IDENTIFYING THE REASON FOR MOBILE CALL DROPS USING DATA MINING TECHNOLOGY

A Thesis Presented

by

Yared Alibo Ayiza

to

The Faculty of Informatics

of

St. Mary's University

In Partial Fulfillment of the Requirements for the Degree of Master of Science

in

Computer Science

November 2018

ACCEPTANCE

IDENTIFYING THE REASON FOR MOBILE CALL DROPS USING DATA MINING TECHNOLOGY

by

Yared Alibo Ayiza

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

Thesis Examination Committee:

Dr. Getahun Semeon
Internal Examiner

Dr. Tibebe Beshah
External Examiner

Dr. Asrat Mulatu
Dean, Faculty of Informatics

November 2018

DECLARATION

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

Yared Alibo Ayiza
Student

_____
Signature

Addis Ababa

Ethiopia

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

Dr. Million Meshesha
Advisor

_____
Signature

Addis Ababa

Ethiopia

November 2018

ACKNOWLEDGMENTS

First and foremost, I would like to thank my ALMAITY GOD for all of things happened in my life and for giving me lovely son Kidus Yared from my lovely wife Kalkidane Alemnew while I was studying this postgraduate program.

Secondly, the success of this thesis is credited to the extensive support and assistance from my advisor Dr. Million Meshesha. I would like to express my grateful gratitude and sincere appreciation to him for his guidance, valuable advice, constructive comments, encouragement and kindness to me throughout this study. Thank you!

Thirdly, I would like to thank my family who supported and encouraged me throughout the time of my study and the research work.

Fourthly, my special thanks go to ethio telecom Analysis and Reporting staffs; specially Yared Tafesse, Tesfaye Belachew and Tarekegne Yohannes; who helped me in providing necessary information and materials which are crucial in this study and gives me support in different area during the research.

Finally, I would like to thank all of you who support me to complete this thesis work as well as my study, even if I didn't mention your name here!

Table of Contents

List of Acronyms

2G: Second Generation

3G: Third Generation

4G: Fourth Generation

ARFF: Attribute Relation File Format

AuC: Authentication Center

BSC: Base Station Controller

BSS: Base Station Subsystem

BTS: Base Transceiver Station

CDR: Call Detail Records

CHAID: Chi-squared Automatic Interaction Detection

CLS: Concept Learning System

CSSR: Call Setup Success Rate

DBMS: Data Base Management System

EIR: Equipment Identity Register

ET: Ethio Telecom

FMS: Fault Management System

GMSC: Gateway Mobile Switching Centre

HLR: Home Locator Register

IMT-2000: International Mobile Telecommunication-2000

KDP: knowledge Discovery Process

LTE: Long Term Evolution

MDL: Minimum Description Length

ME: Mobile Equipment

MS: Mobile Station

MSC: Mobile Switching Center

NSS: Network Switching Subsystem

OSS: Operation Support Subsystem

PBX: Private Branch Exchange

PRF: Premium Rate Fraud

QOS: Quality of Services

RF: Random Forest

RIPPER: Repeated Incremental Pruning to Produce Error Reduction

SEMMA: Sample, Explore, Modify, Model and Assess

SIM: Subscriber Identity Module

SMC: Service Management Center

SMS: Short Message Service

VAS: Value Added Services

VLR: Visitor Location Register

WEKA: Waikato Environment for Knowledge Analysis

List of Figures

List of Tables

ABSTRACT

Access to telecommunication service is critical to the development of all aspects of a nation economy including manufacturing, banking, education, agriculture and government. However, this telecommunication services are not free from problems. Mobile call drop is the main problems of all telecom operators. In telecommunication, mobile call drop is a situation where calls have been cut off before the speaking parties had finished their conversation tone and before one of them had hung up. This research therefore aims to design a predictive model that can determine mobile call drops from ethio telecom mobile network data. To overcome the drawback of simple statistical method, we proposed data mining techniques, methods and methodologies and used in this research.

We selected around 20,000 records of one year and six months collection of Fault Management data. After eliminating irrelevant and unnecessary data, a total of 16996 datasets with 8 attributes are used for the purpose of conducting this study. Data preprocessing was done to clean the datasets. After data preprocessing, the collected data has been prepared in a format suitable for the DM tasks.

The study was conducted using WEKA software version 3.8.1 and four classification techniques; namely, J48 and Random forest algorithm from decision tree as well as PART and JRIP algorithm from rule induction are used. As a result, J48 decision tree algorithm with 10-fold cross validation registered better performance and processing speed of 95.43% and 0.06 sec respectively. The algorithm also used 8 attributes for this experimentation of the research.

Unavailability of related works on telecommunication mobile call drop area was one of the major challenges during the study. Another challenge includes the FMS mobile network server can hold only two years data since the data is huge due to this, we can't get the data before two years.

Finally, we recommend ethio telecom to apply data mining techniques on mobile network data for identifying the reason for call drops.

Keywords: Data Mining, Knowledge discovery, FMS, QoS, Classification, Hybrid, CRISP-DM

CHAPTER ONE: INTRODUCTION

1.1. Background of the Study

"We are living in the information age" is a popular saying; however, we are actually living in the data age [1]. Terabytes or petabytes of data pour into our computer networks, the World Wide Web (WWW), and various data storage devices every day from telecommunications, business, society, science and engineering, medicine, and almost every other aspect of daily life. This explosive growth of available data volume is a result of the computerization of our society and the fast development of powerful data collection and storage tools [1,2].

Information or knowledge has a significant role on human activities. Data mining is the knowledge discovery process by analyzing the large volumes of data from various perspectives and summarizing it into useful information [2]. Due to the importance of extracting knowledge or information from the large data repositories, data mining has become an essential component in various fields of human life. Advancements in Statistics, Machine Learning, Artificial Intelligence, Pattern Recognition and Computation capabilities have evolved the present day's data mining applications and these applications have enriched the various fields of human life including business, education, medical and scientific [3,4].

As more users become reliant upon mobile technology, network operators and service providers face the challenge of increasing capacity to meet user demands, providing services and delivering value added services to their customers with Quality of Services (QoS) [4]. The growth in wireless data, along with the associated increase in content and traffic, necessitates the timely deployment of effective data and telecommunications infrastructure. The telecommunications industry provides communication services to the world through different mechanisms, systems, hardware, software and networks [4,5].

Powerful and versatile tools are badly needed to automatically uncover valuable information from the tremendous amounts of data and to transform such data into organized knowledge. This necessity has led to the birth of data mining [6]. The field is young, dynamic, and promising. Data mining has and will continue to make great strides in our journey from the data age towards information age.

Data Mining, also popularly known as Knowledge Discovery in Databases (KDD), refers to the nontrivial extraction of implicit, previously unknown and potentially useful information from data in databases [6]. While data mining and knowledge discovery in databases are frequently treated as synonyms, data mining is actually part of the knowledge discovery process [7].

The objective of data mining is identifying understandable correlations and patterns from existing data. In order to achieve the objective, the tasks of data mining can be modeled as either Predictive or Descriptive in nature [2,6]. 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 while descriptive data mining focus on finding patterns describing the data that can be interpreted by humans and produces new, nontrivial information based on the available data set.

The goal of predictive data mining is to produce a model that can be used to perform tasks such as classification, prediction or estimation, while the goal of descriptive data mining is to gain an understanding of the analyses by uncovering patterns and relationships in large data sets [2,6].

1.2. The Telecommunication Service

Nowadays the telecommunications industry generates and stores a tremendous amount of data during giving telecom services. These data include call detail data, which describes the calls that traverses the telecommunication networks, network data, which describes the state of hardware and software components in the network, and customer data, which describe telecom customers. This huge amount of data should be handled properly for different purposes like fraud detection, network performance analysis, call drop analysis, customer churn prediction, reporting for higher officials, for network planning and optimization and to support decision making [7,8].

The telecommunication industry was an early adopter of data mining technology [7] and therefore many applications exist in this telecommunication industry like ethio telecom. Major applications include: marketing customer profiling, fraud detection, churn management and network fault isolation [7,8].

Currently, ethio telecom (ET), previously known as Ethiopian Telecommunications Corporation (ETC), is an integrated telecommunications services provider in Ethiopia, providing Internet, data, Value Added Services (VAS), and voice services. It is owned by the Ethiopian government and maintains a monopoly over all telecommunication services in Ethiopia [6].

The introduction of telecommunication in Ethiopia dates back to 1894. The Company is the oldest public telecommunications operator in Africa. In those years, the technological scheme contributed to the integration of the Ethiopian society when the extensive open wire line system was laid out linking the capital with all the important administrative cities of the country [4,8].

After the end of the war against Italy, during which telecommunication network was destroyed, Ethiopia re-organized the Telephone, Telegraph and Postal services in 1941. The Imperial Board of Telecommunications of Ethiopia (IBTE) was established by proclamation No. 131/52 in 1952, which became the Ethiopian Telecommunications Authority in 1981, was placed in charge of both the operation and regulation of telecommunication services in the wake of the market reforms [4].

In 1996, the Government established a separate regulatory body, the Ethiopian Telecommunication Agency (ETA) by Proclamation 49/1996, and during the same year, by regulation 10/1996, the Council of Ministers set up the Ethiopian Telecommunications Corporation (ETC). ETC was replaced by 'ethio telecom' on December 2, 2010 under the proclamation no 197/2010. Under the supervision of the ETA, the principal duty of ET is maintaining and expanding telecommunication services in the country and providing domestic and international telephone, Internet, data, VAS, and other communication services [4,8].

ET provides voice, Internet and data services to the public through fixed line, mobile network and satellite communication throughout the country. From different services ET provides, mobile communication service is one of the biggest and main services with more than 70 million customer uses mobile networks. This mobile service on a mobile network supports customers based on 2G (Second Generation), 3G (Third Generation) network equipment and Long-Term Evolution (LTE) or 4G (Fourth Generation) network equipment [4,8].

The second generation, 2G system fielded in the late 1980s and finished in the late 1990s, was planned mainly for voice transmission with digital signal and the speeds up to 64kbps. 2G wireless cellular mobile service was a step ahead of 2G services by providing the facility of short message

service (SMS). The third-generation, 3G mobile technology is based on wide band wireless network fulfilling the International Mobile Telecommunication-2000 (IMT-2000) specifications by the International Telecommunication Union. 3G offers a vertically integrated, top-down, service-provider approach to delivering wireless Internet access. 4G is a wireless access technology which is a successor of 3G and 4G mobile communications have a transmission rates up to 20Mbps which is higher than 3G [27].

In ethio telecom, mobile performance analysis and mobile quality analysis are sections which are under the Service Management Center (SMC) department in network division are responsible for collecting and analyzing data which are related to call drops.

1.3. Motivation

Currently in ethio telecom, mobile network data analysis is done by traditional simple statistical methods for network quality analysis, network performance analysis and network optimization. However, the application of simple statistical techniques for huge data analysis is time consuming, error prone, tedious, routine and tiresome activities.

In this regard, to overcome the problem of simple statistical analysis an attempt has been made to apply data mining techniques, to show how to discover hidden knowledge from ethio telecom mobile network element data and give effort to evaluate the Quality of Service given to customers.

According to Weiss [32], in telecommunications sector there are many areas of active research interests. Some of these are telecom fraud detection, call drop analysis and prediction, customer churn analysis and prediction, network fault analysis and network element analysis for optimization.

In this research, the application of data mining for telecom mobile network is proposed specifically for identifying the reason for call drops. An attempt has been made to understand the behavior of call drops using data mining process. Since data mining is knowledge discovery process by analyzing the large volume of data from various perspectives and summarizing it in to useful information.

This research aims to design a predictive model that can determine mobile call drops from ethio telecom mobile network data. To overcome the drawback of simple statistical method, we proposed data mining techniques, methods and methodologies and used in this research.

1.4. Statement of the Problem

A disruption or outage in network can lead to call drops and poor sound quality which harms the reputation of the telecom provider and can increase the attrition among its customers. Telecom companies should continuously monitor their networks for such disruptions and resolve root causes at the very early stages [5,7,8].

Nowadays, ethio telecom encounters different problems because of high volume of data generated from the mobile network and need to be analyzed. Based on the data, survey and statistical analysis takes place and the result of the analysis is used for network performance analysis, quality of service analysis, network optimization and maintenance, as well as a report for further decision-making purpose [8].

However, the result of this analysis is not adequate to evaluate the performance of the network, to assess the quality of services delivered to the customers and to use the analysis for further maintenance and expansion purposes. Because of the drawbacks of the result, the performance of the service is degraded and not satisfy customers need. Due to this, the quality of the service is under question and these lead to harms company reputation and finally facilitates customer attrition [4,7,8].

With call detail data collected at a rate of millions of records per second, employing this millions of data for analysis is quite challenging along with analyzing the exact reasons for call drops using manual or simple statistical techniques.

Using data mining technology in telecommunications industry like ethio telecom has become increasingly important since high volume of data is generated every day from mobile network elements which needs further analysis and investigation. Data Mining gives an opportunity to discover hidden knowledge and unknown patterns from telecom mobile network data, which helps better in decision-making [1,9].

Dereje [42] attempt to discover hidden knowledge from ethio telecom GSM mobile network data to determine call setup success rate. Gebremeskel [21] and Jember [17] attempted to extract patterns to support mobile fraud detection on ethio telecom mobile services. As to the researcher's knowledge there is no work that attempt to predict reasons for mobile call drops using ethio telecom mobile network data.

This study therefore tries to apply different Data Mining techniques for constructing a predictive model that helps in determining the reason for mobile call drops. To this end, the study attempts to explore, investigate and answer the following main research questions.

- Which set of data attributes are relevant for call drop analysis?
- Which Data Mining algorithm is suitable to predict the type of call drops?
- To what extent the model enables in identifying the reasons for mobile call drops?

1.5. Objective of the study

1.5.1. General Objective

The general objective of this research is to design a predictive model that can determine mobile call drops from ethio telecom mobile network data using data mining techniques.

1.5.2. Specific Objectives

With the aim of achieving the general objective of the study, the following specific objectives are identified:

- To understand the reasons for call drops based on the review of the related literatures in the area of data mining for telecommunication industry.
- To study and understand mobile network architecture and operation, so as to identify the relevant attributes.
- To understand and get familiar with the data, identify data quality problems and prepare data sets for the experiment.
- To select and prepare data set for experimentation.
- To select data mining algorithms for analysis of mobile network data.
- To construct a predictive model that determine the reasons for mobile call drops.
- To evaluate the performance of the predictive model.
- To report the result and forward recommendation for further research.

1.6. Scope and Limitation of the study

The scope of this research is to develop a predictive model that identifies the reason for mobile call drops from ethio telecom mobile network data. Mobile services on a mobile network supports customers based on 2G (Second Generation), 3G (Third Generation) and Long-Term Evolution (LTE) 4G network equipment. Today in Ethiopia there are more than 70 million mobile users. From these 70 million mobile users more than 15 million users are found in Addis Ababa [4].

Hence this research is limited to call drops happening in Addis Ababa mobile networks. In this research call drops of all 2017 and half year of 2018 (January - June) mobile network data is used because the FMS mobile network server can hold only two years data since the data is huge.

In this research a predictive model is used. A predictive model makes a prediction about values of data using known results found from different historical data. Prediction methods use existing variables to predict unknown or future values of other variables. Predictive model includes classification, prediction, regression and time series analysis [6,]. In this research classification data mining approach is used to predict the mobile call drop reasons from ethio telecom mobile network data.

Unavailability of related works on telecommunication mobile call drop area was one of the major challenges during the study. Another challenge includes, due to time, the depth of the research is on mobile network data which generated from FMS server.

1.7. Methodology

Methodology is the systematic, theoretical analysis of the methods applied to a field of study. It comprises the theoretical analysis of the body of methods and principles associated with a branch of knowledge. Typically, it encompasses concepts such as paradigm, theoretical model, phases and quantitative or qualitative techniques [10].

In Data Mining, Methodology is a way that deals with data collection, analysis and interpretation that shows how to achieve the objective and answer the research questions. Hence, in order to achieve the general and specific objectives of the study the following methods are used.

1.7.1. Research Design

This study follows experimental research. In order to apply this experimental research, we use the six-step process of Hybrid Model. This model was developed, by adopting the CRISP-DM model to the needs of academic research community. Unlike the CRISP-DM process model, which is fully industrial, the Hybrid process model is both academic and industrial.

