATM,TiruneshDibaba Branch ATM,DHGeda ATM &etc. | Addisababa |
| Ambo Branch ATM | Ambo |
| Assela branch ATM | Assela |
| Bahirdar Branch ATM | Bahirdar |
| Horaarsade Branch ATMs | Bishoftu |
| Chiro branch ATM | Chiro |
| Dire Dawa branch ATM,SabinaATM,HafatessaATM,diredawaRashotel ATM | Diredawaa |
| Gonder branch ATM | Gonder |
| Hundene Branch ATM | Harar |
| Jimma Branch ATM, | Jimma |
| Djibruk Branch ATM,Mekele Branch ATM | Mekele |
| Nekemte Branch ATM,Leqa ATM | Nekemte |
| Sebetaayo Branch ATM | Sebeta |
| Shashemene Branch ATM, | Shashemene |
| Woliso branch ATM | Woliso |
| Hawassa branch ATM | Hawassa |

Table 4.2. Terminal Name mapped to Terminal Locations

| **Status Description** | **Mapped value** |
|---|---|
| /Batch not opened ABBTRB2 | 1 |
| /dbdRec read error ABBTRB2 | 2 |
| A/c Validation failure | 3 |
| Account status : Debit Freeze. | 4 |
| Account status : Total freezed. | 5 |
| Account status not Normal | 6 |
| ATM ID NOT FOUND IN D350048 TABLE | 7 |
| Batch Close Error | 8 |
| Batch Close Error in HO Branch | 9 |
| Batch Close Error in Cash Branch | 10 |
| Batch Close Ho in Branch | 11 |
| Credit Already Pass | 12 |
| Debit Transaction Already Found in Table | 13 |
| Duplicate 1200 Message | 14 |
| Insufficient Fund | 15 |
| No Debit Transaction Found | 16 |
| No Original Message Received | 17 |
| Record Creation In D350066 Table Failure | 18 |
| Success | 19 |
| System Error Unable to Load m001994.dll | 20 |

Table 4.3. Mapped status Description

## 4.3. Format conversion

WEKA (Waikato) software was used. The datasets provided to WEKA were prepared in a format that is acceptable by WEKA software. The software accepts records whose attribute values are separated by commas and saved in an ARFF (Attribute Relation File Format).

After preprocessing was done, excel file was first changed into a comma delimited (CSV) file format. Then the file was saved with ARFF (Attribute Relation File Format) file extension. Finally, the dataset in ARFF format is ready to be used in the WEKA software. The declaration of the ARFF format is presented in Annex 2 and Annex3.

# CHAPTER Five

# EXPERMENTATION

## OVERVIEW

In this chapter the researcher experimented the preprocessed data using several data mining algorithms. For conducting this study Data mining unsupervised clustering technique (K-means algorithm) applied. In the second step supervised classification algorithm namely Multilayer perceptron, J48 Decision Tree j48 and NaiveBayes are experimented. Since the ATM card transactions dataset used for the study does not have a characteristic indicating whether a transaction was a suspicious fraudulent transaction or suspicious non-fraudulent transaction clustering applied to label the data set. One of the challenge in conducting this study is, that the application of data mining technique like classification relies on accurate identification of fraudulent and non-fraudulent transactions, however these information usually do not exists or limited or confidential. These happen because of financial institutions prefer to not disclose this kind of information and categorize them in high-risk data, so accessing to this kind of data is very restricted. Having that, the researcher planned to conduct experiments using clustering technique of k-means which is used to group credit card transactions that are suspected to be fraudulent, into a similar cluster and the non fraud to the other cluster then the clustering result interpreted with the help of the domain expert. The output of this model is then used to train MLP(Multi layer perceptron), NaiveBayes and j48 decision tree which can then be used to Build the classification model to classify incoming transactions. So, this chapter discus steps and procedures conducted during the experimentation and evaluation of the model generated by those data mining techniques.

## 5.1. Experimental design

The experiment performed using computer having hardware specification of Intel(R) core (TM) i3-2310M CPU @2.10GHz 2.10 GHz, 3GB RAM and the operating system is window 7. In order to manipulate this data WEKA has been used as machine learning algorithms for data mining tasks. Weka supports several standard data mining tasks, more specifically, data preprocessing, clustering, classification, regression, visualization, and feature selection. WEKA contains many inbuilt algorithms for data mining and machine learning. It is open source and freely available platform-independent software. All of Weka's techniques are predicated on the assumption that the data is available as one flat file or relation, where each data point is described by a fixed number of attributes (normally, numeric or nominal attribute.

Microsoft Excel is used to manipulate the data and to populate the clustering result to the entire data set, In addition to the above mentioned software Notepad ++ also used to easily manage and truck the error generated on the process of feeding the available data into Weka.

The Following steps are followed in order to conduct this Experiment

- First the researcher selected Weka tool for mining the data and the software successfully installed on the specified computer.
- The data adjusted to weka acceptable data format which is ARFF. Iin order to perform this task the data is first converted to csv format then the data defined according to arff format, then the csv data opened using word pad and copied to the ARFF format.
- Data Preprocessing performed to fill missing values and to map some of the record values to numeric.
- Before the data pulled to the classifier the researcher need to label the available data so clustering conducted using K-means algorithm.
- The clustering result is used to label the data for classification
- The ARF and Labeled data is Feed into Weka classifier

## 5.2. Weka interface

Weka is an open source software for data mining under the GNU General public license. This system is developed at the University of Waikato in New Zealand. “Weka” stands for the Waikato Environment for knowledge analysis.

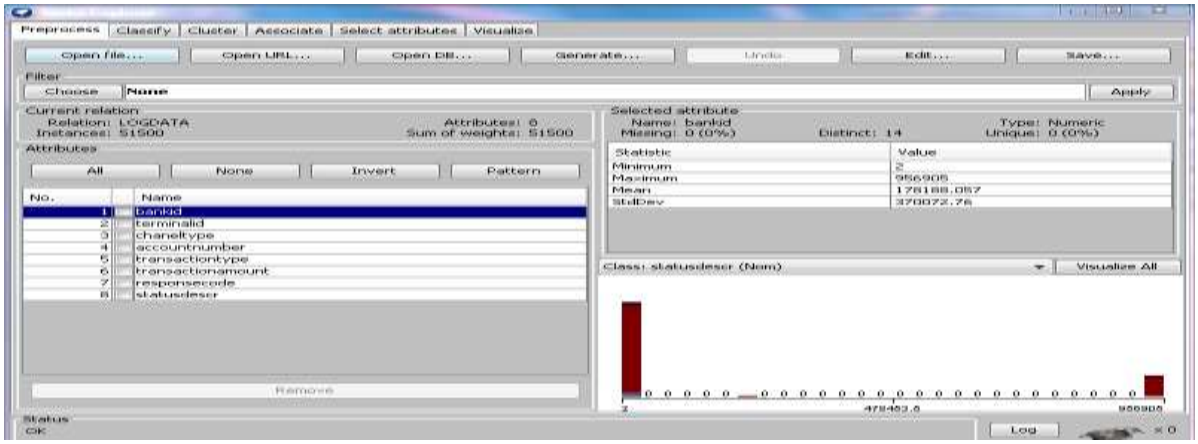



Figure 5.1. Weka Interface after data set fetched into it

## 5.3. Modeling using clustering technique

Clustering is the process in which we divide the available data instances into a given number of sub-groups. These sub-groups are called clusters, and hence the name “Clustering”. In the context of understanding data, clusters are potential classes and cluster analysis is the study of techniques for automatically finding classes. To train the classifier the researcher applied the clustering technique of K-Means algorithm, K-means is a prototype based partitioned clustering techniques attempt to find user specified (K) number of clusters which are represented by their centroid. The K-means algorithm outlines a method to cluster a particular set of instances into K different clusters, where K is a positive integer. It should be noted here that the K-means clustering algorithm requires knowing the number of clusters from the user. It cannot identify the number of clusters by itself.

The K-means clustering algorithm starts by placing K centroid as far away from each other as possible within the available space. Then each of the available data instances is assigned a particular centroid, depending on a metric like Euclidian distance and Manhattan distance.

The position of the centroid is recalculated every time an instance is added to the cluster and this continues until all the instances are grouped into the final required number of clusters. Since recalculating the cluster centroids may alter the cluster membership, the cluster memberships are also verified once the position of the centroid changes. This process continues till there is no further change in the cluster membership, and there is as little change in the positions of the centroids as possible. The initial position of the centroids is thus very important since this position affects all the future steps in the K-means clustering algorithm.