The main differences and extensions include [11,12]: hybrid model provide more general, research-oriented description of the steps. It introduces a data mining step instead of the modeling step. In addition, it introduces several new explicit feedback mechanisms and modification of the last step, since in the hybrid model the knowledge discovered for a particular domain may be applied in other domains. The overall design issues using Hybrid Data mining Methodology are represented diagrammatically as shown in figure 1.1 below.

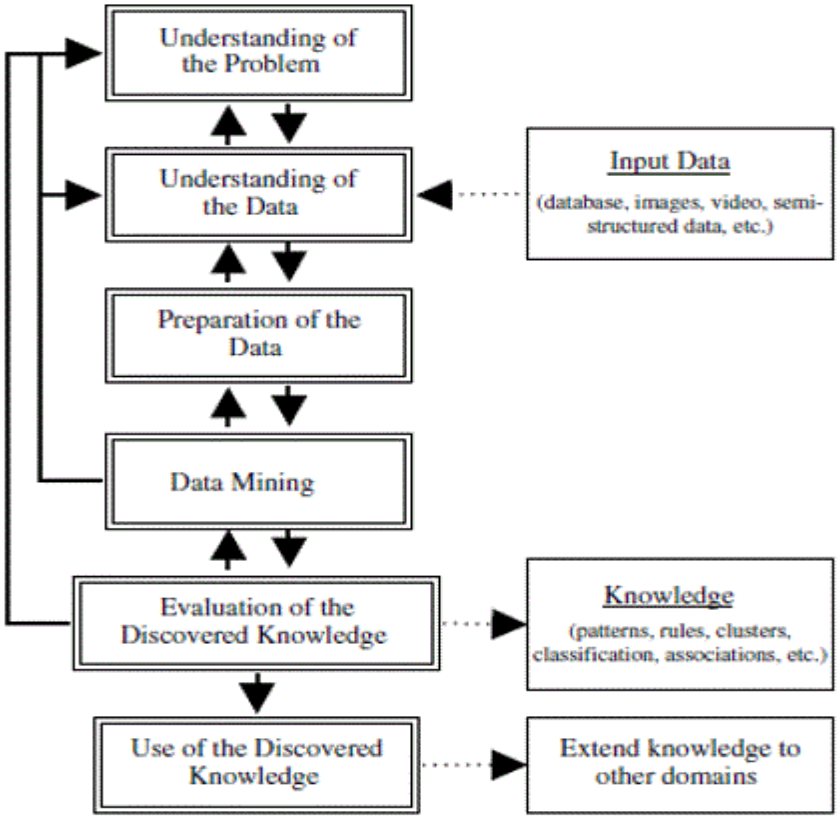


Figure 1.1 The Hybrid DM Process Model [12]

1.7.2. Understanding of the problem domain

This step helps to work closely with domain experts to define the problem and determine the project goals by selecting key people and learning about current solutions to the problem. It also involves learning domain-specific terminology. Finally, project goals are translated into DM goals, and the initial selection of DM tools to be used later in the process is performed.

In this research, in order to identify, define, understand and formulate the problem domain different discussion points reflect the mobile call drops reasons are used, so as to closely works with the domain experts of ethio telecom, then determine attribute feature selection and understanding business processes. In collaborating with the domain experts, the Fault Management System (FMS) data is selected as the main source of data collection. Based on the insight and knowledge gained about the domain of telecom business, data mining problem is defined.

1.7.3. Understanding of the data

This step is used for collecting sample data and deciding which data is important. Data are checked for completeness, redundancy, missing values, plausibility of attribute values. Finally, the step includes verification of the usefulness of the data with respect to the DM goals.

In this research, in order to understand the data, brief discussion on the FMS data are conducted with the domain experts of ethio telecom. The discussion includes listing out initial attributes, their respective values and evaluation of the importance of the FMS data for this research.

Oracle Database: The telecommunication records were very huge data and it is saved in damp file. To read this file, Oracle Database is used and retrieve necessary data in collaboration with domain experts.

Out of the total 16996 datasets, 46.2% of them are belongs to the class of EEU POWER, 32.26% of them are ETHIO POWER and the rest of 21.54% are FIBER&TRANSMISSION classes. With respect to zones where they found EAAZ, WAAZ, NAAZ and SAAZ holds 24.9%, 22.8%, 27.9% and 24.4% respectively. According to their generator status 49.7% of the device has their own stand by generator and the rest of 50.3% has no stand by generator. To understand the nature of the data, descriptive statistics is used. Based on the appropriateness of the problem domain, document analysis, literatures reviewed in chapter two section 2.5 and 2.6 and based on the information obtained from domain experts attributes which are relevant to the study are chosen.

1.7.4. Preparation of the data

This phase is critical for preparing necessary data for subsequent operations. It involves sampling, running correlation and data cleaning, which includes checking the completeness of data records, removing or correcting for noise and missing values. The cleaned data are fed to next operations like reducing dimensionality, discretization and data granularization. The end results are data that meet the specific input requirements for the DM tools selected in first step.

In this research, together with domain expert decision is made to use the FMS data for applying the DM techniques. The cleaned data is further processed by feature selection consulting the domain experts and the Weka attribute selection preprocessing techniques to reduce dimensionality and by derivation of new attributes. The result of these processes generates datasets for training and testing of the classification algorithms selected in this study.

File Splitter: File splitter tool was used because it is a free Windows program that has been used to split the large Excel file into pieces of small size.

MS-Excel: it was used for data preparation, pre-processing and analysis task because it has the capability of filtering attribute with different values. Besides, it is a very important application software to make ready the data and easily convert into a file format acceptable by the WEKA tool.

1.7.5. Data mining for predictive modeling

The main purpose of this research is to develop a predictive model for identifying the call drop reasons using data mining techniques. In this research classification technique is selected because the datasets in FMS data has clear and simplified labeled class.

WEKA version 3.8.1 DM tool has been used to create models using the classification algorithms, such as decision tree and rule induction. WEKA version 3.8 is chosen because [27]:

- It is easy to use by a novice user due to the graphical user interfaces it contains
- It is very portable because it is fully implemented in the Java programming language and thus runs on almost any computing platform
- It contains a comprehensive collection of data preprocessing and modeling techniques, and
- It is freely available under the General Public License (GNU)

1.7.6. Evaluation of the discovered knowledge

In this research different classification models are developed and evaluated using training and testing dataset. The experimental output of the classification models is analyzed and evaluated for performance accuracy using confusion matrix.

After performing the confusion matrix, the results are evaluated by measuring its Accuracy and Error Rate. Furthermore, the effectiveness and efficiency of the model is also computed in terms of recall and precision.

Here in collaboration with domain experts of ethio telecom, understanding the results of the models, checking whether the discovered knowledge is new and interesting, and checking the contribution of the discovered knowledge is evaluated.

1.7.7. Use of Knowledge

After evaluating the discovered knowledge, the last step is using this knowledge for the industrial purposes. In this step the knowledge discovered is incorporated in to performance system and take this action based on the discovered knowledge.

In this research the discovered knowledge is used by integrating the user interface which is designed by java programming language with a Weka system in order to show the prediction of call drop reasons.

Java programming language is chosen because [30]:

- Java is easy to learn: it was designed to be easy to use and is therefore easy to write, compile, debug, and learn than other programming languages.
- Java is object-oriented: This allows us to create modular programs and reusable code.
- Java is platform-independent: One of the most significant advantages of Java is its ability to move easily from one computer system to another.
- Freedom of pointers: JAVA is free from pointers hence we can achieve less development time and less execution time.
- Java is Architectural Neutral: A language or technology is said to be architectural neutral which can run on any available processors in the real world.
- Java is Portable: A portable language is one which can run on all operating systems and on all processors irrespective their architectures and providers.

1.8. Significance of the study

In this research, the application of data mining and knowledge discovery methods and processes in telecommunications network operations especially to ethio telecom is studied. The objective of this research is to design a predictive model, so as to determine mobile call drops from ethio telecom mobile network data using data mining techniques which further used to assist telecom operators in solving everyday problems on mobile call drops. The result of this study helps the network providers to improve the number of calls dropped so as to satisfy its customers. This research also helps for the researchers as a ground root to study on mobile call drops. Generally, this study is important since its application can reduce the revenue losses due to mobile call drops.

1.9. Document Organization

This thesis is organized into five chapters. The first chapter briefly discusses background to the problem area and DM technology, states the problem statement, objective of the study, research methodology, scope and limitation of the study and significance of the results of the research.

The second chapter presents literature review on data mining, data mining overview, knowledge discovery process models, data mining tasks, classification algorithms and critically reviews literatures on telecommunication area like telecom fraud, customer segmentation, customer churn and network fault isolation. In this chapter, international and local works which are related with mobile call drops also reviewed in detail.

The third chapter discusses problem understanding and data preparation processes. In this chapter the general mobile architecture and mobile call drops reason also discussed in detail.

The fourth chapter deals with experimentations and result interpretations. In this chapter building of model with training dataset and validating the result with testing datasets and interpretation of the result of the experimentation was the major concern. This chapter also shows how to use the final experimental results in real world application by doing prototype user interfaces.

The last chapter, chapter five is devoted to concluding remarks and recommendations forwarded based on the research findings of the present study.

CHAPTER TWO: LITERATURE REVIEW

In this chapter, different points that are related to the current work are reviewed from literatures. Overview of data mining, knowledge discovery process models in general and hybrid data mining process model in particular, data mining tasks, classification algorithms, application of data mining in Telecommunication areas and different works related with data mining for telecommunication are presented.

2.1. Overview of Data Mining

Information is available at any time anywhere because of the rapid growth of World Wide Web and electronic information services. The invented machines are faster in producing, manipulating and disseminating information. In information age, an appropriate usage and organization of the information helps to be powerful and to achieve goals. Hence, information processing mechanisms such as automatic data summarization, information extraction and discovering hidden knowledge are very important [33].

2.1.1. What is Data Mining?

Data mining is the process of extracting or mining knowledge from large data sets. But, knowledge mining from data can describe the definition of data mining even if it is long. data mining have similar or a bit different meaning with different terms, such as knowledge mining from data, knowledge extraction, data/pattern analysis, data archaeology, and data dredging [14].

According to Olson [22], Data mining also considered as an exploratory data analysis. Generally, Data mining uses advanced data analysis tools to find out previously unknown (hidden), valid patterns and relationships among data in large data sets. It is the core field for different disciplines such as database, machine learning and pattern recognition.

It is a common practice to refer to the idea of searching applicable patterns in data using different names such as data mining, knowledge extraction, information discovery, information harvesting, data archaeology, and data pattern processing. Among these terms KDD and data mining are used widely [3].

Knowledge discovery was coined at KDD to emphasize the fact that knowledge is the end product of a data-driven discovery and that it has been popularized in the artificial intelligence and machine learning fields. According to Fayyad [35], KDD and data mining are two different terms. KDD

refers to the overall process of discovering useful knowledge from data and data mining refers to a particular step in the process. Furthermore, data mining is considered as the application of specific algorithms for extracting patterns from data.

2.1.2. Why Data Mining?

Nowadays, massive amount of data is produced and collected incrementally. The possibility of gathering and storing huge amount of data by different organizations is becoming true because of using fast and less expensive computers. When organizational data bases keep growing in number and size due to the availability of powerful and affordable database systems the need for new techniques and tools became very important. These tools are used for helping humans to automatically identify patterns, transform the processed data into meaning full information in order to draw concrete conclusions. In addition, it helps in extraction of hidden knowledge from huge amount of digital data [34].

In the private sector industries such as banking, insurance, medicine, telecommunication and retailing use data mining to reduce costs, enhance research, and increase sales. Different organizations worldwide are used data mining techniques for applying and locating higher value customers and to reconfigure their product offerings to increase sales. In the public sector, data mining applications initially were used as a means for detecting fraud and waste of materials, but it grown for different purposes such as measuring and improving program performance [34].

2.2. Knowledge Discovery Process Models

The basic steps of data mining for knowledge discovery are: defining business problem, creating a target dataset, data cleaning and pre-processing, data reduction and projection, choosing the functions of data mining, choosing the data mining algorithms, data mining, interpretation and using the discovered knowledge [35].

The process consists of many steps. Each step attempts to complete a particular discovery task and is accomplished by the application of a discovery method.

Knowledge discovery process includes Knowledge extraction processes like data access and storage, analysis of massive datasets using scalable and efficient algorithms and the interpretation and visualization of the result [11,12].

A Knowledge discovery process model provides an overview of the life cycle of project or research in a data mining, containing the corresponding phases of a project and research, their respective tasks, and relationships between these tasks [11].

There are different process models originated either from academic or industrial environments, all of which have in common the fact that they follow a sequence of steps which more or less resemble between models. Although the models usually emphasize independence from specific applications and tools, they can be broadly divided into those that take into account industrial issues and those that do not. However, the academic models, which usually are not concerned with industrial issues, can be made applicable relatively easily in the industrial setting and vice versa [11,12].

2.2.1. The KDD Process

The knowledge discovery process (KDP), also called knowledge discovery in data base is defined as the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data [35].

As described in figure 2.1 below, Knowledge discovery in database (KDD) has five stages, such as selection, preprocessing, transformation, Data Mining and Interpretation or Evaluation.

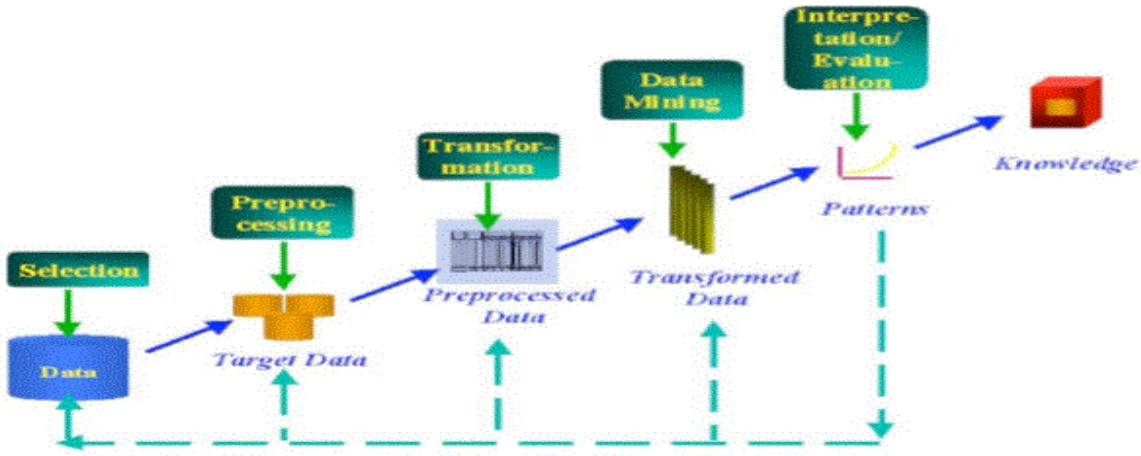


Figure 2.1 The five stages of KDD [35]

Selection: This stage is concerned with creating a target data set or focusing on a subset of variables or data samples, on which discovery is to be performed by Understanding the data and the business area. Because, Algorithms alone will not solve the problem without having clear statement of the objective and understanding.

Pre-processing: This phase is concerned in removing noise or outliers if any, collecting the necessary information to model or account for noise, deciding on strategies for handling missing data fields, and accounting for time sequence information and known changes. On top of these tasks, deciding on DBMS issues, such as data types, schema, and mapping of missing and unknown values are parts of data cleaning and pre-processing.

Transformation: The transformation of data using dimension reduction or transformation methods is done at this stage. Usually there are cases where there are large numbers of attributes in the database for a particular case. With the reduction of dimension there will be an increase in the efficiency of the data-mining step with respect to the accuracy and time utilization.

Data Mining: This phase is the major stage in data KDD because it is all about searching for patterns of interest in a particular representational form or a collection of such representations. These representations include classification rules or trees, regression, clustering, sequence modeling, dependency, and line analysis. Therefore, selecting the right algorithm for the right area is very important.

Evaluation: In this stage the mined data is presented to the end user in a Human viewable format. This involves data visualization, where the user interprets and understands the discovered knowledge obtained by the algorithms.

Using the Discovered Knowledge: Incorporating this knowledge into a performance system, taking actions based on the knowledge, or simply documenting it and reporting it to interested parties, as well as checking for and resolving for conflicts with previously acquired knowledge are tasks in this phase.

Knowledge discovery in database (KDD), as a process consists of an iterative sequence of steps as discussed above. It is also clear that data mining is only one step in the entire process, though an essential one, it uncovers hidden patterns for evaluation.

2.2.2. CRISP-DM Process Model

Data mining needs a standard approach which will help to translate business problems into data mining tasks, suggest appropriate data transformations and data mining techniques, and provide means for evaluating the effectiveness of the results and documenting the experience.

The CRoss Industry Standard Process for Data Mining (CRISP-DM) project addressed parts of these problems by defining a process model which provides a framework for carrying out data. The CRISP-DM methodology provides essential support for those seeking to understand and practice data mining. The required process for success in data mining has been invented independently by many practitioners. By standardizing terminology, CRISP-DM has made it easy for practitioners to communicate about specific data mining projects and about the process in general. CRISP-DM also improves the training of new data miners by providing a detailed and standardized answer to the question "How should data mining be performed?" [6,31,43].

As described in figure 2.2 CRISP-DM has six steps, such as Business understanding, Data understanding, Data preparation, Modeling, Evaluation and Deployment [31,43].

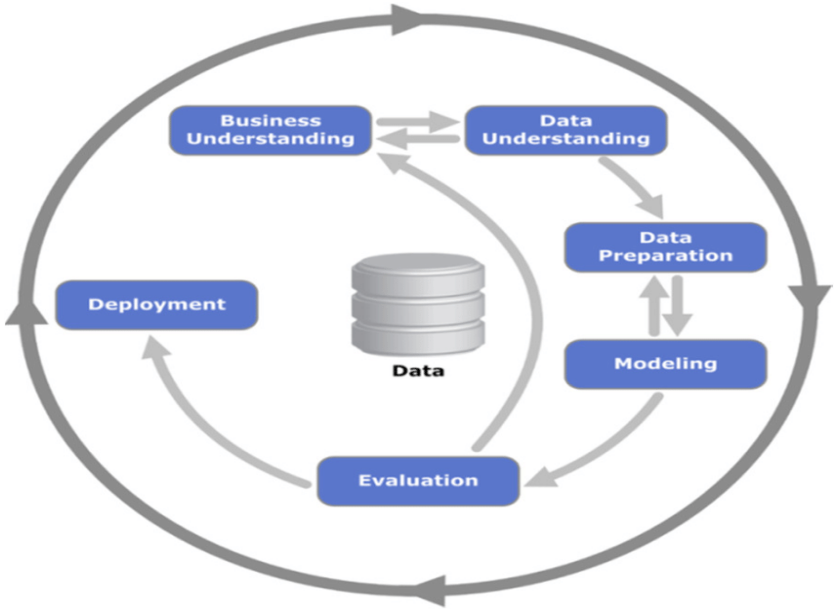


Figure 2.2 CRISP-DM Process Model [22,34]

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

Data Understanding: The data understanding phase starts with an initial data collection and proceeds with activities in order to get familiar with the data, to identify data quality problems, to discover first insights into the data, or to detect interesting subsets to form hypotheses for hidden information. There is a close link between Business Understanding and Data Understanding.

The formulation of the data mining problem and the project plan require at least some understanding of the available data.

Data Preparation: Data are the backbone of data mining and knowledge discovery. However, real-world business data usually are not available in data-mining- ready form. The biggest challenge for data miners, then, is preparing data suitable for modeling [31,43].

Data preparation comprises techniques concerned with analyzing raw data so as to yield quality data, mainly including data integration, data transformation, data cleaning, data reduction, and data discretization.

Modelling: In this phase, various modelling techniques are selected and applied, and their parameters are calibrated to optimal values. Typically, there are several techniques for the same data mining problem type. Some techniques require specific data formats. There is a close link between data preparation and modelling. Often, one realizes data problems while modelling or one gets ideas for constructing new data.

Evaluation: At this stage in the project have built one or more models that appear to have high quality, from a data analysis perspective. Before proceeding to final deployment of the model, it is important to more thoroughly evaluate the model, and review the steps executed to construct the model, to be certain it properly achieves the business objectives. A key objective is to determine if there is some important business issue that has not been sufficiently considered.

Deployment: Creation of the model is generally not the end of the project. Usually, the knowledge gained will need to be organized and presented in a way that the customer can use it. Depending on the requirements, the deployment phase can be as simple as generating a report or as complex as implementing a repeatable data mining process. In many cases it will be the user, not the data analyst, who will carry out the deployment steps.

2.2.3. Hybrid DM Process Model

Hybrid DM process model combine aspects of both academic and industrial areas [43]. It was developed based on the CRISP-DM model by adopting it to academic research. The main differences and extensions include: Hybrid model provide more general, research-oriented description of the step (see figure 1.1), it introduces a data mining step instead of the modeling step. In addition, it introduces several new explicit feedback mechanisms and modification of the last step, since in the hybrid model, the knowledge discovered for a particular domain may be applied in other domains. Hybrid data mining model includes the following six steps [12].

Understanding of the problem domain: This step helps to work closely with domain experts to define the problem and determine the project goals by selecting key people and learning about current solutions to the problem. It also involves learning domain-specific terminology. Finally, project goals are translated into DM goals, and the initial selection of DM tools to be used later in the process is performed.

Understanding of the data: This step is used for collecting sample data and deciding which data is important. Data are checked for completeness, redundancy, missing values, plausibility of attribute values. Finally, the step includes verification of the usefulness of the data with respect to the DM goals.

Preparation of the data: This phase uses for preparing necessary data for subsequent operations. It involves sampling, running correlation and significance tests, and data cleaning, which includes checking the completeness of data records, removing or correcting for noise and missing values. The cleaned data are fed to next operations like reducing dimensionality, discretization and data granularization. The end results are data that meet the specific input requirements for the DM tools selected in first step.

Data mining: The main purpose of this research is to develop a model for identifying the call drop reasons using data mining techniques. In this research classification technique is selected because the datasets in FMS data has clear and simplified labeled class.

Evaluation of the discovered knowledge: In this research different classification models will be developed and evaluated using training and testing dataset. The experimental output of the classification models is analyzed and evaluated the performances accuracy using confusion matrix.

Use of Knowledge: After evaluating the discovered knowledge, the last step is using this knowledge for the industrial purposes. In this step the knowledge discovered is incorporated in to performance system and take this action based on knowledge or simply document it and report it to the interested parties and also check and resolve conflicts with previously acquired knowledge if any.

2.3. Data Mining Tasks

The objective of data mining is identifying understandable correlations and patterns from existing data. In order to achieve the objective, the tasks of data mining can be modeled as either Predictive or Descriptive in nature [35,43], as shown in figure 2.3 below.

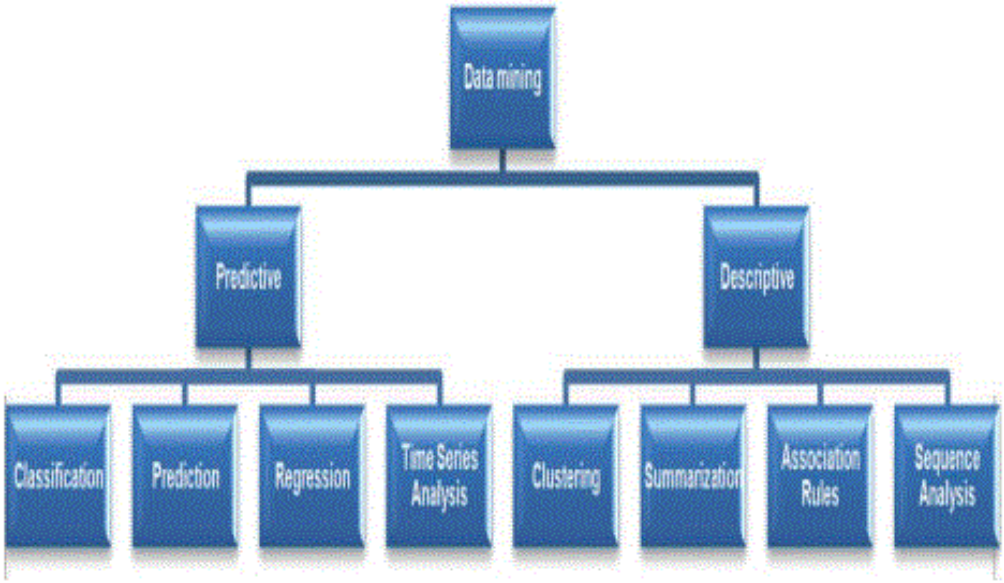


Figure 2.3 Data Mining Tasks and Models [35,43]

According to Williams [13], modeling is the first thing that comes to mind when someone thinks about data mining. Modeling is the process of taking some data and building implied description of the processes that might have generated it. The description is most of the time a computer program or mathematical formula. A model captures the knowledge exhibited by the data and encodes it in some language. Often the aim is to address a septic problem through modeling the world in some form and then use the model to develop a better understanding of the world.

There are basically two type of models constructed by data mining: predictive and descriptive modeling [12,35]. Predictive data mining task perform inference of the current data in order to make prediction while descriptive data mining task characterizes the general property of the data in the database. 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 while descriptive data mining focus on finding patterns describing the data that can be interpreted by humans and produces new, nontrivial information based on the available data set.

The goal of predictive data mining model is to produce a model that can be used to perform tasks such as classification, prediction or estimation while the goal of descriptive data mining is to gain an understanding of the analyses system by uncovering patterns and relationships in large data sets. The objective of predictive data mining model is to predict the future outcomes based on passed records with known answers while the objective of a descriptive data mining model is to discover patterns in the data and to understand the relationships between attributes represented by the data [12,43].

2.3.1. Predictive Model

A predictive model makes a prediction about values of data using known results found from different historical data. Prediction methods use existing variables to predict unknown or future values of other variables [35,36]. As shown in figure 2.3, Predictive model includes classification, prediction, regression and time series analysis.

According to Cios [12], Classification is the best understood of all data mining approaches among all Predictive models. Classification is commonly characterized as with classification tasks such as supervised learning, categorical dependent variable and ability of assigning new data in to the set of well-defined classes.

Classification is one of the classic data mining techniques used to classify each item in a set of data into one of predefined set of classes or groups [12]. Classification method makes use of mathematical techniques such as decision trees, support vector machine, neural network and Bayesian learning. In classification, software is developed that can learn how to classify the data items into groups [12].

Classification involves the discovery of a predictive learning function that classifies a data item into one of several predefines classes. It involves examining the features of a newly presented object and assigning it to a predefined class. Classification is a two-step process. First a model is built describing a predetermined set of data classes or concepts and second, the model is used for classification [12,36].

Prediction can be viewed as the construction and use of a model to assess the class of an unlabeled sample, or to assess the value or value range of an attribute that a given sample is likely to have. Any of the techniques used for classification can be adapted for use in prediction by using training examples where the value of the variable to be predicted is already known, along with historical data for those examples [36].

2.3.2. Descriptive Modeling

According to Rokach [40], Descriptive data mining method can be defined as discovering interesting regularities in the data, to uncover patterns and find interesting subgroups in the bulk of data is normally used to generate frequency, cross tabulation and correlation.

Unlike the predictive model, a descriptive model serves as a way to explore the properties of the reasons for mobile call drops, not to predict new call drop reasons. Descriptive task encompasses methods such as Clustering, Association Rules, Summarizations and Sequence analysis. But data mining involves clustering and association rule discovery methods.

Clustering: According to Jackson [6], Clustering is a data mining technique that makes meaningful or useful cluster of objects which have similar characteristics using automatic technique. The clustering technique defines the classes and puts objects in each class, while in the classification techniques, objects are assigned into predefined classes.

In Clustering, a set of data items is partitioned into a set of classes such that items with similar characteristics are grouped together. Clustering is best used for finding groups of items that are similar. For example, given a data set of customers, subgroups of customers that have a similar buying behavior can be identified [38].

Clustering which is also called unsupervised learning, groups similar data together into clusters. It is used to find appropriate groupings of elements for a set of data. Unlike classification, clustering

is a kind of undirected knowledge discovery or unsupervised learning; that is, there is no target field, and the relationship among the data is identified by bottom-up approach [6].

Association rule discovery: Association is one of the best-known data mining technique. In association, a pattern is discovered based on a relationship between items in the same transaction for example calling number and peak status. That's the reason why association technique is also known as relation technique. The association technique is used in market basket analysis to identify a set of products that customers frequently purchase together. Telecom can also use association technique between calling date and time with zone to identify its customer calling habits [6].

In this study an attempt is made to create a predictive model using classification algorithms that attempts to determine the reason for call drops.

2.4. Classification Algorithms

The classification task can be seen as a supervised technique where each instance belongs to a class, which is indicated by the value of a special goal attribute or simply the class attribute [1]. The goal attribute can take on categorical values, each of them corresponding to a class. Each example consists of two parts, namely a set of predictor attribute values and a goal attribute value. The former is used to predict the value of the latter. The predictor attributes should be relevant for identifying the class of an instance. In the classification task the set of examples being mined is divided into two mutually exclusive and exhaustive sets, called the training set and the test set [1].

The classification model is built from the training set, and then the model is evaluated on the test set. During training, the classification algorithm has access to the values of both predictor attributes and the other attribute for all examples of the training set, and it uses that information to build a classification model. This model represents classification knowledge essentially, a relationship between predictor attribute values and classes that allows the prediction of the class of an example given its predictor attribute values. For testing, the test set the class values of the examples is not shown. In the testing phase, only after a prediction is made is the algorithm allowed to see the actual class of the just-classified example. One of the major goals of a classification algorithm is to maximize the predictive accuracy obtained by the classification model when classifying examples in the test set unseen during training [1,22].

There are different classification algorithms that are used for constructing a predictive model. Common classification algorithm includes Decision Tree, rule induction, K-Nearest Neighbor, Support Vector Machines, Naive Bayesian Classification and Neural Networks [22]. In this study decision tree and rule induction algorithms are used.

2.4.1. Decision tree

Decision tree is one of the most used data mining techniques because its model is easy to understand for users. In decision tree technique, the root of the decision tree is a simple question or condition that has multiple answers. Each answer then leads to a set of questions or conditions that help us determine the data so that we can make the final decision based on it. The algorithms that are used for constructing decision trees usually work top-down by choosing a variable at each step that is the next best variable to use in splitting the set of items [40].

A decision tree is a classifier expressed as a recursive partition of the instance space. The decision tree consists of nodes that form a rooted tree (see figure 2.4 below), 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 frequent case, each test considers a single attribute, such that the instance space is partitioned according to the attribute's value. In the case of numeric attributes, the condition refers to a range.

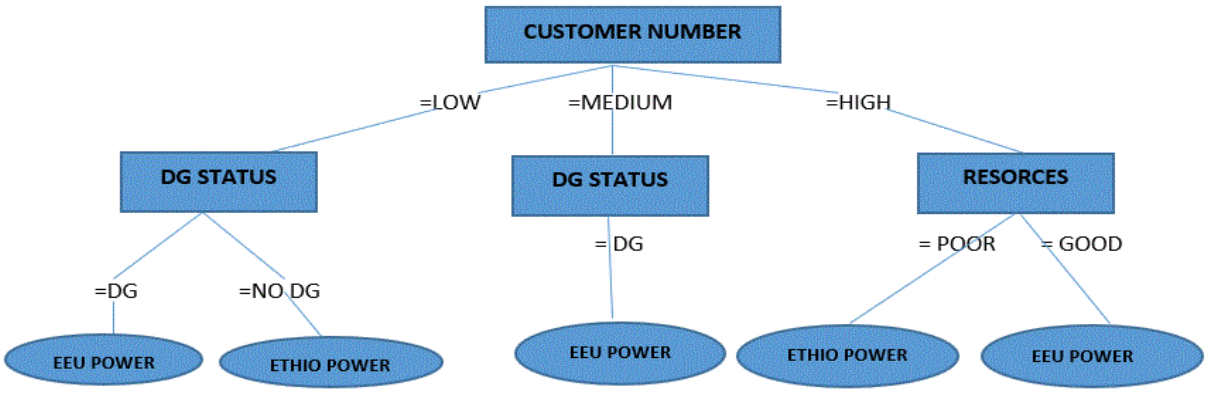


Figure 2.4 Decision tree classifier for mobile call drop reason

Each leaf is assigned to one class representing the most 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.

The many benefits in data mining that decision trees offer include the following [28,40]:

- Decision trees require very little data preparation whereas other techniques often require data normalization, the creation of dummy variables and removal of blank values.
- Uses a white box model i.e. the explanation for the condition can be explained easily by Boolean logic because there are mostly two outputs. For example, yes or no.
- Self-explanatory and easy to follow when compacted
- Able to handle a variety of input data: nominal, numeric and textual
- Able to process datasets that may have errors or missing values
- High predictive performance for a relatively small computational effort
- Available in many data mining packages over a variety of platforms
- Useful for various tasks, such as classification, regression, clustering and feature selection.