In this study a thorough analysis and interpretation of clusters is made by the researcher and the domain experts

**Experiment 1**

For this study the value of K set to 2 which are the default values of the expected result fraud or non-fraud, after the clustering applied the distribution of the clustered instances showed in table, the percentage distribution of the clustered instances from table 5.1: cluster 0 (9456), cluster 1(42044), Incorrectly clustered instances : 12454.0 24.1825 %.

Cluster 0 <-- 51

Cluster 1 <-- 0

| **Clustered as** | **Cluster0** | **Cluster1** | **Sum** |
|---|---|---|---|
| 0 | 8104 | 38477 | 46581 |
| 5 | 640 | 263 | 903 |
| 51 | 1245 | 640 | 1814 |
| 78 | 374 | 869 | 1243 |
| 94 | 17 | 205 | 222 |
| 96 | 129 | 608 | 737 |

Table 5.1. Result of Experiment

Cluster 0 --→ 51, cluster 1-→0, 0, 51, 78, 94, and 96 are status /response code

The result obtained from descriptive modeling seen by domain expert and interpreted as follows: Cluster 0 stands for response code of 51, that means transaction not completed success fully because there are one of the following problems; Application related issues, balance issues, number of bad login issues, so that is highly suspected to be a fraudulent transaction. Status 0 stands for successfully committed transaction without Application related issues, balance issues, and any other chance of occurrence of fraudulent transaction, the domain expert stated that this transaction is less expected to be fraudulent. As we can see status code 5 and 51 is more grouped under cluster 0 both code indicate that there is application related or balance related issues in that account and more of the status code 0 is grouped under cluster one as the suggestion given by the expert the transaction grouped under cluster 0 is clustered as Fraud and cluster 1 clustered as non fraud. Finally the descriptive modeling experimentation refined based on the sound comments and suggestions of the domain experts. The new data set labeled based on the centroid values generated from k-means for the predictive model building.

For classification modeling experiments, decision tree (in particular the J48 algorithm), Multiplayer perceptron algorithms and NaiveBaye will **explored** in the semi – supervised modeling.

## 5.4. Modeling using classification techniques

Classification is the process of building a model of classes from a set of records that contain class labels (Gaganjot Kaur, 2014) , Modeling is a crucial tasks performed on data mining, as the objective of the study shows the ultimate goal of this study is to develop a predictive model which helps to detect ATM frauds.

Once the clustering model obtained the next steps of this study is developing the predictive model. Before proceeding into developing the predictive model the data set need to be labeled according to the clustering result obtained and the result which have high intra-class similarity and low intra-class similarity selected then the data set will be experimented using two data

mining classification algorithm those are Decision Tree J48 and Multilayer perceptron of Neural network model

### 5.4.1.J48 Decision tree

Decision Trees model has two goals: producing an accurate classifier and understanding the predictive structure of the problem (Petri, 2010). Decision trees are a class of data mining techniques that have roots in traditional statistical disciplines such as linear regression. Decision trees also share roots in the same field of cognitive science that produced neural networks.

In specific to this study J48 is used, which is an extension of ID3. The additional features of J48 are accounting for missing values, decision trees pruning, continuous attribute value ranges, derivation of rules, etc. In the WEKA data mining tool, J48 is an open source Java implementation of the C4.5 algorithm

J48 algorithm has some parameters that can be changed to further improve classification accuracy. Initially the classification model is built with the default parameter values of the J48 algorithm

The WEKA tool provides a number of options associated with tree pruning. In case of potential over fitting pruning can be used as a tool for précising. In other algorithms the classification is performed recursively till every single leaf is pure, that is the classification of the data should be as perfect as possible. This algorithm it generates the rules from which particular identity of that data is generated. The objective is progressively generalization of a decision tree until it gains equilibrium of flexibility and accuracy.

A decision tree is a predictive machine-learning model that decides the target value (dependent variable) of a new sample based on various attribute values of the available data. The internal nodes of a decision tree denote the different attributes; the branches between the nodes tell us the possible values that these attributes can have in the observed samples, while the terminal nodes tell us the final value (classification) of the dependent variable.

**Experiment 2 (Using J48 decision tree) with 10-fold cross validation**

The J48 Decision tree classifier follows the following simple algorithm. In order to classify a new item, it first needs to create a decision tree based on the attribute values of the available training data. So, whenever it encounters a set of items (training set) it identifies the attribute that discriminates the various instances most clearly. This feature that is able to tell us most about the data instances so that we can classify them the best is said to have the highest information gain. Now, among the possible values of this feature, if there is any value for which there is no ambiguity, that is, for which the data instances falling within its category have the same value for the target variable, then we terminate that branch and assign to it the target value that we have obtained.

For the other cases, we then look for another attribute that gives us the highest information gain. Hence we continue in this manner until we either get a clear decision of what combination of attributes gives us a particular target value, or we run out of attributes. In the event that we run out of attributes, or if we cannot get an unambiguous result from the available information, we assign this branch a target value that the majority of the items under this branch possess.

Now that we have the decision tree, we follow the order of attribute selection as we have obtained for the tree. By checking all the respective attributes and their values with those seen in the decision tree model, we can assign or predict the target value of this new instance.

The first experimentation is performed with the default parameters. The default 10-fold cross validation test option is employed for training the classification model.

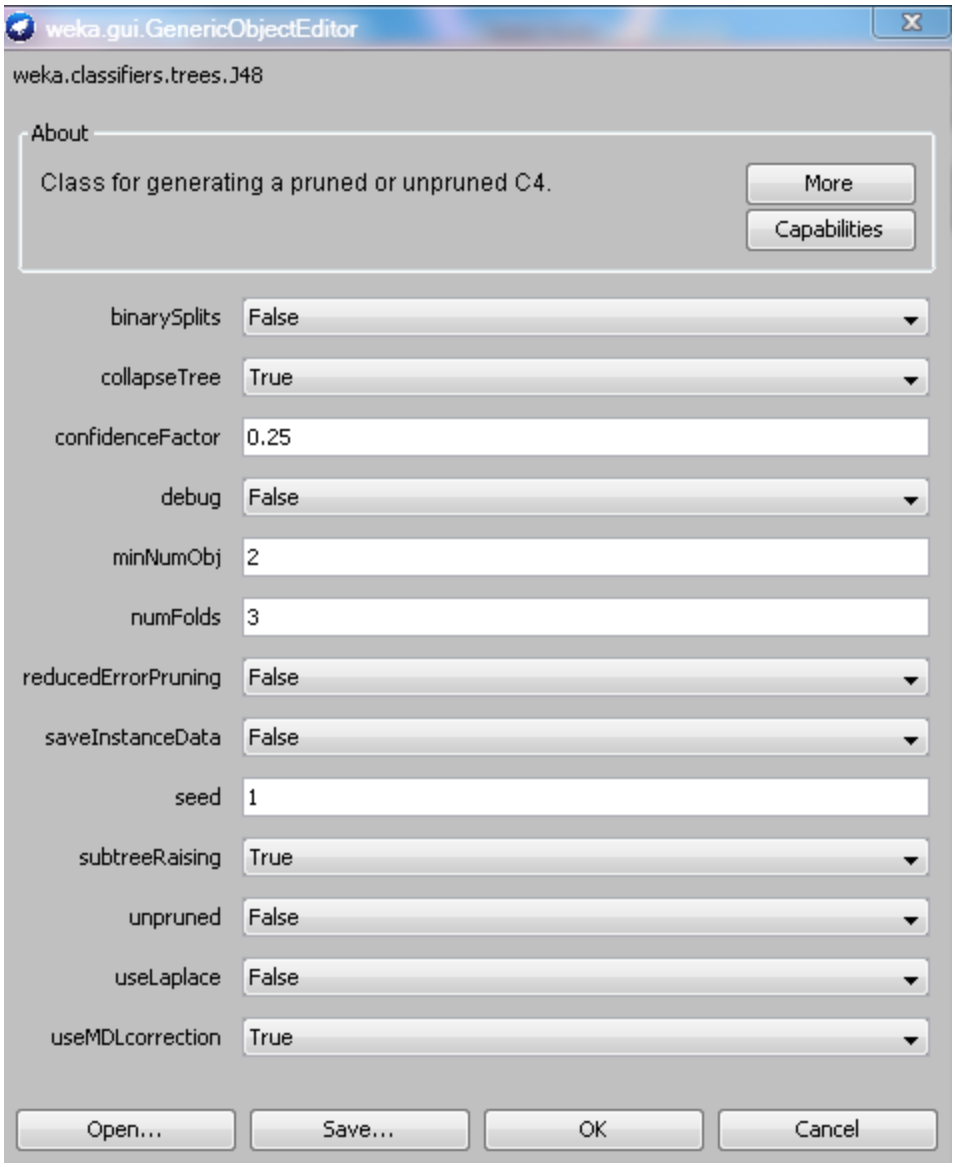



Figure 5.2. The Default parameter setting of J48 Algorithm in weka

Using these default parameters the classification model is developed with a J48 decision tree having 32 numbers of leaves and 63 tree size with result depicted in table 5.2.

| The Number of Total instance used | Correctly classified | Incorrectly classified |
|---|---|---|
| 51500 | 46594(90.473%) | 4906(9.5262) |

Table 5.2. Experimentation using J48 and 10- fold cross validation

As shown in the resulting confusion matrix, the J48 learning algorithm scored an accuracy of 90.473%. This result shows that out of the total datasets 46594 **(**90.473**%)** records are correctly classified, while 4906 **(**9.5262**%)** of the records are incorrectly classified.