Some of the weaknesses of DT are [28,40]:

- Some DT can only deal with binary valued target classes, others are able to assign records to an arbitrary number of classes, but errors are prone when the number of training examples per class gets small. This can happen rather quickly in a tree with many levels and many branches per node.
- The process of growing a DT is computationally expensive. At each node, each candidate splitting field is examined before its best split can be found.
- Decision tree are less appropriate for estimation tasks where the goal is to predict the value of continuous such as income, blood pressure, or interest rate.
- Decision tree are also problematic for time-series data values a lot of effort is put into presenting the data in such a way that tends and sequential patterns are made visible.

The construction of the decision tree involves the following three main phases [28,40]:

- **Construction Phase:** The initial decision tree is constructed in this phased based on the entire training data set. It requires recursively partitioning the training set in to two or more sub-partition using a splitting criterion until a stopping criterion is met.
- **Pruning Phase:** The tree constructed in the previous phase may not result in the best possible set of rules due to over-fitting. The pruning phase removes some of the lower branches and nodes to improve its performance.
- **Processing the Pruned tree:** To improve understandability.

When a decision tree is built, many of the branches will reflect anomalies in the training data due to noise or outliers. Tree pruning methods address this problem of over fitting the data. There are two common approaches to tree pruning [28].

I. **Pre-pruning: -** In the pre-pruning approach, a tree is "pruned" by halting its construction early. Upon halting, the node becomes leaf. The leaf may hold most frequent class among the subset tuples or the probability distribution of those tuples. When constructing a tree, attribute selection measures such as statistical significance, information gain, gini index and so on can be used to access the goodness of a split. If partitioning the tuple at node would result in a split that falls below a prespecified threshold, then further partitioning of a given subset is halted. There are difficulties, however in choosing an appropriate threshold. High threshold could result in oversimplified trees; whereas low thresholds could result in very little simplification.

II. **Post-pruning: -** Post-pruning removes sub trees from a "fully grown" tree. A sub tree at a given node is pruned by removing its branches and replacing it with leaf. The leaf is labeled with the most frequent class among the sub tree being replaced. The "best" pruned tree is one that minimizes the encoding bits. This method adopts Minimum Description Length (MDL) principle. The basic idea is that the simplest solution is preferred.

Alternatively, pre-pruning and post-pruning may be interleaved for a combined approach. Post-pruning requires more computation than pre-pruning, yet generally leads to a more reliable tree. Although pruned tree tends to be more compact than their unpruned counterparts, they may still be rather large and complex. Decision tree can suffer from repetition and replication.

2.4.1.1. Decision tree Basic Principle (Hunt's method)

All DT induction algorithms follow the basic principle, known as CLS (Concept Learning system), given by Hunt. A CLS tries to mimic the human process of learning a concept, starting with examples from two classes and then inducing a rule to distinguish the two classes based on other attributes. Let the training dataset be T with class-labels {C1, C2... Ci}. The decision tree is built by repeatedly partitioning the training data using some splitting criterion till all the records in a partition belong to the same class. The steps to be followed are [28]:

I. If T contains no cases (T is trivial), the decision tree for T is a leaf, but the class to be associated with the leaf must be determined from information other than T.
II. If T contains cases all belonging to a single class Cj (homogeneous), corresponding tree is a leaf identifying class Cj.
III. If T is not homogeneous, a test is chosen, based on a single attribute, that has one or more mutually exclusive outcomes {O1, O2..., On}. T is partitioned into subsets T1, T2, T3... Tn, where Ti contains all those cases in T that have the outcome Oi of the chosen test.

The decision tree for T consists of a decision node identifying the test, and one branch for each possible outcome. The same tree building method is applied recursively to each subset of training cases.

2.4.1.2. Measures of Diversity in Decision tree

The diversity index is a well-developed topic with different names corresponding the various fields. To statistical biologist, it is Simpson diversity index. To cryptographers, it is one minus the repeat rate. To econometricians, it is the Gini index that is also used by the developers of the Classification and Regression Trees (CART) algorithm. A high index of diversity indicates that the set contains an even distribution of classes whereas a low index means that members of a single class predominate [28]. The best splitter is the one that decreases the diversity of the record sets by the greatest amount. The three common diversity functions are discussed here. Let there be a dataset S (training data) of C outcomes. Let P(I) denotes the proportion of S belonging to a class I where I varies from 1 to C for the classification problem with C classes.

$$Simple\ Diversity\ index = Min(p(I)) \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \textbf{(2.1)}$$

Entropy provides an information theoretic approach to measure the goodness of a split. It measures the amount of information in an attribute.

$$Entropy(S) = \sum_{I=1}^{C}(-p(I)\log_2 p(I)) \qquad \text{..... (2.2)}$$

Gain(S,A), the information gain of the example set S on an attribute A, defined as

$$Gain(S,A) = Entropy(S) - \sum\left(\left(\frac{|S_V|}{|S|}\right) * Entropy(S_V)\right) \qquad \text{..... (2.3)}$$

Where $\sum$ is over each value V of all the possible values of the attribute A,

SV = subset of S for which attribute A has value V, |SV| = number of elements in SV, and

|S| = number of elements in S.

The above notion of gain tends to favor the attributes that have a larger number of values. To compensate this, it is suggested using the gain ratio instead of gain, as formulated below.

$$Gain\,Ratio(S,A) = \frac{Gain(S,A)}{SplitInfo(S,A)} \qquad \text{..... (2.4)}$$

Where SplitInfo(S, A) is the information due to the split of S on the basis of the value of the categorical attribute A. Thus SplitInfo(S,A) is entropy due to the partition of S induced by the value of the attribute A.

Gini index measures the diversity of population using the formula

$$Gini\ Index = 1 - \sum\left(p(I)^2\right) \qquad \text{..... (2.5)}$$

Where P(I) is the proportion of S belonging to class I and $\sum$ is over C.

A number of different algorithms may be used for building decision trees including Chi-squared Automatic Interaction Detection (CHAID), Classification and Regression Trees (CART), C4.5, J48 and Random forest [48].

2.4.1.3. J48 decision tree algorithm

Decision tree models are constructed in a top-down recursive divide-and-conquer manner. J48 decision tree algorithms have adopted this approach. The training set is recursively partitioned into smaller subsets as the tree is being built [6].

According to Rokach [40], J48 decision tree algorithm is a predictive machine learning model that decides the target value of a new sample based on various attribute values of the available data. Having the capability of generating simple rules and removing irrelevant attributes, the J48 decision tree can serve as a model for classification.

According to Hemalatha [41], J48 decision tree algorithm performs the following sequence of steps to accomplish its classification task.

- It checks for base cases
- It finds normalized information gain
- It selects the attribute with the highest normalized information gain
- It Creates decision nodes
- It recourses on the sub lists obtained by splitting and add those nodes as children of node

J48 Decision Tree Classifier uses two phases [41]: tree construction and tree pruning.

Tree construction starts with the whole data set at the root. It then checks the attribute of the data set and partition them based on the following cases

Case I: - If the attribute value is clear and has a target value, then it terminates the branch and assigns the value as Target value (classification)

Case II: - If the attribute gives the highest information, then continue till we get a clear decision or run out of attributes.

Case III: - If we run out of attributes or we are presented with ambiguous result, then assign the present branch as target value.

Case IV: - Ignore missing values.

The second phase Tree Pruning identifies and remove branches that reflects noise and outliers to reduce classification errors.

J48 algorithm has various advantages. Some of these advantages are the following [41]:

- Gains of balanced flexibility and accuracy
- Capability of limiting number of possible decision points
- Higher accuracy

2.4.1.4. Random forest (RF)

Random forests (RF) are a combination of tree predictors that each tree depends on the values of a randomly selected vector samples and it distributes equal values of vector samples to all of the trees in the forest. The strength of individual trees in the forest and the correlation between them determines the generalized error of a forest and its tree [15].

As primary form knowledge representation, the random forest algorithm is presented in terms of decision trees. Though, the random forest algorithm is thought of as a meta-algorithm. Decision tree algorism or any one of the other model building algorisms could be the actual model builder [15].

RF Algorithm performs the following sequence of steps to accomplish its classification [41].

- Choose T number of trees to grow
- Choose m number of variables used to split each node. m<<M, where M is the number of input variables and m holds constant while growing the forest.
- Grow T trees (when growing each tree do)
- Construct a bootstrap sample of size n sampled from Sn with the replacement and grow a tree from this bootstrap sample
- When growing a tree at each node select m variables at random and use them to find the best split
- Grow the tree to a maximal extent and there is no pruning
- To classify point X, collect votes from every tree in the forest and then use majority voting to decide on the class label

Random forest algorithm has various advantages. Some of these advantages are the following [15],

- It has the ability of running on large data bases.
- It has the capability of handling thousands of input variables without variable deletion and fast learning.
- Random forest algorithm has an effective method of estimating missed data and highest maintenance accuracy.
- The algorithm has the ability of saving generated forests for future use.

2.4.2. Rule Induction system

According to Witten [28], the rule-based induction method is one of the most important machine learning techniques as it can express the regularities regarding rules that are frequently hidden in the data. It is the most fundamental tool in the data mining process. Generally, rules are expressions of the form: If (condition), then conclusion.

if (characteristic 1 is equal to value 1) and (characteristic 2 is equal to value 2) and ....and

(characteristic n is equal to value n), then (decision will be equal to the result).

Witten [28] stated that some rule induction systems provoke more complex rules, in which the characteristic values are expressed by the contradiction of some other values or by a value of the overall subset of the characteristic domain. It was further explained that the data by which the rules are provoked are generally presented in a form similar to a table that shows different cases (rows) against the variables (characteristics and decisions).

Rajput [46] said that the rule induction belongs to supervised learning, and all of it cases are pre-classified by experts. In simple words the decision values are assigned by the experts in this process. Anil further elaborated that the characteristics represent independent values, while the decisions represent the dependent variables. The covering method represents classification of knowledge in the form of a set of rules which represent or give a description of each class.

This procedure makes use of the following search process to produce the rules for each class in the training set T:

While the stopping criterion is not satisfied:

- Form a new rule to cover examples belonging to a target class employing the Rule Forming Process;
- Add this rule to the Rule Set;
- Remove all examples from T which are covered by this new rule.
- Stop the procedure when there are no more classes to classify.

2.4.2.1. JRIP rule classifier

According to Rajput [46], JRIP is one of the basic and most popular rule induction algorithms. Classes are examined in increasing size and an initial set of rules for the class is generated using incremental reduced error JRIP (RIPPER) proceeds by treating all the examples of a particular judgment in the training data as a class and finding a set of rules that cover all the members of that class. Thereafter it proceeds to the next class and does the same, repeating this until all classes have been covered.

JRip implements a propositional rule learner, Repeated Incremental Pruning to Produce Error Reduction (RIPPER). It is based in association rules with reduced error pruning (REP), a very common and effective technique found in decision tree algorithms [46].

2.4.2.2. PART

According to Rajput [46], PART is a separate-and-conquer rule learner. The algorithm producing sets of rules called decision lists which are planned set of rules. A new data is compared to each rule in the list in turn, and the item is assigned the class of the first matching rule. PART builds a partial C4.5 decision tree in each iteration and makes the “best” leaf into a rule.

PART is a partial decision tree algorithm, which is the developed version of C4.5 and RIPPER algorithms. The main specialty of the PART algorithm is that it does not need to perform global optimization like C4.5 and RIPPER to produce the appropriate rules. However, decision trees are sometime more problematic due to the larger size of the tree which could be oversized and might perform badly for classification problems [46].

2.5. Data Mining Tools

Within data mining, there is a group of tools that have been developed by a research community and data analysis enthusiasts; they are offered free of charge using one of the existing open-source licenses. An open-source development model usually means that the tool is a result of a community effort, not necessary supported by a single institution but instead the result of contributions from an international and informal development team. This development style offers a means of incorporating the diverse experiences.

Data mining provides many mining techniques to extract data from databases. Data mining tools predict future trends, behaviors, allowing business to make proactive, knowledge driven decisions. The development and application of data mining algorithms requires use of very powerful software tools. As the number of available tools continues to grow, the choice of most suitable tool becomes increasingly difficult [47]. Open source tools available for data mining are briefed as below. In order to conduct this research, we select WEKA data mining tools.

2.5.1. WEKA (Waikato Environment for Knowledge Analysis)

Weka is a collection of machine learning algorithms for data mining tasks. These algorithms can either be applied directly to a data set or can be called from your own Java code. The Weka (pronounced Weh-Kuh) workbench contains a collection of several tools for visualization and algorithms for analytics of data and predictive modeling, together with graphical user interfaces for easy access to this functionality.

Weka is a Java based open source data mining tool which is a collection of many data mining and machine learning algorithms, including pre-processing on data, classification, clustering, and association rule extraction. It provides three graphical user interfaces i.e. the Explorer for exploratory data analysis to support preprocessing, attribute selection, learning, visualization, the Experimenter that provides experimental environment for testing and evaluating machine learning algorithms, and the Knowledge Flow for new process model inspired interface for visual design of KDD process. A simple Command-line explorer which is a simple interface for typing commands is also provided by WEKA.

Weka loads data file in formats of ARFF, CSV, C4.5, binary. Though it is open source, Free, Extensible, can be integrated into other java packages.

2.5.2. KEEL (Knowledge Extraction based on Evolutionary Learning)

Knowledge Extraction based on Evolutionary Learning is an application package of machine learning software tools. KEEL is designed for providing solution to data mining problems and assessing evolutionary algorithms. It has a collection of libraries for preprocessing and post-processing techniques for data manipulating, soft-computing methods in knowledge of extracting and learning and providing scientific and research methods.

Keel is a software tool to assess evolutionary algorithms for Data Mining problems. It includes regression, classification, clustering, and pattern mining and so on. It also contains a big collection of classical knowledge extraction algorithms, preprocessing techniques (instance selection, feature selection, discretization, imputation methods for missing values etc.), Computational Intelligence based learning algorithms, including evolutionary rule learning algorithms based on different approaches (Pittsburgh, Michigan and IRL), and hybrid models such as genetic fuzzy systems, evolutionary neural networks etc [47].

2.5.3. R (Revolution)

Revolution is a free software programming language and software environment for statistical computing and graphics. The R language is widely used among statisticians and data miners for developing statistical software and data analysis. One of R's strengths is the ease with which well-designed publication-quality plots can be produced, including mathematical symbols and formulae where needed.

R is a well-supported, open source, command line driven, statistics package. There are hundreds of extra “packages” freely available, which provide all sorts of data mining, machine learning and statistical techniques. It allows statisticians to do very intricate and complicated analyses without knowing the blood and guts of computing systems.

2.5.4. KNIME (Konstanz Information Miner)

Konstanz Information Miner, is an open source data analytics, reporting and integration platform. It has been used in pharmaceutical research but is also used in other areas like CRM customer data analysis, business intelligence and financial data analysis. It is based on the Eclipse platform and, through its modular API, and is easily extensible. Custom nodes and types can be implemented in

KNIME within hours thus extending KNIME to comprehend and provide firsttier support for highly domain-specific data format.

Knime, pronounced “naim”, is a nicely designed data mining tool that runs inside the IBM’s Eclipse development environment. It is a modular data exploration platform that enables the user to visually create data flows (often referred to as pipelines), selectively execute some or all analysis steps, and later investigate the results through interactive views on data and models. The Knime base version already incorporates over 100 processing nodes for data I/O, preprocessing and cleansing, modeling, analysis and data mining as well as various interactive views, such as scatter plots, parallel coordinates and others.

2.5.5. RAPIDMINER

RAPIDMINER is a software platform developed by the company of the same name that provides an integrated environment for machine learning, data mining, text mining, predictive analytics and business analytics. It is used for business and industrial applications as well as for research, education, training, rapid prototyping, and application development and supports all steps of the data mining process. Rapid Miner uses a client/server model with the server offered as Software as a Service or on cloud infrastructures.

Rapid Miner provides support for most types of databases, which means that users can import information from a variety of database sources to be examined and analyzed within the application.

2.5.6. ORANGE

Orange is a component-based data mining and machine learning software suite, featuring a visual programming frontend for explorative data analysis and visualization, and Python bindings and libraries for scripting. It includes a set of components for data preprocessing, feature scoring and filtering, modeling, model evaluation, and exploration techniques. It is implemented in C++ and Python. Its graphical user interface builds upon the cross-platform framework.

Orange is a component-based data mining and machine learning software suite. It includes a set of components for data preprocessing, feature scoring and filtering, modeling, model evaluation, and exploration techniques. Data mining in Orange is done through visual programming or Python scripting.

No.	Tool Name	Release Date	License	Language	Operating System	Type
1.	RAPID MINER	2006	AGPL Proprietary	Language Independent	Cross platform	Statistical analysis, data mining, predictive analytics.
2.	ORANGE	2009	GNU General Public License	Python C++, C	Cross Platform	Machine learning, Data mining, Data visualization
3.	KNIME	2004	GNU General Public License	Java	Linux, OS X, Windows	Enterprise Reporting, Business Intelligence, Data mining
4.	WEKA	1993	GNU General Public License	Java	Cross Platform	Machine Learning
5.	KEEL	2004	GNU GPL v3	Java	Cross Platform	Machine Learning
6.	R	1997	GNU General Public License	C, Fortran and R	Cross Platform	Statistical Computing

Table 2.1 Data Mining Tools Summary

2.6. WEKA Interfaces

WEKA (Waikato Environment for Knowledge Analysis) is a machine learning and data mining software tool written in Java and distributed under the GNU Public License. The goal of the WEKA project is to build a state-of-the-art facility for developing machine learning techniques and to apply them to real-world data mining problems. It contains several standard data mining techniques, including data preprocessing, classification, regression, clustering, and association. Although most users of WEKA are researchers and industrial scientists, it is also widely used for academic purposes [13,35].

Figure 2.5 WEKA interface

WEKA version 3.8.1 has five interfaces, which start from the main GUI Chooser window, as shown in figure 4.1 each interface has a specific purpose and utility. Whereas the Explorer and Knowledge Flow are tailored to beginning users, the experimenter, workbench and simple CLI target more advanced users. The buttons can be used to start the following applications:

- **Explorer:** An environment for exploring data with WEKA (the rest of this documentation deals with this application in more detail).
- **Experimenter:** An environment for performing experiments and conducting statistical tests between learning schemes.
- **Knowledge Flow:** This environment supports essentially the same functions as the Explorer but with a drag-and-drop interface. One advantage is that it supports incremental learning.

- **Workbench:** provides as a working area for creating a model same as experimenter.
- **Simple CLI:** Provides a simple command-line interface that allows direct execution of WEKA commands for operating systems that do not provide their own command line interface.

The Explorer is possibly the first interface that new users will run simulations in. It allows for data visualization and preprocessing. In this study we use Explorer environment to conduct the experiment.

2.7. Application of Data Mining in Telecommunication Sector

The telecommunications industry generates and stores a tremendous amount of data. These data include call detail data, which describes the calls that traverse the telecommunication networks, network data, which describes the state of the hardware and software components in the network, and customer data, which describes the telecommunication customers. The amount of data is so great that manual analysis of the data is difficult, if not impossible.

Globally, the development of telecommunication industry is one of the important indicators of social and economic development of a given country. In addition to this, the development of communication sector plays a vital role in the overall development of all sectors related to social, political and economic affairs. This sector is very dynamic in its nature of innovation and dissemination [6].

In telecommunication sector, data mining is applied for various purposes. Some of the Data mining applications in this sector and some research areas are discussed here:

2.7.1. Telecom Fraud Detection

Fraud is a serious problem for telecommunication companies, leading to revenue loss of billions of dollars each year [5]. Fraud can be divided into two categories: subscription fraud and superimposition fraud. Subscription fraud occurs when a customer opens an account with the intention of never paying for the account charges. Superimposition fraud involves a legitimate account with some legitimate activity, but also includes some “superimposed” illegitimate activity by a person other than the account holder [17].

Fraud detection methods are continuously being developed to checkmate criminals who also adopt new strategies regularly. The development of new fraud detection methods is made more difficult due to the severe limitation imposed by restricted information flow about the outcome of fraud detection efforts. Fraud detection has been implemented by a number of methods such as data mining, statistics, and artificial intelligence [5,16,17].

According to Jember [17], there are different types of fraud. However, the most common types of fraud are Subscription Fraud, Subscriber Identity Module (SIM) Box or Bypass Fraud, Premium Rate Fraud (PRF) Internal Fraud, Prepaid Fraud, Postpaid Fraud, Cloning Fraud, Roaming Fraud, Private Branch Exchange (PBX) fraud. Subscription is among the most common frauds due to the low technical knowledge required to perform the fraud.

2.7.2. Marketing/Customer Profiling

Telecommunication companies maintain a great deal of data about their customers. In addition to the general customer data that most businesses collect, telecommunication companies also store call detail records, which precisely describe the calling behavior of each customer. This information can be used to profile the customers and these profiles can then be used for marketing and/or forecasting purposes [32].

According to Pradeep [7], a serious issue with telecommunication companies is customer churn. Customer churn involves a customer leaving one telecommunication company for another. Customer churn is a significant problem because of the associated loss of revenue and the high cost of attracting new customers.

Data mining techniques now permit companies the ability to mine historical data in order to predict when a customer is likely to leave. These techniques typically utilize billing data, call detail data, subscription information (calling plan, features and contract expiration data) and customer information (e.g., age). Based on the induced model, the company can then take action, if desired.

2.7.3. Network Fault Isolation

Telecommunication networks are extremely complex configurations of hardware and software. Most of the network elements are capable of at least limited self-diagnosis and these elements may collectively generate millions of status and alarm messages each month.

In order to effectively manage the network, alarms must be analyzed automatically in order to identify network faults in a timely manner or before they occur and degrade network performance. A proactive response is essential to maintain the reliability of the network. Because of the volume of the data and a single fault may cause many different seemingly unrelated alarms to be generated, the task of network fault isolation is quite difficult. Data mining has a role to play in generating rules for identifying faults [32].

2.7.4. Event Log Analysis

Weiss [32] discussed the problems of event correlation and data mining in the context of event log analysis and presents novel tools and techniques for addressing the problems. Event logs play an important role in modern IT systems, since they are an excellent source of information for monitoring the system in real-time and for conducting retrospective event analysis. Event correlation is one of the most prominent event processing techniques today. It has received a lot of attention in the context of network fault management over the past decade and is becoming increasingly important in other domains as well, including event log analysis.

Data mining techniques are a common choice for knowledge discovery from event logs, and the mining of patterns from event logs has been identified as an important system and network management task. Recently proposed mining approaches for accomplishing this task have often been based on some well-known algorithm for mining frequent item sets, and they have focused on detecting frequent event type patterns.

2.8. Related Works

In this section, different works which are related with data mining for telecommunication sectors are presented. The papers presented are knowledge discovery from mobile network and different works in telecom sector using data mining techniques are selected and presented based on the relevance and similarities of the papers with this thesis work.

2.8.1. International Works

Nelson [44] studies the development of segmentation models that helps to acquire some knowledge about the clients' behavior according to specific characteristics, which can be grouped, allowing the company to define specific relationship actions for each of the identified groups, associating the distinct approaches with some specific characteristics. The analysis of the different

groups allows identifying the distribution of the main characteristics of each of them according to business perspectives, allowing the Company to develop market intelligence. This intelligence, which is extremely analytical, helps defining clearer and more objective business rules, allowing the Company to achieve better results with less cost and efforts. The classification models allow the Company, based on the knowledge obtained from identified grouping, to anticipate specific events, becoming more proactive and, consequently, more efficient in the business processes.

Nelson [44] stated that, the insolvent clients' behavior segmentation model Brazil Telecom to have more specific knowledge related to the insolvency scenario of the Company. This knowledge helps defining more focused actions against insolvency, as well as creating more efficient collecting procedures. The insolvency behavior segmentation model has identified five characteristic groups of clients, which were separated in distinct classes. The paper reports on the findings of a research project that had the objective to build a decision support system to handle customer insolvency for a large telecommunication company. Prediction of customer insolvency and with an accuracy that could make this prediction useful in business terms was one of the core objectives of the study. In the paper the process of building such a predictive model through knowledge discovery and data mining techniques in vast amounts of heterogeneous as well as noisy data is described. A cluster model was used based on an unsupervised learning, using Kohonen's self-organizing maps. Additionally, an MLP neural network was used for the predicting models.

According to Ogwueleka [19], user profiling and classification are important tasks in data intensive environments where computer assisted decision-making are sought for. The calling behavior is described by the subscriber's call data and was used in this study as a basis for modeling. The goal for the learning methods in this study was to learn user profiles from the call data in order to make decisions about fraud occurrence.

The methods presented in this study learn to detect fraud from partially labeled data, in that a call account was known to be defrauded but not exactly when. The data is therefore a mixture of legitimate and fraudulent data with an unknown mixing mechanism.

The models used in fraud detection were probabilistic models and neural networks. The ability to learn from data was considered an important asset of these models, and the capability to process uncertainty, which was present in the fraud domain. Modeling was performed on user level, user profile level and class level, of which the user profile level was seen to be the most appropriate.

Discriminative training was utilized for tuning the models for best diagnostic accuracy. Discriminative training was utilized for tuning the models for best diagnostic accuracy [19].

According to Kusaksizoglu [20], fraud is a significant source of lost revenue to the telecom industry. Efficient fraud detection and analysis systems can save telecom operators a lot of money. Automated fraud detection systems enable operators to respond to fraud by detection, service denial and prosecutions against fraud. In this study, the objective was to examine the call detail records (CDRs), demographic data and payment data of mobile subscribers in order to develop models of normal and fraudulent behavior via data mining techniques.

First, they have done some Exploratory Data Analysis (EDA) on the data set and discovered that some variables like Account length, Package type, Gender, Type, Total Charged Amount showed important tendency for fraudulent use. They applied k-means cluster method to cluster the customer, based on their call behaviors. Standard variables with ranked attributes and variables obtained from factor analysis due to some correlated variables were used as two different set of variables.

Finally, they experimented data mining techniques such as Decision trees, Rule based methods, and Neural Networks. They discussed the collected results based on performance measures such as accuracy, sensitivity, specificity and precision [20]. In their experimental results, random forest decision tree algorithm gives better accuracy and precision of 99.647% and 99. 785% respectively.

2.8.2. Local Works

Jember [17], conducted a research to explore the potential application of data mining in fraud detection on mobile communication service in the case of ethio telecom mobile data. According to the researcher, the application of data mining methods and tools to large quantities of data generated by the Call Detail Record (CDR) of telecommunication switch machine has been the art of the day to address the serious problems of telecommunication operators. The CDR consists of a vast volume of data set about each call made and it is a major resource of data for research works to find out hidden patterns of calls made by customers in addition to the typical use for bill processing activities.

The data used by the researcher was a three-month data of October and November of 2003 and March of 2004. The researcher used three months of data from October and November of 2003

and March of 2004. From the total of 9153 customers, 900 customers who made an international call of more than one minute per call were selected using stratified sampling technique in order to get representative data from all the postpaid mobile subscribers. The methodology used had three basic steps, data collection, data preparation, model building and testing. MATLAB tool and neural network data mining technology were employed to build the models from which an accuracy of 89% was achieved.

The research selected the attributes Minimum number of calls, Maximum number of calls, Average number of calls, Standard deviation of number of calls, Total number of calls, Minimum duration, Minimum duration, Average duration, Standard deviation of duration and Total duration.

The research scope was limited to exploring the possibility of application of the DM technology to supply fraud detection in mobile communication network using artificial neural network of the postpaid mobile phone. The researcher recommended giving sufficient time to get representative samples of the data and sufficient time to build appropriate model to get the best performance [17].

Another study was done by Gebremeskel [21]. In this research work, an attempt had been made to assess the possible application of data mining technology to support mobile fraud detection on Ethio-Mobile Services. The data source of this study was taken from call detail record (CDR). The researcher used a two-month data from the month of February and March of 2005.

The methodology followed by Gebremeskel [21] included data collection, data preparation, and model building and testing. SPSS software tool and artificial neural network algorithm were used as data mining techniques. The researcher took 29,463 records as a sample and selected 9 attributes for that process. From these 29,463 call records, 9186 were identified as fraudulent calls. In general, the numbers of fraudulent call became 31.18 percent of the total sample size (total record). The following attributeswere selected during his experimentation. CallType, ChargeType, Roamflag, CallingPartyNumber, CalledPartyNumber, CallBeginTime, CallDuration, CallCost and CalledArealndicator.

Additionally, the study was targeted at prepaid mobile service. The number of post-paid mobile subscribers was 53614 and the prepaid ones were 381417. The researcher further recommended that additional research could be done by including other attributes of the call detail record so as to build models with better performance and better accuracy [21].

Dereje [42] discovered hidden knowledge from ethio telecom mobile network data specifically on GSM mobile so as to determine call setup success rate. According to the research, to overcome the drawback of simple statistical method they proposed data mining techniques, methods and methodologies and used in their research.

In order to discover knowledge from the data they use divisive hierarchical clustering method to cluster the data. The K-means algorithm, Waikato Environment for Knowledge Analysis (WEKA) tool and CRoss Industry Standard Process (CRISP) data mining process model was used during the research work.

Data preprocessing was done using different data mining methods to prepare the data for analysis. After data preprocessing, they found relevant attributes and clustering was conducted on the preprocessed data. During clustering, first they clustered the data set into two, then further cluster those clusters containing dissimilar instances. Finally, they got clusters with similar behavior to analyze and interpret.

As a result, they show the relationship between attributes against Call Setup Success Rate (CSSR). The data shows that most of the Call Setup are failed and categorized under Very Poor CSSR category. To enhance Call Setup Success Rate, emphasize should be given to attributes used as KPIs. The research result reveals which attribute should enhance to improve the call setup success rate. Enhancing CSSR leads giving (QoS) to customers and it implies customer satisfaction and increases company revenue [42].

In summary, during the review of related works, most of the researches are focused on customer classification, customer churn prediction and telecom fraud detection. Hence, there is a gap to study the data generated from telecommunications mobile network to discover patterns that determine call success and drops. The output of this research gives new patterns, new information and new insight to the network which helps to improve the quality of service provided, the performance of the network, optimization and expansion of the network. The improvement leads to customer satisfaction and revenue maximization. This research also fills the gap by creating a predictive model using data mining classification algorithms to discover patterns determining the reason for call drops.

CHAPTER THREE

PROBLEM UNDERSTANDING AND DATA PREPARATION

In this chapter, problem understanding, data understanding and preparation of the raw data to be suitable for analysis are discussed in detail. In this research Hybrid Data Mining methodology is used to understand the problem and the data as well as to prepare the data for analysis.

3.1. Understanding of the problem domain

In this section, we tried to understand the domain to set the data mining goal and the objective of this research as per the real situation of ethio telecom working area and as per the domain perspective.

Telecommunication Company worldwide suffers from calls which are dropped without reaching the end users. The estimated losses amount to several billions of dollars due to this call drops. Since these call drops highly affect the telecom operator's revenue, it is still a significant loss. The mobile telecommunication industry stores and generates tremendous amounts of raw and heterogeneous data that provides rich fields for analysis.

During this research we study the domain and how the domain experts analyze mobile network elements data. They are doing the analysis using simple statistical tools. The output from the analysis is used to enhance the quality of the service and to take appropriate actions to deliver the service to its customers as well as sent as a report for higher officials for further decisions. It has its own drawback to analyze huge amount of data using simple statistical tools.

Providing a Quality of Service based on current network status information is the success criteria for ethio telecom. If the information used is not reliable the decision given is also unreliable.

In order to have detail understanding and knowledge about the problem domain, we discuss with domain experts based on the discussion points as shown below concerning the telecom mobile call drop reasons.

- How do you explain mobile call drop in telecom sector?
- Who and why people commit mobile call drops?

- Which attributes are call drop reason indicators from the fields of Fault Management System (FMS) data?
- How can we identify and prevent mobile call drops?
- What is the current practice to prevent mobile call drops?

According to the domain experts of ethio telecom and documents, the Ethiopian Government has decided to transform the telecommunication infrastructure and services to world class standard. Thus, ethio telecom is born from this ambition in order to bring about a paradigm shift in the development of the telecom sector to support the steady growth of our country.

The vision of the company is to be a world-class telecommunications service provider so that the following points are the mission of the company [4]:

- Connect every Ethiopian through Information Communication Technology.
- Provide telecommunication services and products that enhance the development of our Nation.
- Build reputable brand known for its customers' consideration.
- Build its managerial capability that enables to operate an international standard.

In line with its ambitious mission, ethio telecom has the following ambitious goals [4]:

- Being a customer centric company
- Offering the best quality of services
- Meeting world-class standards
- Building a financially sound company

To deliver different services to its customers, there are projects to cover 85% of the country through land lines and wireless communication technologies. Whereas to deliver quality of services (QoS) to the customer by conducting different analyses there are situations to upgrade the service, optimization of the network elements and new installation of network equipment for existing infrastructures are on progress.