**Experiment 3 Using J48 decision tree with percentage split.**

The second of J48 experiment is conducted, by changing the default testing parameter (the 10-fold cross validation). In this learning method a percentage split is used to partition the dataset into training and testing data set. The purpose of using this parameter was to evaluate the performance of the learning scheme by increasing the proportion of testing dataset if it could achieve better classification accuracy than the first experimentation. Even if different experiments were tested using percentage splitting mechanisms the one with 70 % training and 30 % testing has scored better classification accuracy. The result of this learning scheme is summarized and presented in Table 6

| Total number of instance for testing | Correctly Classified instance | Incorrectly classified instance |
|---|---|---|
| 15450 | 14030(90.8091%) | 1420(9.1909) 9.1909 |

Table 5.3. Using J48 decision tree with percentage split

The size of the tree and the number of leaves produced from this training were 32 and 63 respectively.

In this experiment out of the 51,500 total records 36,050 (70%) of the records are used for training purpose while 15,450 (30%) of the records are used for testing purpose. As we can see from the confusion matrix of the model developed with this proportion, out of the 15,450 testing records 91.1% of them are correctly classified and 8.9% records are incorrectly classified. Resulting confusion matrix is presented in Appendix D.

### 5.4.2. Artificial Neural Network Modeling (Using MLP )

Artificial Neural Network is a field in Artificial Intelligence (AI) where we use data structures and algorithms for learning and classification of data, inspired by Human brain working element of neural system.

For the purpose of conducting this experiment multilayer perceptron (MLPs) is selected it is one of the Logistic Regression, MLP is feed-forward architecture which uses inner nodes. The MLP model not only gives a perceptron structure for representing, it also defines a learning rule for this kind of network. The MLP is divided into three layers: the input layer, the hidden layer and the output layer, where each layer in this order gives the input to the next.

**Experiment 4. MLP with 10-fold cross validation**

After all necessary pre-processing is done; the neural network model experimentation is tested by changing the values of the hidden layer of Multilayer perceptron algorithm with 10-fold.Table 5.3 shows the result of Multilayer perceptron algorithm with 10-fold cross validation of the model.

| Total number of instance for testing | Correctly Classified instance | Incorrectly classified instance |
|---|---|---|
| 51500 | 46581(90.4485%) | 4919(9.5515%) |

Table 5.4.Experimentation using MLP and cross validation using 10 folds

**Experiment 5 MLP with percentage split**

The second of MLP experiment is conducted, by changing the default testing parameter (the 10-fold cross validation). In this learning method a percentage split is used to partition the dataset into training and testing data set. The purpose of using this parameter was to evaluate the

performance of the learning scheme by increasing the proportion of testing dataset if it could achieve better classification accuracy than the first experimentation. Even if different experiments were tested using percentage splitting mechanisms the one with 70 % training and 30 % testing has scored better classification accuracy. The result of this learning scheme is summarized and presented in Table 5.5.

| Total number of instance for testing | Correctly Classified instance | Incorrectly classified instance |
|---|---|---|
| 15,450 | 13477 (87.2298%) | 1973 (12.7702%) |

Table 5.5.Experimentation using MLP and with 70% data set to training and 30% for testing

### 5.4.3. Experimentation using Naïve Bayes

The Naive Bayes classification algorithm is a probabilistic classifier. It is based on probability models that incorporate strong independence assumptions. The independence assumptions often do not have an impact on reality. Therefore they are considered as naïve [ (Center)].

The Naive Bayes classification algorithm is a probabilistic classifier. It is based on probability models that incorporate strong independence assumptions; the independence assumptions often do not have an impact on reality. Therefore they are considered as naïve.

**Experiment 6. Naïve Bayes with 10-fold cross validation**

The first Experiment is conducted with the default parameter and cross validation with 10 fold the result gained from this experiment are stated as follow.

| Total number of instance for testing | Correctly Classified instance | Incorrectly classified instance |
|---|---|---|
| 51500 | 44717(86.8291%) | 4919(13.1709%) |

Table 5.6. Experimentation NaiveBayes and with 70% data set to training and 30% for testing

The above result shows us Naïve bayes Algorithm experimented with cross validation of 10 folds 86.8291 accuracy achieved which is the lowest accuracy percentage out of the experiment conducted previously.

### Experiment 7.NaiveBayes with percentage split

The second of NaiveBayes experiment is conducted, by changing the default testing parameter (the 10-fold cross validation). In this experiment percentage split is used to partition the dataset into training and testing data set the training set given 70 % of the data set and the testing set given 30 % of the data set so accordingly the result obtained from weka described below under table 10

| Total number of instance for testing | Correctly Classified instance | Incorrectly classified instance |
|---|---|---|
| 15,450 | 13477 (87.2298%) | 1973 (12.7702%) |

Table 5.7. Experimentation NaiveBayes and 70% data set to training and 30% for testing

The second experimentation which tested using percentage split of 70% for training and 30 % of testing achieved 87.2298% of accuracy it seems better when comparing to the first experiment but it's still too low when comparing to the above experiment which conducted using decision tree of j48 and Artificial Neural Network of MPL.

### Selecting and Comparison of K-Means with J48 and K-Means with MPL

Selecting a better classification technique for building a model, selecting the one which performs best in handling the prediction and identifying the ATM Transaction frauds are one of the aims of this study. For that reason, two classification models are selected. For each algorithm the best performance accuracy are listed in table 5.8 and table 5.9 below.

| Classification Algorithm | Test Model | Accuracy | |
|---|---|---|---|
| | | Correctly Classified | Incorrectly Classified |
| Decision Tree(J48) | 10-Fold cross validation | 46710(90.699%) | 4790(9.301) |
| | Percentage split(70/30) | 14030(90.8091%) | 1420(9.1909) |
| Artificial Neural Network (MLP) | 10-Fold cross validation | 46581(90.4485%) | 4919(9.5515%) |
| | Percentage split(70/30) | 13477 (87.2298%) | 1973 (12.7702%) |
| NaiveBayes | 10-Fold cross validation | 44717(86.8291%) | 4919(13.1709%) |
| | Percentage split(70/30) | 13477 (87.2298%) | 1973 (12.7702%) |

Table 5.8. Comparison of Experimented classifiers

As mentioned above one of the basic aims of data mining is to compare different models with respect to developing predictive model and to select the better classification accuracy accordingly. Therefore, detailed experimentation for different models has been conducted. As a result, the best classification algorithm which is appropriate for this problem domain has been selected.

The detailed classification accuracy and confusion matrix for each algorithm conducted in this thesis were presented in the Appendix.

## 5.5. Evaluation of the Discovered Knowledge

Under this section the result obtained from those conducted experiment evaluated and the best one chosen (the one with the high accuracy result) is selected to develop the model.

As it has been discussed above the experiment conducted with percentage split has a better result rather that the cross validation and K-means with J48 decision tree algorithm selected with the accuracy of 90.8% selected so having that rules are generated from the model.

Rules generated from the predictive model are described as follows:.

Rule 1. if channel Type > 1210(ATM machine) and Transaction type <= 100 and transaction amount <= 15555 and terminal id <=2163 the transaction is fraud. Then the transaction is not fraud, the above rule gives correct result for 18.0 of 1. Cases it covers.