Generally, we identify key people at this domain from ethio telecom and focuses on understanding the research objective and requirements from a business perspective. Then we convert this knowledge into a DM problem and a preliminary plan is designed to achieve the objective and learn about current solutions to the problem domain.

3.1.1. Mobile Architecture

Mobile communications are an open, digital cellular technology used for transmitting mobile voice and data services. The mobile telecommunications technical specifications define the different elements within the mobile network architecture and the ways in which they interact to enable the overall network operation to be maintained [24,25].

As shown in Figure 3.1, mobile network can be divided into the following broad parts [27]:

Mobile Station (MS), Base Station Subsystem (BSS), Network Switching Subsystem (NSS) and Operation Support Subsystem (OSS).

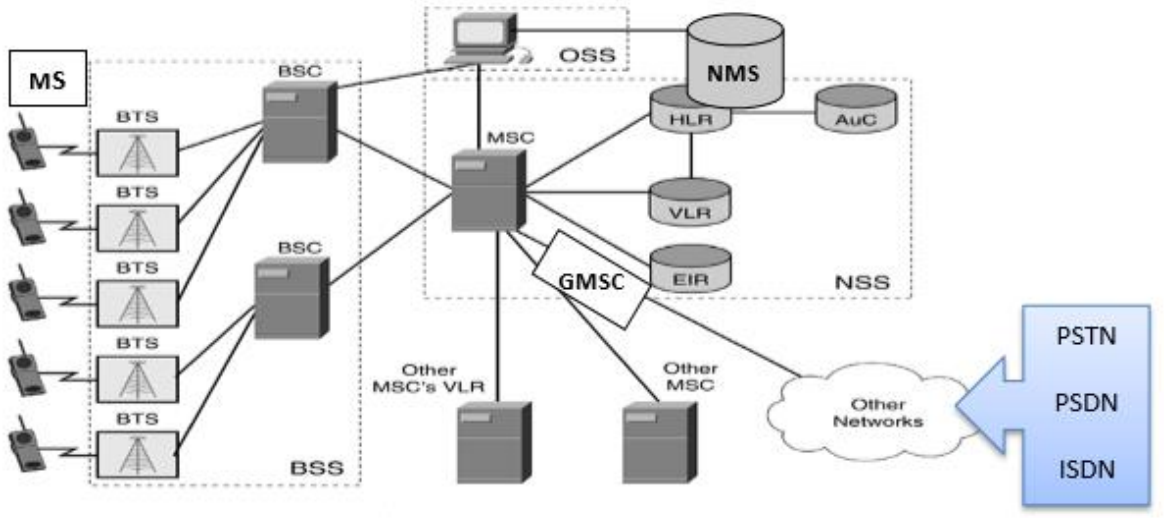


Figure 3.1 Mobile Architecture [27]

- ***The Mobile Station (MS):*** Mobile Equipment (ME) or as they are most widely known, cell or mobile phones and Subscriber Identity Module (SIM) are the section of a mobile cellular network that the user operates and get service. The two users which are the caller and receiver are communicated by using their SIM cards inside the ME.

- ***The Base Station Subsystem (BSS):*** provides the interface between the ME and the Network Switching Subsystem (NSS). It is in charge of the transmission and reception. It may be divided into two parts:
    - Base Station Controller (BSC): it controls a group of Base Transceiver Stations (BTSs) and manages their radio resources. A BSC is principally in charge of handover or handoffs, frequency hopping, exchange functions and power control over each managed Base Transceiver Stations.
    - Base Transceiver Station (BTS) or Base Station: it maps to transceivers and antennas used in each cell of the network. It is usually placed in the center of a cell. Its transmitting power defines the size of a cell.
- ***The Network Switching Subsystem (NSS):*** The mobile network subsystem contains a variety of different elements and is often termed the core network. It provides the main control and interfacing for the whole mobile network.

The major elements within the core network include:

- ***Mobile Switching Center (MSC):*** the central component of the Network Subsystem. The MSC performs the switching of calls between the mobile and other fixed network (such as the PSTN or ISDN) or mobile network users, as well as the management of mobile services such as registration, authentication, location updating, handovers, and call routing to a roaming subscriber. It also performs such functions as toll ticketing, network interfacing, common channel signaling, and others.
- ***Home Locator Register (HLR):*** is a database used for storage and management of subscriptions. The HLR is considered the most important database, as it stores permanent data about subscribers, including a subscriber's service profile, location information, and activity status.
- ***Visitor Location Register (VLR):*** is a database that contains selected information from the HLR and stores information about all the subscribers that are registered (visiting) in that MSC service area at the moment that enables the selected services for the individual subscriber to be provided.
- ***Authentication Center (AuC):*** provides authentication and encryption parameters that verify the user's identity and ensure the confidentiality of each call.

- ***Gateway Mobile Switching Centre (GMSC):*** A gateway between the MSC and other networks. The GMSC is responsible for routing the call.
- ***Equipment Identity Register (EIR):*** The EIR is a database that contains information about the identity of mobile equipment that prevents calls from stolen, unauthorized, or defective mobile stations.

- ***Operation Support Subsystem (OSS):*** is the functional entity from which the network operator monitors and controls the system. The purpose of OSS is to provide network overview and to offer the customer cost-effective support for centralized, regional and local operational and maintenance activities that are required for a mobile network. OSS contains Network Management System database which stores the data generated from network elements.

The above mobile Network architecture elements are used to support mobile services. One of the operations performed by mobile network elements is call set-up. There are two types of call set-up: call from MS (Mobile Originated) and call to MS (Mobile terminated) [24,25,27].

3.1.2. Mobile Call Drops

The drop-call probability is one of the most important quality of service indexes for monitoring performance of cellular networks. For this reason, mobile phone operators apply many optimization procedures on several service aspects for its reduction. As an example, they maximize service coverage area and network usage; or they try to minimize interference and congestion; or they exploit traffic balancing among different frequency layers [27].

In telecommunications, the dropped-call rate (CDR) is the fraction of telephone calls due to technical reasons. Call drop is a situation where calls were cut off before the speaking parties had finished their conversational tone and before one of them had hung up (dropped calls). This fraction is usually measured as a percentage of all calls. A call attempt invokes a call setup procedure, which, if successful, results in a connected call. A connected call may be terminated (disconnected) due to a technical reason before the parties making the call would wish to do so (in ordinary phone calls this would mean before either of the parties has hung up). Such calls are classified as dropped calls [27].

In many practical cases this definition needs to be further expanded with a number of detailed specifications describing which calls exactly are counted as dropped, at what stage of the call setup procedure a call is counted as connected. In modern telecommunication systems, such as cellular (mobile) networks, the call setup procedure may be very complex and the point at which a call is considered successfully connected may be defined in a number of ways, thus influencing the way the dropped-call rate is calculated [32].

3.1.3. What Causes Call Drops?

According to William Staling [27], the prime reasons for dropped calls are:

- **Inadequate coverage** which can be due to: Lack of tower infrastructure, improper network planning and non-optimization of network. Resources like receiving and transmitting number of antennas, specialized human skills highly affects the calls to be dropped.
- **Overloaded cell towers –** number of subscribers are growing day by day and most of them are on smartphones. The network capacity is simply not being ramped up at the same pace resulting in overloaded networks. The number of customers is directly related to the reason for call drops in that when there is high number of customers in a specific area, the transmission tower becomes congested and the towers become unable to receive additional users. So, the calls which initiated to this tower are dropped.
- **Cityscape changes -** there have been instances where a new multistoried building comes up & the adjacent building's subscribers lose cell reception. Such instances are very common with rapidly changing cityscapes and call for routine network data analysis from service providers. Some zones in Addis Ababa, there are high number of big buildings which masks the transmission towers not to transfer signals with each other. So, these building are also the major reasons to drop the calls.
- **Switching between towers –** this situation occurs when a person is traveling or moving around while talking. If a call handover takes place from one BTS to another, especially in case of overloaded networks, there are chances of dropped calls.
- **Technical Failures –** this is beyond anyone's control and operators generally monitor downtimes through well-equipped network operation centers, can be due to transmission path or power outages. Some of the reason for technical problems include seasonal conditions. In Ethiopia, broadly speaking there are three seasons. The short rains season known as Belg, runs

from February to May. This is then followed by the long rains season known as Kiremt, which is between June and Mid-September. The Bega typically occurs between October and January and is characterized by generally dry weather over the bulk of the country, with wet weather and a secondary peak in rainfall over the south. Natural effects like flooding and thunderstorm happens based on this season. So, this natural effect is one of the reasons to drop the call.

Based on the assessment of existing problems in ethio telecom mobile network data, the main call drop reasons are ETHIO POWER, EEU POWER and FIBER&TRANSMISSION. On this assessment, the following attributes are considered useful for identifying mobile call drop reasons.

1. LOCATION: identifies where the device is located. The device can be located East Addis Ababa Zone (EAAZ), West Addis Ababa Zone (WAAZ), North Addis Ababa Zone (NAAZ) or South Addis Ababa Zone (SAAZ),
2. DG STATUS: identifies whether the device has its own generator or not. In ethio telecom, currently there are some devices which have their own standby generator which operates in the condition when the power is off. Some of the device have no any generator.
3. TECHNOLOGY: identifies the technology of the down sites whether it is 2G, 3G or 4G.
4. RESOURCES: identifies the capacity of the available resources whether it has poor resources, good resources or excellent resources. The capacity of the available resources has inverse relation with call drop rates. This means, when there are high resources available, the call drops rates become decreases and vice versa.
5. SEASONS: identifies the weather condition of the environments. As discussed above, In Ethiopia, broadly speaking there are three seasons which are kiremnt, belg and bega.
6. CONGESTION_TYPE: identifies how much the devices are congested. It can be highly congested, medium or critically congested. The congestion type has direct relationship with the call drop. This means, when the device is critically congested, the call drops increases and vice versa.
7. CUSOMER_NUMBER: identifies how much users are there in a specific area. Based on the real work situation and domain expert justification, customer numbers are registered in the system as low, medium and high amount of customer numbers.
8. RCA CATEGORY: this attribute is a class used to classify call drop reasons. In this research, call drop reasons can be due to ETHIO POWER, EEU POWER or FIBER & TRANSMISSION.

3.2. Understanding of the Data

Next to identifying the problem and building a simple plan for solving the problem, we proceed to the central item in data mining process which is data understanding. This includes listing out attributes with their respective values and evaluation of their importance for this research and careful analysis of the data and its structure is done together with domain experts by evaluating the relationships of the data with the problem at hand and the particular DM tasks to be performed. Finally, we verify the usefulness of the data with respect to the DM goals.

3.2.1. Data Collection

The telecommunications industry generates and stores a tremendous amount of data during giving telecom services. These data include call detail data, which describes the calls that traverses the telecommunication networks, network data, which describes the state of hardware and software components in the network and customer data, which describe telecom customers.

Call detail data is generated when every user makes a call and each calls detail information stored in the database. The network data is generated from all network elements about network element status, call setup information, and the like. Customer data includes detail customer information and customer location information [7,32].

These huge amounts of data should be handled properly for different purposes like fraud detection, network performance analysis, customer churn prediction, reporting for higher officials, for network planning and optimization and to support decision making [7,32].

From the above telecommunications data, network data is used for this research. The database of this network data was manipulated by using oracle software system. The first and the major source of data was network alarm data's which are sent from each network elements to the Fault Management System during failures.

High volume of data generated from different network elements and stored into different databases for different purposes [7,32]. Those databases are:

- Home Location Register (HLR) database contains customer profile, having all the detail, like customer ID, customer number, billing detail and for prepaid with intelligent network it has detail of current recharge.

- Visitor Location Register (VLR) database(s) stores updated user location and contains a copy of most of the data stored at the HLR. It is, however, temporary data which exists for only as long as the subscriber is active in the particular area covered by the VLR. This database will therefore contain some duplicate data as well as more precise data relevant to the subscriber remaining within the VLR coverage or location area.
- Fault Management System (FMS) stores the whole network element status and network operations data. It includes all call setup data and alarm indication data.

During this research we use the data stored in Fault Management System (FMS) database in a period of 2017 and half of 2018 (January - June). This data can be extracted to various file formats like Excel, PDF and CSV file type; for this research we used the data in Excel and CSV formats. These formats can be used directly or by converting the format to WEKA tool for analysis.

The data from the FMS server is only extracted by authorized personnel or domain expert. So, to get the data official permission is needed. Managing the huge FMS data was one of the challenges and time consuming because as the size of the data increases, preprocessing this huge amount data becomes more complex and tedious. After eliminating irrelevant and unnecessary data, a total of 16996 datasets are used for the purpose of conducting this study. We also select 8 attributes for this study based on their relevant for this research.

3.2.2. Description of the collected data

Description of the data is very important in data mining process in order to clearly understand the data. Without such an understanding, useful application cannot be developed. As indicated before this research is done by collecting the data from ethio telecom FMS mobile network server of 2017 and half of 2018 records.

In this section, from the source described above, the attributes with their data types and descriptions are shown in the following table 3.1 below. The attributes have different data types like date, string, nominal and numeric data type.

No	Attribute Name	Data Type	Description	Missing value
1.	Device Name	String	Name of the device which is down	0%
2.	Location	Nominal	Region of the device where it is located	0%
3.	DG Status	Nominal	Whether the device has generator or not	5%
4.	Technology	Nominal	The technology of the down site	3%
5.	Site ID	Numeric	Uniquely identifies the device name	0%
6.	Specific RCA	Nominal	Specific down cause of the device	12%
7.	RCA SUB Category	String	Sub category of specific RCA	8%
8.	Occurrence Time	Date	when the device is down	0%
9.	Cleared Time	Date	when the device is up	0%
10.	Acknowledge Operator	String	who acknowledged for the down site	0%
11.	Assignment Operator	String	who assigned to repair the down site	20%
12.	Acknowledge Time	Date	when the site acknowledges	0%
13.	TT ID	Alpha Numeric	Trouble Ticket ID for individual devices	0%
14.	Duration	Numeric	The total time it takes to recover (in days)	0%
15.	RCA Category	String	The main reason for the device to down	0%
16.	Comment Text	String	Individual comment on each alarm	25%
17.	Specific RCA Old	Nominal	The previous root cause of the alarm	8%
18.	Severity	Nominal	How the alarms are critical	0%
19.	Alarm Name	String	The name of sources of each alarm	6%
20.	Alarm Source	Nominal	Where the alarms are originated	0%
21.	NE Type	String	Network element types	15%
22.	Vendor	String	The technology vendor	0%
23.	Season	Nominal	Seasonal conditions	0%
24.	Customer_Number	Nominal	How much user are there in specific area	0%
25.	Alarm Type	String	Type of the alarms generated	100%
26.	Index	Numeric	Increasing counting number	0%

Table 3.1 All attributes with their Description of mobile call drop reasons

3.3. Preparation of the Data (Data Pre-processing)

Today's real-world databases are highly susceptible to noisy, missing and inconsistent data due to their typically huge size and their likely origin from multiple, heterogeneous sources. Low quality data will lead to low-quality mining results [26,45]. Hence, Data preprocessing is required to have a data set which is suitable for analysis.

Preprocessing of the data in preparation for classification and prediction 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 [26].

The purpose of data preprocessing is to clean selected data for better quality. Data quality is a multifaceted issue that represents one of the biggest challenges for data mining. It refers to the accuracy and completeness of the data. Data quality can also be affected by the structure and consistency of the data being analyzed. The presence of duplicate records, the lack of data standards, the timeliness of updates and human error can significantly impact the effectiveness of the more complex data mining techniques, which are sensitive to understated differences that may exist in the data. To improve data quality, it is sometimes necessary to clean the data, which can involve the removal of duplicate records, normalizing the values used to represent information in the database [26,45].

Some selected data may have different formats because the data is very huge, and they stored the data in damp files. Then, in ordered to use the data it needs to convert in to suitable format. Because the purpose of the preprocessing stage is to cleanse the data as much as possible and to put it into a format that is suitable for use in later stages.

3.3.1. Data selection

This phase uses to create a target dataset. The whole target dataset may not be taken for the DM task. Irrelevant or unnecessary data are eliminated from the DM database before starting the actual DM function. For this research, from the year 2017 and half of 2018 (January - June) 16996 records are used. Since the dataset contains irrelevant and unnecessary data, the 16996 records are selected after eliminating irrelevant and unnecessary data for the purpose of conducting this study.