Rule 2. if channel Type > 1210(ATM machine) and Transaction type <= 100 and channel type> 1230 and bank id >5 and distance from branch is NTB(Near to branch) and terminal id<=215054 and terminal id <=90001 and transaction amount >350 and transation amount<=1500 and location = addis ababa Then the transaction is fraud, the above rule gives correct result for 1 of 0. cases it covers

Rule 3.if channel Type > 1210(ATM machine) and Transaction type <= 100 and transaction amount > 1555 the transaction is not fraud. Then the transaction is not fraud, the above rule gives correct result for 12.0 of 1. cases it covers.

Rule 4.If channel Type <= 1210 Then the transaction is not fraud, the above rule gives correct result for 28613.0 of 3957.0 cases it covers. The transaction passed through channel type of 1210 (POS) is not fraud

Rule 5. if channel Type > 1210(ATM machine) and Transaction type <= 100 and channel type<= 1230 Then the transaction is not fraud, the above rule gives correct result for 22103 of 864. cases it covers

Rule 6. if channel Type > 1210(ATM machine) and Transaction type <= 100 and channel type> 1230 and bank id >6 Then the transaction is not fraud, the above rule gives correct result for 138 of 0. cases it covers

Rule 7. if channel Type > 1210(ATM machine) and Transaction type <= 100 and channel type> 1230 and bank id <=5 Then the transaction is not fraud, the above rule gives correct result for 112 of 2. cases it covers

Rule 8. if channel Type > 1210(ATM machine) and Transaction type <= 100 and channel type> 1230 and bank id >5 and distance from branch is NTB(Near to branch) and terminal id<=215054 and terminal id >90001 Then the transaction is not fraud, the above rule gives correct result for 135 of 3. cases it covers

Rule 9. if channel Type > 1210(ATM machine) and Transaction type <= 100 and channel type> 1230 and bank id >5 and distance from branch is NTB(Near to branch) and terminal id<=215054 and terminal id <=90001 and transaction amount <=350 Then the transaction is not fraud, the above rule gives correct result for 105 of 6. cases it covers

Rule 10. if channel Type > 1210(ATM machine) and Transaction type <= 100 and channel type> 1230 and bank id >5 and distance from branch is NTB(Near to branch) and terminal id<=215054 and terminal id <=90001 and transaction amount >350 and transation amount>1500 Then the transaction is not fraud, the above rule gives correct result for 17 of 1. cases it covers

Rule 11. if channel Type > 1210(ATM machine) and Transaction type <= 100 and channel type> 1230 and bank id >5 and distance from branch is NTB(Near to branch) and terminal id<=215054 and terminal id <=90001 and transaction amount >350 and transation amount<=1500 and location = adama Then the transaction is not fraud, the above rule gives correct result for 8 of 1. cases it covers

Using the generated rule programs developed using Microsoft Visual Studio 2008(For I Application Side) and Microsoft SQL 2008 to store the data set it is assumed as this software is going to be deployed on Card Banking Application Server and the data Evaluated before Dispensing the Money to the customer.

## 5.6. User Interface Design

The system will take sample data set from the testing data and investigate the transaction against the model to validate it as a fraud or non Fraud Transaction, it was developed using C# code and SQL Data server as persistence data layer. For real environment application it can be installed on the ATM application server and triggered when new transaction called.

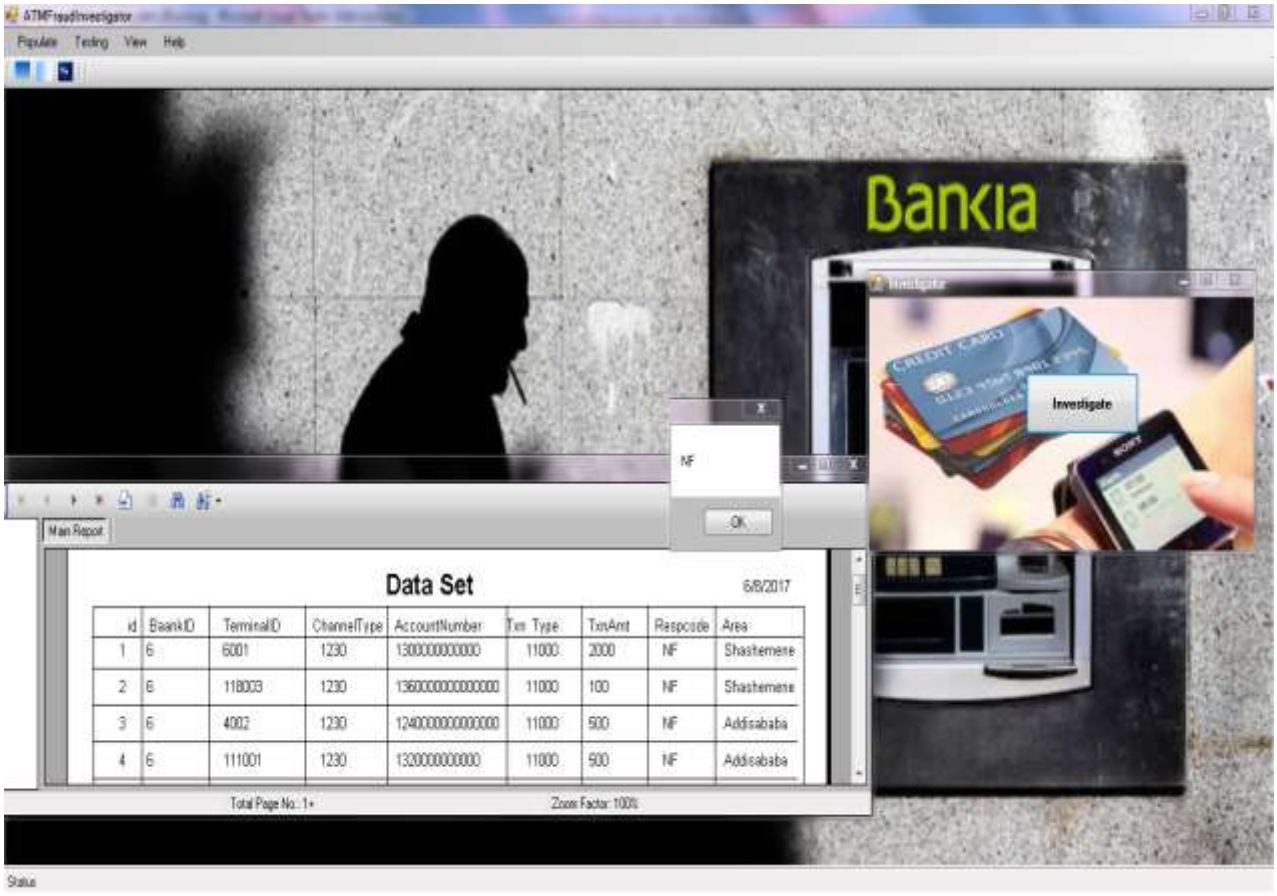



Figure 5.3. User Interface of Fraud detection System

## 5.7. User acceptance testing Result

At the end of this study Alpha testing methods used to test users expectation toward the result of this study. **Alpha testing** is used when UAT is carried out in the software builder's environment and is more significant in the context of commercial off the shelf software.

**The result obtained is summarized in the following table:**

| UAT | Strongly Agree | Agree | Neutral | disagree | Strongly disagree |
|---|---|---|---|---|---|
| 1. Is the System easily accessible | 8 | 2 | 0 | 0 | 0 |
| 2. Is the Rule generated by the system important for detecting frauds | 6 | 3 | 1 | 0 | 0 |
| 3. Is the attribute selected by the Data mining software relevant for detecting real time ATM card frauds | 7 | 2 | 1 | 0 | 0 |
| 4. The system able to detect different kind of frauds | 6 | 3 | 1 | 0 | 0 |
| 5. The system designed is easily applicable to the real life environment | 5 | 3 | 1 | 1 | 0 |
| 6. The system maintains data set in persistent manner | 9 | 1 | 0 | 0 | 0 |

Table 5.9 UAT summary report

# CHAPTER SIX

# CONCLUSION AND RECOMENDATION

## 6.1. CONCLUSION

Banking and Finance sectors are now becoming more sophisticated; almost all banks have already implemented CBS (Core Banking Solutions) with the advancement of such technologies enabling services like Mobile Banking. Internet Banking and Card Banking are getting easier and those enablement have also simplified almost all operation performed in the Banks. Nowadays going to branch for getting Banking services are unnecessary because already its assessable through E-Banking services , you can Withdraw, make payment, Transfer money through this Technology but less attention is given to security aspect. Because of this analyzing the security gap and developing predictive model to detect Fraud pattern was the central aim of this thesis.

In this study the researchers has identified the gap and refer many literature to fill this gap in respect to ATM Fraud attack. For conducting the study KDD data mining process model was followed. The researcher has assessed different Data mining techniques like K-means of Clustering, as well as classification algorithm such as Multilayer perceptron (MLP), Decision Tree using J48 Algorithm, Naive Bayes.
Sample data were collected from the bank application servers. In order to preprocess the data Microsoft Excel2007 and notepad++ used. After the preprocessing steps completed the data fetched into Weka software to analyze the pattern exist within the data and to develop the predictive model.

In modeling since there is no label to indicate whether a transaction is Fraud or not clustering method used in order to define the classes with in the data set Simple K-means adopted then the data set labeled accordingly, in developing the Classification model three algorithms are tested. The study was conducted first by default parameter setting and Cross validation with 10-fold; then the second experimentation done using percentage splitting, which was 70% of the data set for training and 30% of the data set for testing the one with percentage split have achieved good result in general.

The result from J48 and multilayer is still promising but to propose one model as the best for this study, the model from J48 algorithm is selected with 91% of accuracy. The model created by J 48 converted into if else statements for making the model understandable to the code developed by C# programming language.

In conclusion the result obtained by using Data mining techniques by analyze the Fraud pattern from the history log data and developing data dependent Real time fraud detection models is really achievable. However the volume of the data set and the quality of the data are still challenging tasks for developing the model.

## 6.2. RECOMMENDATION

This Research is conducted for academic purpose. The researcher advice to carry out this study in project form and deploy it on Ethswitch ATM Application servers. In addition to this the Researcher suggests the following recommendation for further research.

- ✓ While conducting this research the researcher have seen some data quality and completeness problem so the researcher advised to push banks through NBE or Ethswitch to enter all information like location and terminal name.
- ✓ Currently information like distance are not included for the future study the researcher recommend to have GPS system to analyze the distance between two ATM and with this it become possible to detect fraud like stolen card easily.
- ✓ As described in the above section the researcher have applied purposive sampling to select the data source, as noticed in this study one of the most important challenge is, using supervised data mining technique like learning machine or classification relies on accurate identification of fraudulent and non-fraudulent transactions so that for the future work the researcher advice to use large sample data from all banks since the size of the data increase the accuracy of the model.
- ✓ In today's competitive environment Finance organizations are not willing to share fraud history , so the extracted knowledge might not cover all fraud scenarios and since it increases the error and decrease the accuracy and finally affect the performance the

researcher recommend Banks to share this information for the purpose of eradicating this complex type of frauds.

- ✓ The very challenge of this study where getting literature written for conducting such types of problem, the issues was very important and urgent we can say that there is no study conducted in this area so in this thesis the researcher would like to call Researcher to dig this concealed problems and suffer our country from Fraud attack such as e-Banking frauds like ATM card frauds, Mobile Banking frauds and Internet Banking frauds.

## Bibliography

Adane, T. (2011). *Mining inurance data for fraud detection.* Addis Ababa University, Addis Ababa, Ethiopia.

Akal, T. D. (June 2012). *Constructing predictive model for Network Intrusion detection.* Addis Ababa, Ethiopia.

Anita, & Ravindra. (2013). Data mining techniques for fraud detection. *International journal of computerscience and Information technology volume4* .

Azeem Ush, N. A. (2014). Real-Time Credit-Card Fraud Detection using Artificial Neural Network. *Proc. of Int. Conf. on Recent Trends in Information, Telecommunication and Computing, ITC* .

Bramer, M. (2007). *Principle of data mining.*

Center, I. K. (n.d.). *IBM Knowledge center page*. Retrieved 5 1, 2017, from IBM web site: https://www.ibm.com/support/knowledgecenter/en/SSEPGG_10.1.0/com.ibm.im.overview.doc/c_naive_bayes_classification.html

Chauhan, H., & Chauhan, A. (2013). Implementation of decision tree algorithm c4.5. *International journal of scientific and research publications* .

Cios, Swiniarsiki, & Kuargan. (2007). A knowledge Discovery Approach. *ISBN:978-0-387-33333-5* .

Dagne, T. A. (2011). Web use patterns Discovery:The casi of Data Mining in Addis Ababa University .

Daniel, M. (2013). *Application of Datamining Technology to support to support fraud protection.* Addis Ababa: Addis Abeba Univeristy.

Do, C. B., & Batzoglou, S. (2008). What is the expectation of Maximization algorithm . *Primer Nature Publishing* .

Falaki, S. A. (2014). An Update Research On Credit Card On-Line Transactions. *Federal University of Technology* .

Fayyad, P.-S. G. (1996). Advances in Knowledge Discovery and Data Mining,. *AAAI Press, Cambridge* .

Gaganjot Kaur, A. C. (2014). Improved J48 Classification Algorithm for the Prediction of Diabetes. *International Journal of Computer Applications (0975 – 8887)* , Volume 98 – No.22.

Gebremeskel, K. (2013). APPLICATION OF DATA MINING TECHNIQUES TO PREDICT CUSTOMERS'CHURN. *A Thesis Submitted to the School of Graduate Studies of Addis Ababa University* , 17.

Gheware, S. (2014). Data Mining Tasks ,tools, techniques and Applications. *Internatinal Journal of Advanced Research in computer and Communication Enginering , 3* (10).

Golmohammadi, K., & Zaine, O. (2015). *Data mining Applicationg for fraud detection in securities market.* Canada: University of Alberta.

Ibrahim, Amit, & Barron. (2010). Fraud detection and control on ATM machines. *ReaserchGate* .

Jiawei, M. J. (2011). *Data mining concepts and techniques.*

kamber, j. H. (2006). *Data Mining Concepts and techniques.* San Fransisco: Dian cerraa.

Kanungo, T. (2002). An efficient K-means Algorithm :Analysis and Implementation. *IEEE Transactions on pattern analysis and Machine Intellegence , 24* (7).

Karlik, B., & Yuksek, K. (2007). Fuzzy clustering neural network for Real-time odor recognition. *Journal of automated methods and management in chemistry* .

Kovalerchuk, B. (2014). *Data mining for financial applications.* Central washington univeristy, USA.

Kumar, A. (2011). *Artificial neural network for data mining.* Pusa New Delhi: I.A.S.R.I.,Library Avenue.

Leni, M. A. (2016). Data mining classification comparision. *International jourgal of Engineering trends and Technology* .

Linda Delamaire (UK), H. A. (2009). Credit card fraud and detection techniques:. *Banks and Bank Systems, Volume 4, Issue 2* .

Maik, T. K. (2014). Automatic ATM Fraud Detection as a. *Wincor Nixdorf International GmbH, DE R&D ACT 53, Paderborn, Germany* .

Mamo, D. (2013). Application of Data Mining to support Fraud detection. *A Thesis Submitted to the School of Graduate Studies of Addis Ababa University in Partial Fulfillment of the Requirement for the Degree of Master of Science in Information Science BY DANIEL MAMO January* .

Mohamadi, M., Muda, Z., & Yasin, W. A hybrid framework based on neural network MLP and K-means clustering for intrusion detection system. *Proceedings of the 4th International conference on computing and informatics.* Malaysia.

Neelamadhab Padhy, D. P. (2012). The survey of Data mining Application and feature scope. *International Journal of Computer Science, Engineering and Information Technology (IJCSEIT), Vol.2, No.3* .

Oguweleka, F. N. (2011). Data mining application in credit card fraud detection system. *Journal of Engineering Science and Technology* , 311-322.

Ombati, M. N. (2010). E-Commerce Products and Services in The Banking Industry: The Adoption and Usage in Commercial Banks in Kenya. *Journal of Electronic Banking Systems, IBIMA Business Review; Vol 1 pgs 70-83 Article ID 678961* .

Petri, C. (2010). Decision Tree. *Cluj Napoca* .

Qaiser, U. S. (2014). A Comparative Study of Data Mining Process Models (KDD, CRISP-DM and SEMMA). *International Journal of Innovation and Scientific Research* , 217-222.

R. Gayathri, A. M. (2013). Investigation of Data Mining Techniques in Fraud Detection: Credit Card. *International Journal of Computer Applications (0975 – 8887)* .