In this phase, we also remove attributes which have no correlation with call drop reasons as shown in table 3.2 below.

No.	Attribute Name	Attribute data type	Reason for Removal
1.	Vendor	String	Have no correlation with call drop reason
2.	Device Name	String	Have no correlation with call drop reason
3.	Cleared Time	Date	Have no correlation with call drop reason
4.	Acknowledge Operator	String	Have no correlation with call drop reason
5.	Acknowledge Time	Date	Have no correlation with call drop reason
6.	TT ID	Alpha Numeric	Have no correlation with call drop reason
7.	Duration	Numeric	Have no correlation with call drop reason

Table 3.2 Reduced Attributes which have no correlation with call drop reasons

Accordingly, DEVICE NAME, SITE ID, EXP SPECIFIC RCA, COMMENT TEXT, SPECIFIC RCA OLD, LEN, OD, CAL_OCC_DATE_TIME, CAL_CLEAR_DATE_TIME, CLEARED TIME, SEVERITY, ALARM NAME, ALARM SOURCE, NE TYPE and VENDOR attributes was not important for this research because they didn't give any input for call drops reason. So, we delete the whole column since it is meaningless to use those attributes.

The importance of reducing the number of attributes not only speed up the learning process, but also prevents most of the learning algorithms from getting fooled into generating an inferior model by the presence of many irrelevant or redundant attributes [43]. So that very limited numbers of attributes that are most important for the study at hand are selected because of this reason. In order to select the best attributes from this initial collected dataset, we evaluate the information content of the attributes with the help of real working situations, documents and domain expert.

Hence, together with the domain expert and real work situations we remove those attributes which have less important for this research and the remain attributes which are shown in the table 3.3 are the final list of attributes that have been used in this study.

In summary, by applying Data Selection preprocessing method in our data, 16996 datasets and 8 attributes (including the class attribute) are selected for analyses which are important to reason out the call drops.

No	Attribute Name	Data Type	Description	Data Values
1.	Location	Nominal	Region of the device which it is located	EAAZ, WAAZ, NAAZ, SAAZ
2.	DG Status	Nominal	Whether the device has its own generator or not	DG, NO DG
3.	Technology	Nominal	The technology of the down site	2G, 3G, 4G
4.	Resources *(derived)*	Nominal	The capacity of resources	POOR, GOOD, EXCELLENT
5.	Season	Nominal	Seasonal conditions	KIREMT, BELG, BEGA
6.	Congestion_Type *(derived)*	Nominal	How much the sites are congested	CRITICAL, MAJOR, MINOR
7.	Customer_Number	Nominal	How much user are there in specific area	LOW, MEDIUM, HIGH
8.	RCA Category *(class attribute)*	Nominal	Shows the main reason for the device to down	ETHIOPOWER, EEUPOWER, FIBER&TRANSMISSION

Table 3.3 Final list of attributes used in this research

In this research, based on the assessment of existing situation and discussion with domain experts, we derived two attributes (Resources and Congestion_Type) which are necessary and have direct relationship with call drop reasons.

The attribute resources are derived from location and number of customers attributes in a concept that if there are locations where there is high number of customers, there must be sufficient amount of resources for this location otherwise the call become dropped. So, amount of resources has direct relationship with call drop reasons.

The other derived attribute was Congestion_Type. It also derived from number of customers and resources in a concept that if there are high number of customers and do not have enough resources, the calls become dropped. So, Congestion_Type has also direct relationship with call drop reasons.

As discussed in section 3.1.4, RCA CATEGORY is a class attribute which identifies the reason for each mobile call drops. The descriptive statistics of selected variables for each class are shown in table 3.4 below.

No	Attributes		Number of records for each class		
			EEU POWER	**ETHIO POWER**	**FIBER&TRANS MISSION**
			7851 (46.2%)	**5483 (32.26%)**	**3662 (21.54%)**
1.	Location	EAAZ	977	1850	1401
		WAAZ	3024	864	0
		NAAZ	1304	1175	2261
		SAAZ	2546	1594	0
2.	DG Status	DG	1655	4629	2170
		NO DG	6196	854	1492
3.	Technology	2G	2446	2659	1628
		3G	3871	1996	1797
		4G	1534	828	237
4.	Resources	POOR	1079	1880	3662
		GOOD	3698	2407	0
		EXCELLENT	3074	1196	0
5.	Season	KIREMT	2144	2347	0
		BELG	3543	1842	2834
		BEGA	2164	1294	828
6.	Congestion_Type	CRITICAL	1079	1880	3662
		MAJOR	3698	2407	0
		MINOR	3074	1196	0
7.	Costomer_Number	HIGH	215	1923	3662
		MEDIUM	4734	1343	0
		LOW	2902	2217	0

Table 3.4 Descriptive statistics of the selected variables

3.3.2. Data cleaning

This phase is used for making sure that the data is free from different errors. Otherwise, different operations like removing or reducing noise by applying smoothing techniques, correcting missing values by replacing with the most commonly occurring value for that attribute are done [26].

The selected data are cleaned further by removing the records that had incomplete (invalid) data and/or missing values under each column. Removing such records was done as the records with this nature are few and their removal does not affect the entire dataset. Data cleaning is done by using MS-Excel 2013. Some of data cleaning preprocesses are discussed below.

- Remove attributes with no (zero) information:
  - Attribute with no information or irrelevant are removed as shown below in table 3.5.

No.	Attribute Name	Attribute Value	Reason for Removal
1.	Index	Numeric	Has no information
2.	Alarm Type	Numeric	Has no information
3.	Site ID	Numeric	Has no information

Table 3.5 Reduced Attributes with no Information

- Remove attributes which have the same value for all instances:
  - Attributes with the same values are removed as shown in in table 3.6. If the values are the same, these values cannot be used for comparison or interpret to get new information.

No.	Attribute Name	Attribute Value	Reason for Removal
1.	Severity	Text	They have the same Values for all instances
2.	Alarm Id	Numeric	They have the same Values for all instances
3.	Alarm Source	Text	They have the same Values for all instances

Table 3.6 Reduced Redundant Attributes

- Remove correlated attributes:
    - Table 3.7 below shows removed attributes which are highly correlated with each other using dimensionality reduction.

No.	Attribute Name	Attribute Value	Reason for Removal	Removed Attribute
1.	Specific RCA old & RCA Category	Text	They have correlation with each other	Specific RCA old
2.	Occurrence Time & Season	Date/Text	They have correlation with each other	Occurrence Time
3.	Specific Zone & Location	Text	They have correlation with each other	Specific Zone

Table 3.7 Reduced Attributes which are highly correlated with each other

3.3.3. Data formatting

The datasets provided to WEKA were prepared in a format that is acceptable by WEKA tools. It accepts records whose attribute values are separated by commas and saved in an ARFF (Attribute Relation File Format). The excel file was first changed into a comma delimited (CSV) file format. Next the file was saved with ARFF file extension. Now the dataset, which is in ARFF format, is ready to be used in the WEKA software.

```
LOCATION,RESOURCES,DGSTATUS,TECHNOLOGY,SEASON,CONGESTIONTYPE,COST
OMERNUMBER,RCA CATEGORY
EAAZ,POOR,DG,GU,KIREMT,CRITICAL,HIGH,EEU POWER
EAAZ,POOR,DG,GU,KIREMT,CRITICAL,HIGH,EEU POWER
EAAZ,POOR,DG,GU,KIREMT,CRITICAL,HIGH,EEU POWER
EAAZ,POOR,DG,GU,KIREMT,CRITICAL,HIGH,EEU POWER
EAAZ,POOR,DG,GU,KIREMT,CRITICAL,HIGH,EEU POWER
EAAZ,POOR,DG,GU,KIREMT,CRITICAL,HIGH,EEU POWER
EAAZ,POOR,DG,GU,KIREMT,CRITICAL,HIGH,EEU POWER
EAAZ,POOR,DG,GU,KIREMT,CRITICAL,HIGH,EEU POWER
EAAZ,POOR,DG,GU,KIREMT,CRITICAL,HIGH,EEU POWER
EAAZ,POOR,DG,GU,KIREMT,CRITICAL,HIGH,EEU POWER
EAAZ,POOR,DG,GU,KIREMT,CRITICAL,HIGH,EEU POWER
EAAZ,POOR,DG,GU,KIREMT,CRITICAL,HIGH,EEU POWER
EAAZ,POOR,DG,GU,KIREMT,CRITICAL,HIGH,EEU POWER
EAAZ,POOR,DG,GU,KIREMT,CRITICAL,HIGH,EEU POWER
```

Figure 3.2 Sample CSV format data sets prepared for WEKA

CHAPTER FOUR: EXPERIMENTATION AND MODELING

This chapter discusses the experiments, experimental results and model building conducted during the research. As stated in previous chapters, the domain experts in ethio telecom have been doing the analysis using simple statistical method on the data which is extracted from FMS database server. This simple statistical method is unable to utilize the whole data and reach at good analysis result. To overcome this problem, we propose the data mining techniques, algorithms and methods. During the experimentation we discover hidden knowledge from the data extracted from FMS database server which enables to enhance the call success rate.

4.1. Model Building

Modeling is one of the major tasks undertaken under the phase of data mining in hybrid process model. In this phase, different techniques can be employed for the data mining problems. Some of the tasks include: selecting the modeling technique, experimental setup, building a model and evaluating the model.

The output of a series of experiments of classification models are analyzed and evaluated in terms of the details of the confusion matrix of the model. Furthermore, models of the different classification algorithms, such as decision tree and rule induction were compared with respect to their performance measures such as Precision, Recall, F-measure and accuracy.

4.1.1. Selecting Modeling Technique

In this research, the supervised classification techniques are adopted. Selecting appropriate model depends on data mining goals. Consequently, to attain the objectives of this research, four classification techniques have been selected for model building. The analysis was performed using WEKA environment. Among the different available classification algorithms in WEKA, J48 and Random forest algorithms from decision tree as well as, PART and JRIP algorithms from rule induction are selected for experimentation of this study. In this work an attempt was done to build a model using selected algorithms for classification of call drop reasons.

Firstly, the J48 decision tree algorithm is chosen because it is one of the most common decision tree algorithms that are used today to implement classification techniques using WEKA.

Second, Random Forest (RF) is an ensemble classifier that consists of many decision trees and outputs the class that is the mode of the class's output by individual trees and because of the following advantages no need for pruning trees, Accuracy and variable importance generated automatically, Overfitting is not a problem, not very sensitive to outliers in training data and easy to set parameters [15].

Third, PART is selected similarly as of J48 and simplicity and generated list of rules easy to interpret and associate the problem domain. Finally, the applying JRIP rule induction techniques because it is one of the basic and most popular rule induction algorithms. Classes are examined in increasing size and an initial set of rules for the class is generated using incremental reduced error.

4.2. Experiment Design

The model is built based on the default 66% percentage split and 10-fold cross validation. The default ratio is 66% for training and 34% for testing. In 10-fold cross validation, the initial data are randomly partitioned into 10 mutually exclusive subsets or folds, 1,2,3,4 ...10, each approximately equal size. The training and testing are performed 10 times. In the first iteration, the first fold is reserved as a test set, and the remaining 9 folds are collectively used to train the classifier.

Train classifier on folds: 2 3 4 5 6 7 8 9 10; Test against fold: 1

Train classifier on folds: 1 3 4 5 6 7 8 9 10; Test against fold: 2

Train classifier on folds: 1 2 4 5 6 7 8 9 10; Test against fold: 3

Train classifier on folds: 1 2 3 5 6 7 8 9 10; Test against fold: 4

Train classifier on folds: 1 2 3 4 6 7 8 9 10; Test against fold: 5

Train classifier on folds: 1 2 3 4 5 7 8 9 10; Test against fold: 6

Train classifier on folds: 1 2 3 4 5 6 8 9 10; Test against fold: 7

Train classifier on folds: 1 2 3 4 5 6 7 9 10; Test against fold: 8

Train classifier on folds: 1 2 3 4 5 6 7 8 10; Test against fold: 9

Train classifier on folds: 1 2 3 4 5 6 7 8 9; Test against fold: 10

The accuracy estimate is the overall number of correct classifications from the 10 iterations divided by the total number of samples in the initial dataset [18].

Generally, a procedure or mechanism was used to test the model's quality and validity is needed to be set before the model is actually built. In order to perform the model building process of this study, we use 16996 datasets with 8 attributes to investigate the model. The training and testing dataset were prepared by purposive sampling technique from the original dataset. The original dataset is presented in annex 1.

In this study, we perform eight experiments with decision tree and rule induction algorithms. Since this research is experimental research design, we use J48 and Random Forest algorithms from decision tree and PART and JRIP algorithms from rule induction. In order to validate and compare the classification performance of the techniques the 10-fold cross validation and percentage split are used. Both methods are tested with default value by using 10-fold cross validation and 66% percentage split.

4.3. J48 Decision tree model building

A decision tree is a classifier expressed as a recursive partition of the instance space. 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 frequent case, each test considers a single attribute, such that the instance space is partitioned according to the attribute value.

Generally, this research is more interested in generating rules that best predict the call drop reasons and to come to an understanding of the most important factors (variables) affecting the call drops.

One of the classification techniques applied for building the classification model in this thesis is the J48 algorithm. As shown in the figure 4.2, J48 is one of the most common decision tree algorithms that are used today to implement classification technique using WEKA. This algorithm is implemented by modifying parameters such as confidence factor, pruning and up running, hanging the generalized binary split decision classification and other option available in table 4.1. Therefore, it is very crucial to understand the available options to experiment the algorithms, as it can make a significance difference in the quality of the result.

In many cases, the default setting will prove adequate, but to compare results or models and attain the research objectives other options are considered [6].

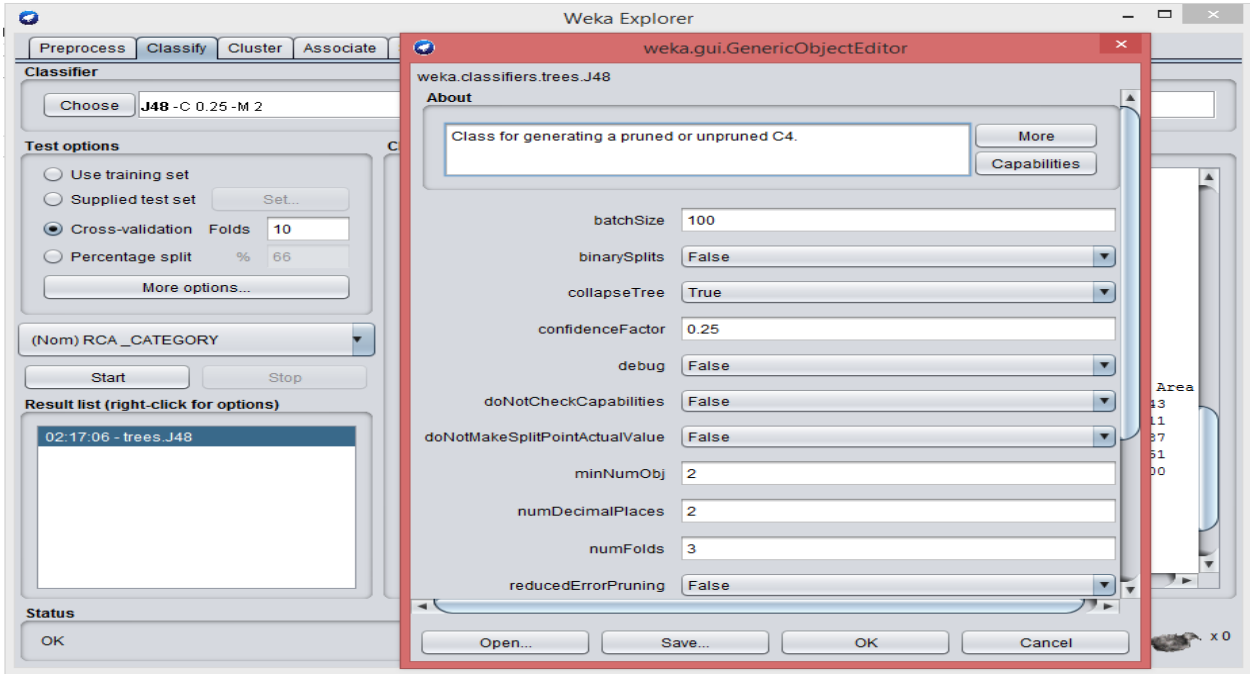


Figure 4.2 Snapshot of J48 algorithms setting

Some of the major parameters and default values of J48 decision tree algorithm are presented in table 4.1 below.