Raghavendra Patidar, L. S. (2011). Credit Card Fraud Detection Using Neural Network. *International Journal of Soft Computing and Engineering (IJSCE)* .

Rajpreetl, D. R. (2013). Comparitive Study of Online Banking Security System of various Banks in India. *International Journal of Engineering, Business and Enterprise* .

Raval, U., & Jani, C. (2015). Implementation and Improvisation of K-means clustering. *IJCSMS Volume4* .

Reganie, B. (2013). *Customer segmentation and prediction: the case of busa gonofa michro finance.* Addis Ababa: AAU, Website.

S. O. Falaki1, B. K. (2012). Probabilistic Credit Card Fraud Detection System in Online Transactions. *International Journal of Software Engineering and Its Applications* .

Santhosh Baboo, N. P. (2015). Analysis of Spending Pattern on Credit Card Fraud Detection. *IOSR Journal of Computer Engineering (IOSR-JCE)* .

Sowjanya, & Jysotsna. (2013). Application of data mining for Financial accounting fraud detection. *IJARCSSE* .

Tayal, B. B. (2009). *Data Mining and Data warehousing.* Banglore, India: University Science press.

Tej paul, V. &. (2003). Understanding Credit Card Frauds. *Card business review Tata Consultancy Services* .

Tesfaye, M. (2013). Application of Data mining in credit risk assesment. *AAU website* .

Vajk Istvan, I. R. (2006). Frequent Pattern Mining in web log data . *Adaptive polytechinical , 3* (1).

Wadsworth, J. E. (1944). The Bank of England. *Cabridge Univeristy press* .

Worku, G. (2009). *Electronic Banking in Ethiopia : practices, challenges and opportunities.* Gondor Univeristy, Gondor, Ethiopia.

Xie, J., Jiang, S., Xie, W., & Gao, X. (2011). An efficient global K-means clustering algorithm. *Journal of computers volume6* .

Yirdaw, K. T. (2013). Information Security Managment Framework for Banking Industry. *Addis Abeba Univeristy Web site* , 20.

Yongjian, F. (2012). Data mining: Tasks, Techniques and applications. Rolla, Univeristy of Missouria, China.

Zeineba, D. (2005). The Impact of E-commerce on Developed and Developing Countries. *International conference of globalization.*

# Annex

## 1. ISO 8583 Financial Transaction code

**ISO 8583** *Financial transaction card originated messages — Interchange message specifications* is the International Organization for Standardization standard for systems that exchange electronic transactions made by cardholders using payment cards. It has three parts:

A card-based transaction typically travels from a transaction acquiring device, such as a point-of-sale terminal or an automated teller machine (ATM), through a series of networks, to a card issuing system for authorization against the card holder's account. The transaction data contains information derived from the card (e.g., the account number), the terminal (e.g., the merchant number), the transaction (e.g., the amount), together with other data which may be generated dynamically or added by intervening systems. The card issuing system will either authorize or decline the transaction and generate a response message which must be delivered back to the terminal within a predefined time period.

ISO 8583 defines a message format and a communication flow so that different systems can exchange these transaction requests and responses. The vast majority of transactions made at ATMs use ISO 8583 at some point in the communication chain, as do transactions made when a customer uses a card to make a payment in a store (EFTPOS). In particular, both the MasterCard and Visa networks base their authorization communications on the ISO 8583 standard, as do many other institutions and networks. ISO 8583 has no routing information, so is sometimes used with a TPDU header.

Cardholder-originated transactions include purchase, withdrawal, deposit, refund, reversal, balance inquiry, payments and inter-account transfers. ISO 8583 also defines system-to-system messages for secure key exchanges, reconciliation of totals, and other administrative purposes.

Although ISO 8583 defines a common standard, it is not typically used directly by systems or networks. It defines many standard fields (data elements) which remain the same in all systems or networks, and leaves a few additional fields for passing network specific details. These fields are used by each network to adapt the standard for its own use with custom fields and custom usages.

The placements of fields in different versions of the standard varies; for example, the currency elements of the 1987 and 1993 versions are no longer used in the 2003 version, which holds currency as a sub-element of any financial amount element. As of writing, ISO 8583:2003 has yet to achieve wide acceptance.

An ISO 8583 message is made of the following parts:

- Message type indicator (MTI)

- One or more bitmaps, indicating which data elements are present
- Data elements, the fields of the message

**Message Type Indicator**

This is a 4 digit numeric field which classifies the high level function of the message. A message type indicator includes the ISO 8583 version, the Message Class, the Message Function and the Message Origin, each described briefly in the following sections. The following example (MTI 0110) lists what each digit indicates.

```
0xxx -> version of ISO 8583 (for example: 1987 version)
x1xx -> class of the Message (for example: Authorization Message)
xx1x -> function of the Message (for example: Request Response)
xxx0 -> who began the communication (for example: Acquirer)
```

**Message Class**

Position two of the MTI specifies the overall purpose of the message.

| Position | Meaning | Usage |
|---|---|---|
| x0xx | Reserved by ISO | |
| x1xx | Authorization message | Determine if funds are available, get an approval but do not post to account<br>for reconciliation, Dual Message System (DMS), awaits file exchange for posting to account |
| x2xx | Financial messages | Determine if funds are available, get an approval and post directly to the account,<br>Single Message System (SMS), no file exchange after this |
| x3xx | File Actions Message | Used for hot-card, TMS and other exchanges |

| | | |
|---|---|---|
| x4xx | Reversal and Chargeback Messages | Reversal (x4x0 or x4x1): Reverses the action of a previous authorization.<br>Chargeback (x4x2 or x4x3): Charges back a previously cleared financial message. |
| x5xx | Reconciliation Message | Transmits settlement information message |
| x6xx | Administrative Message | Transmits administrative advice. Often used for failure messages (e.g. message reject or failure to apply) |
| x7xx | Fee Collection Messages | |
| x8xx | Network Management Message | Used for secure key exchange, logon, echo test and other network functions |
| x9xx | Reserved by ISO | |

## Message function

Position three of the MTI specifies the message function which defines how the message should flow within the system. Requests are end-to-end messages (e.g., from acquirer to issuer and back with timeouts and automatic reversals in place), while advices are point-to-point messages (e.g., from terminal to acquirer, from acquirer to network, from network to issuer, with transmission guaranteed over each link, but not necessarily immediately).

| Position | Meaning |
|---|---|
| xx0x | Request |
| xx1x | Request response |
| xx2x | Advice |
| xx3x | Advice Response |
| xx4x | Notification |
| xx5x | Notification Acknowledgement |
| xx6x | Instruction (ISO 8583:2003 only) |
| xx7x | Instruction Acknowledgement (ISO 8583:2003 only) |
| xx8x | Reserved for ISO use. (Some implementations use for Response acknowledgment) |
| xx9x | Reserved for ISO use. (Some implementations use for Negative acknowledgment) |

**Message origin**

Position four of the MTI defines the location of the message source within the payment chain.

| Position | Meaning |
|---|---|
| xxx0 | Acquirer |
| xxx1 | Acquirer Repeat |
| xxx2 | Issuer |
| xxx3 | Issuer Repeat |
| xxx4 | Other |
| xxx5 | Other Repeat |

**Processing code**

The following is a table specifying the message type and processing code for each transaction type.

| Transaction | Message Type | Processing Code |
|---|---|---|
| Authorization | 0100 | 00 a0 0x |
| Balance Inquiry | 0100 | 31 a0 0x |
| Sale | 0200 | 00 a0 0x |
| Cash | 0200 | 01 a0 0x |
| Void | 0200 | 02 a0 0x |
| Mobile Topup | 0200 | 57 a0 0x |

## Annex 2. Sample ARRF data for Clustering

@relation LOGDATA

@attribute 'bankid' numeric

@attribute 'terminalid' numeric

@attribute 'chaneltype' numeric

@attribute 'accountnumber' numeric

@attribute 'transactiontype' numeric

@attribute 'transactionamount' numeric

@attribute 'responsecode' {0,5,51,78,94,96,196}

@attribute 'statusdescr' {1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21,22,23,24,25,26,27,28,29,30,31}

@data

6,6001,1230,1300000000000,11000,2000,0,19

6,118003,1230,1360000000000000,11000,100,0,19

6,4002,1230,1240000000000000,11000,500,0,19

6,111001,1230,1320000000000,11000,500,0,19

5,50013001,1210,33274410501,11000,400,0,19

5,50013001,1210,33274410501,11000,200,0,19

3,136008,1210,37348510501,11000,300,0,19

6,100001,1230,1210000000000000,11000,2000,0,19

6,100001,1230,1210000000000000,11000,2000,96,18

6,400001,1230,1320000000000,11000,100,0,19

2,87019,1210,50198110501,11000,100,0,19

6,100001,1210,57585310501,11000,200,0,19

## Annex 3. Sample ARRF data for Classification

@relation LOGDATA

@attribute 'bankid' numeric

@attribute 'terminalid' numeric

@attribute 'chaneltype' numeric

@attribute 'accountnumber' numeric

@attribute 'transactiontype' numeric

@attribute 'transactionamount' numeric

@attribute 'responsecode' numeric

@attribute 'statusdescr' {1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21,22,23,24,25,26,27,28,29,30,31}

@attribute 'fraudstatus' {NF,F}

@attribute 'location' {Adama,Addisababa,Ambo,Assela,Bahirdar,Bishoftu,Chiro,Diredawa,Gonder,Harar,Jimma,Mekele,Nekemte,Sebeta,Shashemene,Woliso,NN}

@attribute 'distancefrombranch' {NTB,FFB,NN}

@data

6,6001,1230,1.3E+12,11000,2000,0,19,NF,Shashemene,NTB

6,118003,1230,1.36E+15,11000,100,0,19,NF,Shashemene,FFB

6,4002,1230,1.24E+15,11000,500,0,19,NF,Addisababa,FFB

6,111001,1230,1.32E+12,11000,500,0,19,NF,Addisababa,NTB

5,50013001,1210,33274410501,11000,400,0,19,NF,Addisababa,NTB

5,50013001,1210,33274410501,11000,200,0,19,NF,NN,NN

3,136008,1210,37348510501,11000,300,0,19,NF,NN,NN

6,100001,1230,1.21E+15,11000,2000,0,19,NF,NN,NN

6,100001,1230,1.21E+15,11000,2000,96,18,F,Addisababa,NTB

6,400001,1230,1.32E+12,11000,100,0,19,NF,Addisababa,NTB

## Annex 6. The original collected sample data set

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| BDS | 20160101 | 555982 | 1.60E+11 | 6 | 6001 | 1220 | C9569057270371127 | 1.30E+12 |
| BDS | 20160101 | 560316 | 1.60E+11 | 6 | 6001 | 1230 | C9569057270371127 | 1.30E+12 |
| BDS | 20160101 | 1223046 | 1.60E+11 | 6 | 118003 | 1220 | C9567059996764645 | 1.36E+15 |
| BDS | 20160101 | 1223428 | 1.60E+11 | 6 | 118003 | 1230 | C9567059996764645 | 1.36E+15 |
| BDS | 20160101 | 5540706 | 1.60E+11 | 6 | 4002 | 1220 | C9567059630305052 | 1.24E+15 |
| BDS | 20160101 | 5541038 | 1.60E+11 | 6 | 4002 | 1230 | C9567059630305052 | 1.24E+15 |
| BDS | 20160101 | 6414569 | 1.60E+11 | 6 | 111001 | 1220 | C9569059969978691 | 1.32E+12 |
| BDS | 20160101 | 6414931 | 1.60E+11 | 6 | 111001 | 1230 | C9569059969978691 | 1.32E+12 |
| BDS | 20160101 | 7323924 | 1.60E+11 | 6 | 100001 | 1220 | C9567051157586024 | 1.21E+15 |
| BDS | 20160101 | 7331473 | 1.60E+11 | 6 | 100001 | 1220 | C9567051157586024 | 1.21E+15 |
| BDS | 20160101 | 7332130 | 1.60E+11 | 6 | 100001 | 1230 | C9567051157586024 | 1.21E+15 |
| BDS | 20160101 | 7332170 | 1.60E+11 | 6 | 100001 | 1230 | C9567051157586024 | 1.21E+15 |
| BDS | 20160101 | 7440853 | 1.60E+11 | 6 | 400001 | 1220 | C9569052277738790 | 1.32E+12 |
| BDS | 20160101 | 7441177 | 1.60E+11 | 6 | 400001 | 1230 | C9569052277738790 | 1.32E+12 |
| BDS | 20160101 | 7570486 | 1.60E+11 | 6 | 100001 | 1200 | C9572052402816672 | 5.76E+10 |
| BDS | 20160101 | 7570756 | 1.60E+11 | 6 | 100001 | 1210 | C9572052402816672 | 5.76E+10 |
| BDS | 20160101 | 7571614 | 1.60E+11 | 6 | 2001 | 1220 | C9569059151883071 | 1.32E+12 |
| BDS | 20160101 | 7571895 | 1.60E+11 | 6 | 2001 | 1230 | C9569059151883071 | 1.32E+12 |
| BDS | 20160101 | 8070039 | 1.60E+11 | 6 | 118003 | 1220 | C9567058292766189 | 1.24E+15 |
| BDS | 20160101 | 8073300 | 1.60E+11 | 6 | 118003 | 1230 | C9567058292766189 | 1.24E+15 |
| BDS | 20160101 | 8073529 | 1.60E+11 | 6 | 118003 | 1220 | C9567058292766189 | 1.24E+15 |
| BDS | 20160101 | 8073830 | 1.60E+11 | 6 | 118003 | 1230 | C9567058292766189 | 1.24E+15 |
| BDS | 20160101 | 8143198 | 1.60E+11 | 6 | 111001 | 1220 | C9567050900989840 | 1.10E+15 |

## Annex 7. Partial View of Decision tree generated with percentage splitting

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

J48 pruned tree

------------------

chaneltype <= 1210: NF (28613.0/3957.0)

chaneltype > 1210

| transactiontype <= 1000

| | transactionamount <= 1555

| | | terminalid <= 2163: F (18.0/1.0)

| | | terminalid > 2163: NF (2.0)

| | transactionamount > 1555: NF (12.0/1.0)

| transactiontype > 1000

| | chaneltype <= 1230: NF (22103.0/864.0)

| | chaneltype > 1230

| | | bankid <= 6

| | | | bankid <= 5: NF (112.0/2.0)

| | | | bankid > 5

| | | | | distancefrombranch = NTB

| | | | | | terminalid <= 215054

| | | | | | | terminalid <= 90001

| | | | | | | | transactionamount <= 350: NF (105.0/6.0)

| | | | | | | | transactionamount > 350

| | | | | | | | | transactionamount <= 1500

| | | | | | | | | | transactionamount <= 750

| | | | | | | | | | | terminalid <= 55001: NF (32.0/5.0)

| | | | | | | | | | | terminalid > 55001: F (2.0)

| | | | | | | | | | transactionamount > 750

| | | | | | | | | | | location = Adama: NF (8.0/1.0)

| | | | | | | | | | | location = Addisababa: F (1.0)

| | | | | | | | | | | location = Ambo: F (0.0)

| | | | | | | | | | | location = Assela: F (0.0)

| | | | | | | | | | | location = Bahirdar: F (0.0)

| | | | | | | | | | | location = Bishoftu: F (0.0)

| | | | | | | | | | | location = Chiro: F (0.0)

| | | | | | | | | | | location = Diredawa: F (0.0)

| | | | | | | | | | | location = Gonder: F (1.0)

| | | | | | | | | | | location = Harar: F (0.0)

| | | | | | | | | | | location = Jimma: F (32.0/11.0)

| | | | | | | | | | | location = Mekele: F (0.0)

| | | | | | | | | | | location = Nekemte: F (0.0)

| | | | | | | | | | | location = Sebeta: F (0.0)

| | | | | | | | | | | location = Shashemene: F (0.0)

| | | | | | | | | | | location = Woliso: F (0.0)

| | | | | | | | | | | location = NN: F (0.0)

| | | | | | | | | | transactionamount > 1500: NF (17.0/1.0)

| | | | | | | terminalid > 90001: NF (135.0/3.0)

| | | | | | terminalid > 215054

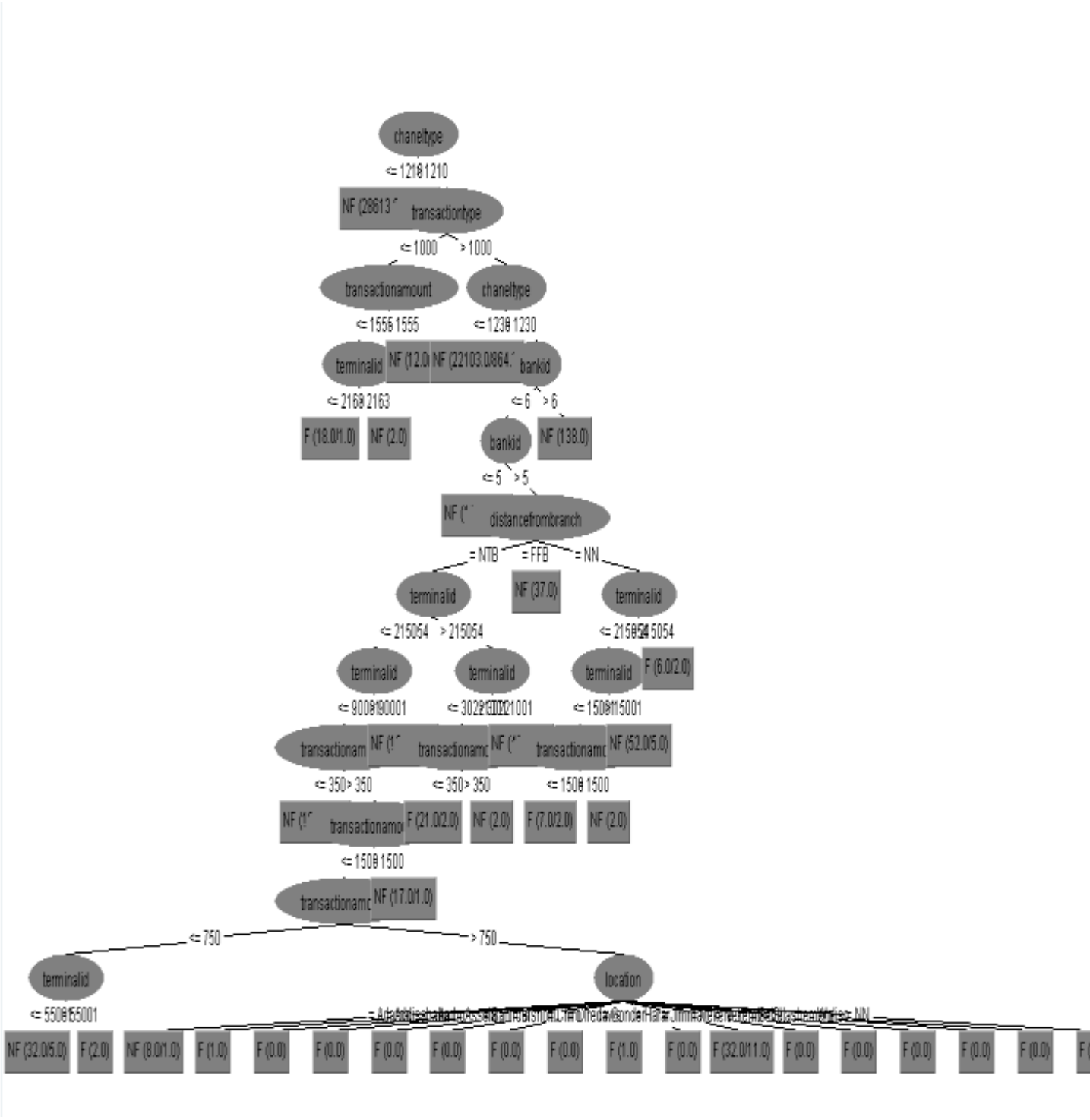
| | | | | | | terminalid <= 30221001

| | | | | | | | transactionamount <= 350: F (21.0/2.0)

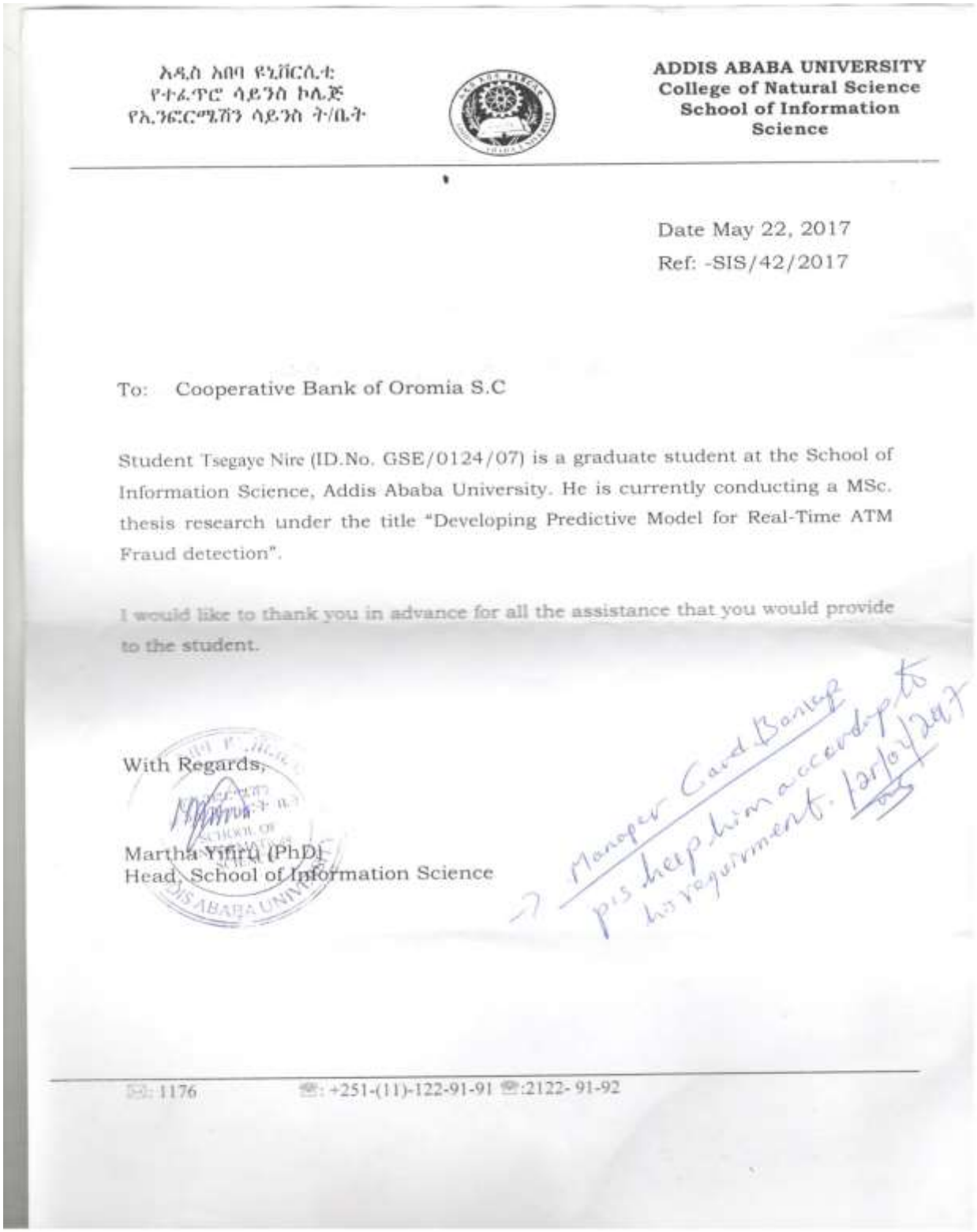
| | | | | | | | transactionamount > 350: NF (2.0)

| | | | | | | terminalid > 30221001: NF (42.0/4.0)

## Annex 8. The Partial View of the Decision Tree J48 Model



## Annex 9. Supporting letter from AAU to the Bank



አዲስ አበባ ዩኒቨርሲቲ
የተፈጥሮ ሳይንስ ኮሌጅ
የኢንፎርሜሽን ሳይንስ ት/ቤት

ADDIS ABABA UNIVERSITY
College of Natural Science
School of Information Science

Date May 22, 2017
Ref: -SIS/42/2017

To: Cooperative Bank of Oromia S.C

Student Tsegaye Nire (ID.No. GSE/0124/07) is a graduate student at the School of Information Science, Addis Ababa University. He is currently conducting a MSc. thesis research under the title "Developing Predictive Model for Real-Time ATM Fraud detection".

I would like to thank you in advance for all the assistance that you would provide to the student.

With Regards,

Martha Yifiru (PhD)
Head, School of Information Science

Manager Card Banking
pls help him accordingly to his requirment. 23/05/2017

1176 +251-(11)-122-91-91 2122- 91-92