Name	Default Value	Possible Value	Description
Confidence Factor	0.25	10-0.5	The confidence factor used for pruning (smaller values incur more pruning)
minNumObj	2	1,2	The minimum number of instances per leaf
Unpruned	False	True/False	Use unpruned tree (the default value “False” means that the tree is pruned).
Sub tree raising	True	True/False	Whether to consider the sub tree raising operation in post pruning.
B-use binary splits	True	True/False	Whether to use binary splits on nominal attributes when building the tree.

Table 4.1 Some of the J48 algorithm parameters and their default values

Experiment 1:

The first experimentation is performed with the J48 default parameters of 10-fold cross validation test option. In 10-fold cross-validation, the initial data are randomly partitioned into 10 mutually exclusive subsets or "folds," 1, 2, 3, …, 10, each approximately equal size. Training and testing are performed 10 times. In the first iteration, the first fold is reserved as a test set, and the remaining 9 folds are collectively used to train the classifier; the classifier of the second iteration is trained on folds 1, 3, 4, …, 10 and tested on the second fold; and so on.

Using these J48 default parameters the classification model was developed. The experiment has been produced a decision tree which have 55 numbers of leaves and 80 tree size. Table 4.2 shows the resulting performance measurement of J48 algorithm with 10-fold cross validation of the model. Running information of J48 algorithm with 10-fold validation technique is provided on annex-2.

Algorithm	Test Option	Accuracy	Time	Recall	precision	F-Measure	Class
J48	10-fold cross-validation	95.43%	0.06 sec	90.5%	95.6%	93.0%	EEU
				95.8%	91.0%	93.3%	ETHIO
				100%	100%	100%	FB&TR
Weighted Average		***95.43%***	***0.06 sec***	***95.4%***	***95.5%***	***95.4%***	

Table 4.2 Performance result for J48 algorithm with 10-fold cross validation

As shown from the resulting performance measurements in table 4.2, the J48 learning algorithm scored an accuracy of 95.43%. This result shows that out of the total training datasets 16218 (95.43%) records are correctly classified instances, while 777 (4.57%) of the records are incorrectly classified.

Experiment 2:

This experiment is performed, by changing the default testing option of 10-fold cross validation to the percentage split (66%). In this learning scheme a percentage split is used to partition the dataset into training and testing data. With 66% of the data used for training and 34% used for testing.

The purpose of using this parameter was to assess the performance of the learning scheme by changing the 10-fold cross validation to the default value of the percentage split (66%) in order to test if the algorithm achieve a better classification accuracy than the first experimentation. The result of this learning scheme is summarized and presented in table 4.3.

Algorithm	Test Option	Accuracy	Time	Recall	precision	F-Measure	Class
J48	Percentage split (66%)	95.26%	0.00 sec	90.9%	94.8%	92.8%	EEU
				94.8%	91.0%	92.9%	ETHIO
				100%	100%	100%	FB&TR
Weighted Average		***95.26%***	***0.00 sec***	***95.3%***	***95.3%***	***95.3%***	

Table 4.3 Performance result for J48 algorithm with default percentage split (66%)

In this experiment out of 16996 total records, 11217 (66%) of the records are used for training purpose and 5779 (34%) of the records are used for testing purpose. As shown from the performance result in table 4.3, the model developed with this proportion, out of 5779 testing records 5504 (95.26%) of them are correctly classified instances and 274 (4.74%) records are incorrectly classified instances.

The J48 decision tree model with 10-fold cross validation has a better classification accuracy of 16218 (95.43%) which correctly classified and 777 (4.57%) that are wrongly classified. Again, the J48 decision tree model with default percentage split of 66% has 5504 testing records 95.26% of them are correctly classified instances and 274 (4.74%) records are incorrectly classified instances.

Generally, from the two experiments conducted before, the model developed with the 10-fold cross validation test option gives a better classification performance of identifying the reasons for call drops to their respective class category. Therefore, among the different decision tree models built in the foregoing experimentations, the first model, with 10-fold cross validation, has been chosen due to its better classification performances of 95.43% accuracy.

4.4. Random Forest model building

In this experiment random forest algorithm is explored. It builds an ensemble of several models instead of building a single one, and prediction of the ensemble model is made as a consensus of predictions made by all its individual members. The model is built based on the default values of 66% percentage split and 10-fold cross validation.

Experiment 3:

The third experiment is performed using random forest with 10-fold cross validation and the outcome of the experiment is presented in table 4.4. The overall performance of this test has lowered from what has scored in the previous experiment of J48 decision tree. The snapshot running information of random forest algorithm with 10-fold validation technique is provided on annex-3.

Algorithm	Test Option	Accuracy	Time	Recall	precision	F-Measure	Class
Random Forest	10-fold cross-validation	95.36%	1.31 sec	90.4%	95.6%	92.8%	EEU
				95.7%	90.9%	93.2%	ETHIO
				100%	100%	100%	FB&TR
Weighted Average		***95.36%***	***1.31 sec***	***95.4%***	***95.5%***	***95.4%***	

Table 4.4 Performance result for RF algorithm with 10-fold cross validation

As shown from table 4.4 Random Forest algorithm with 10-fold cross validation achieved an accuracy of 95.36%. This result shows that out of the total training datasets 160206 (95.36%) records are correctly classified instances, while 789 (4.64%) of the records are incorrectly classified instances.

Experiment 4:

This experiment is performed, by changing the default testing option of 10-fold cross validation to the percentage split (66%) and the outcome of this experiment is presented in table 4.5 below.

Algorithm	Test Option	Accuracy	Time	Recall	precision	F-Measure	Class
Random Forest	Percentage split (66%)	95.22%	0.22 sec	90.8%	94.8%	92.8%	EEU
				94.8%	90.9%	92.8%	ETHIO
				100%	100%	100%	FB&TR
Weighted Average		***95.22%***	***0.22 sec***	***95.2%***	***95.3%***	***95.2%***	

Table 4.5 Performance result of RF algorithm with default percentage split (66%)

As shown from the performance result of Random Forest algorithm with default percentage split of 66% scored an accuracy of 95.22%. This result shows that out of the total training datasets 5502 (95.22%) records are correctly classified instances, while 276 (4.78%) of the records are incorrectly classified instances.

The Random Forest (RF) algorithms with 10-fold cross validation scored an accuracy of 95.36% from the total training datasets 16206 (95.36%) records are correctly classified instances, while 789 (4.64%) of the records are incorrectly classified instances. The second of the Random Forest algorithm use the default of values of percentage split (66%) get the accuracy of 95.22%. This result shows that out of the total training datasets 5502 (95.22%) records are correctly classified instances, while 276 (4.78%) of the records are incorrectly classified instances.

Generally, when we compared from the two experiments conducted, that are the Random Forest algorithms with 10-fold cross validation and Random Forest algorithms with default percentage split of 66%, the model developed with the 10-fold cross validation gives a better classification accuracy of prediction than the default percentage split of 66%. So, the model constructed using 10-fold cross validation has been chosen due to its better classification performance than the model of 66% percentage split. Then the maximum accuracy of Random Forest algorithm is 95.36% with 10-fold cross validation.

4.5. PART Rule Induction model building

The third data mining algorithm applied in this research was PART Rule induction algorithm. PART algorithm is selected for the reason that it has the ability to produce accurate and easily interpretable rules that helps to achieve the research objectives and have the advantages of simpler and has been found to give sufficiently strong rules [28]. To build the Rule induction model, 16996 datasets was used as an input to the system. The test options used for the experiment are 10-fold cross validation and default value of the percentage split (66%).

Experiment 5:

The fifth experiment is performed using PART Rule induction algorithm with 10-fold cross validation and the outcome of this experiment is presented in table 4.6 below. The snapshot running information of PART with 10-fold validation technique is provided on annex 4.

Algorithm	Test Option	Accuracy	Time	Recall	precision	F-Measure	Class
PART	10-fold cross-validation	95.41%	0.11 sec	90.4%	95.6%	92.9%	EEU
				95.8%	90.9%	93.3%	ETHIO
				100%	100%	100%	FB&TR
Weighted Average		***95.41%***	***0.11 sec***	***95.4%***	***95.5%***	***95.4%***	

Table 4.6 Performance result for PART algorithm with 10-fold cross validation

As shown from the performance result of PART Rule induction algorithm with 10- fold cross validation scored an accuracy of 95.41%. This result shows that out of the total training datasets 16215 (95.41%) records are correctly classified instances, while 780 (4.59%) of the records are incorrectly classified instances.

Experiment 6:

This experiment is performed, by changing the default testing option of 10-fold cross validation to the percentage split (66%) and the outcome of this experiment is presented in table 4.7 below.

Algorithm	Test Option	Accuracy	Time	Recall	precision	F-Measure	Class
PART	Percentage split (66%)	95.26%	0.01 sec	90.9%	94.8%	92.8%	EEU
				94.8%	91.0%	92.9%	ETHIO
				100%	100%	100%	FB&TR
Weighted Average		***95.26%***	***0.01 sec***	***95.3%***	***95.3%***	***95.3%***	

Table 4.7 Performance result of PART algorithm with default percentage split (66%)

As shown from the performance result, the PART Rule induction algorithms with default percentage split (66%) scored an accuracy of 95.26%. This result shows that out of the total training datasets 5504 (95.26%) records are correctly classified instances, while 274 (4.74%) of the records are incorrectly classified instances. This shows that the first experiment conducted with the 10-fold cross validation algorithm is better performance than the second experiment of default percentage split 66%.

The PART Rule induction algorithms with 10-fold cross validation scored an accuracy of 95.41% out of the total training datasets 16215 (95.41%) records are correctly classified instances, while 780 (4.59%) of the records are incorrectly classified instances. The second experiment of the PART Rule induction algorithms with default values of percentage split (66%). From the total training datasets 5504 (95.26%) records are correctly classified instances, while 274 (4.74%) of the records are incorrectly classified instances.

Generally, when we compare the two experiments conducted before, the PART Rule induction algorithms with 10-fold cross validation and PART with default percentage split 66%, the model developed with the 10-fold cross validation gives a better classification accuracy of identifying newly arriving mobile call drops reason. So, the model constructed using 10-fold cross validation algorithm has been chosen due to its better classification performance of accuracy 95.41%.

4.6. JRIP Rule Induction model building

JRip (RIPPER) is one of the basic and most popular rule induction algorithms. Classes are examined in increasing size and an initial set of rules for the class is generated using incremental reduced error JRip (RIPPER) proceeds by treating all the examples of a particular judgment in the training data as a class and finding a set of rules that cover all the members of that class. Thereafter it proceeds to the next class and does the same, repeating this until all classes have been covered.

To build JRIP Rule induction model, 16996 datasets was used as an input to the system. The test options used for this experiment are 10-fold cross validation and default value of the percentage split (66%).

Experiment 7:

This experiment is performed using JRIP Rule induction algorithm with 10-fold cross validation and the outcome of this experiment is presented in table 4.8 below. The snapshot running information of JRIP with 10-fold validation technique is provided on annex 5.

Algorithm	Test Option	Accuracy	Time	Recall	precision	F-Measure	Class
JRIP	10-fold cross-validation	95.26%	2.27 sec	90.2%	95.4%	92.7%	EEU
				95.6%	90.8%	93.1%	ETHIO
				100%	100%	100%	FB&TR
Weighted Average		***95.26%***	***2.27 sec***	***95.3%***	***95.3%***	***95.3%***	

Table 4.8 Performance result for JRIP algorithm with 10-fold cross validation

As shown from the performance result, the JRIP Rule induction algorithm with 10- fold cross validation scored an accuracy of 95.26%. This result shows that out of the total training datasets 16189 (95.26%) records are correctly classified instances and 806 (4.74%) of the records are incorrectly classified instances.

Experiment 8:

This experiment is performed, by changing the default testing option of 10-fold cross validation to the percentage split (66%) and the outcome of this experiment is presented in table 4.9.

Algorithm	Test Option	Accuracy	Time	Recall	precision	F-Measure	Class
JRIP	Percentage split (66%)	95.12%	0.00 sec	89.9%	95.4%	92.5%	EEU
				95.5%	90.1%	92.7%	ETHIO
				100%	100%	100%	FB&TR
Weighted Average		***95.12%***	***0.00 sec***	***95.1%***	***95.2%***	***95.1%***	

Table 4.9 Performance result of JRIP algorithm with default percentage split (66%)

As shown from the performance result, JRIP Rule induction algorithms with default percentage split values (66%) scored an accuracy of 95.12%. This result shows that out of the total training datasets 5496 (95.12%) records are correctly classified instances, while 282 (4.88%) of the records are incorrectly classified instances.

JRIP Rule induction algorithms with 10-fold cross validation scored an accuracy of 95.26% from that out of the total training datasets 16189 (95.26%) records are correctly classified instances, while 806 (4.74%) of the records are incorrectly classified instances. The second experiment in this algorithm is applied with default values of percentage split (66%) and scored an accuracy of 95.12% from the total training datasets 5496 (95.12%) records are correctly classified instances, while 282 (4.88%) of the records are incorrectly classified instances.

Generally, when we compare the two experiments conducted before, the JRIP Rule induction algorithm with 10-fold cross validation and the JRIP with default percentage split (66%), the model developed with the 10-fold cross validation algorithm gives a better classification performance of identifying newly arriving mobile call drops reason than default percentage split (66%). So, the model constructed using 10-fold cross validation has been chosen due to its better classification performance with an accuracy of 95.26%.

4.7. Comparison of J48, RF, PART and JRIPP Models

Selecting a better classification technique for building a model, which performs best in handling the prediction of call drop reasons are one of the aims of this study. For this reason, four best performing classification models are compared. Table 4.10 present algorithms and test options that register the best performance.

Type of algorithm	Accuracy	Time taken	Recall	precision	F-Measure
J48	95.43%	0.06 sec	95.4%	95.4%	95.4%
Random Forest	95.36%	1.31 sec	95.4%	95.4%	95.4%
PART	95.41%	0.11 sec	95.4%	95.5%	95.4%
JRIP	95.26%	2.27 sec	95.3%	95.3%	95.3%

Table 4.10 Performance Comparison of the selected models

As shown in the table 4.10, all algorithms are performing well, with an accuracy of all more than 95%. Especially, J48 and PART algorithms register the highest accuracy of 95.43% and 95.41% respectively. But when we compare the time take to build the model between those two algorithms, J48 registered better model building time than PART which means J48 algorithm has high speed to build the model than PART.

Hence J48 decision tree algorithm with 10-fold cross validation is used as a model of this study because it has high performance and fast model building time than all the other algorithms study in this research. The snapshot running information of J48 decision tree algorithm with 10-fold cross validation is shown in annex 2. The confusion matrix output of J48 decision tree algorithm with 10-fold cross validation is also shown in the table 4.11 below.

Confusion Matrix			
a	b	C	Classified as
5125	540	0	a= EEU POWER
237	5428	0	b= ETHIO POWER
0	0	5665	c= FIBER & TRANSMISSION

Table 4.11 Confusion Matrix for J48 algorithm with 10-fold cross validation

As shown in table 4.11, there is no misclassification in FIBER & TRANSMISSION class. But the misclassification happens between EEU POWER and ETHIO POWER classes. This is because as shown in chapter three section 3.1.1 demographic characteristics of the attributes, these two classes have high number of instances than the third class and due to this variation, the prediction model was highly influenced by these two classes. Then in order to avoid this variation between the data, we use resampling techniques.

As discussed with domain experts, the other reason for misclassification between these two classes was there is a relationship between these classes in that if EEU POWER occur, there is also a possibility that ETHIO POWER to be occurred.

4.8. Rules generated by selected algorithms

As discussed before from those experiments conducted in supervised approach, the J48 decision tree algorithm with 10-fold cross validation gives a better classification performance of identifying the newly arriving call drop reasons and high model building time than the other algorithms used in this research. The rules indicate that the possible conditions in which the FMS record could be classified in each of the classes. From this model a set of rules are extracted by traversing the decision tree and generating a rule for each leaf and making a combination of all the tests found on the path from the root to the leaf node.

One of the interesting rules detected is how much resource is critical in order to predict the call drop reasons. As we discussed in chapter three section 3.3.1, resource is derived attribute and in real situation the domain experts don't consider it during call drop statistical method analysis.

Another interesting rule detected was the generator status, even though there is a stand by generator in the sites one of the major reasons for call drop is power.

The following are some of the interesting rules extracted from the decision tree. Therefore, those which cover more cases and have better accuracy are chosen. The following rules indicate the possible conditions in which a call drop reasons could be classified in each of EEU POWER, ETHIO POWER and FIBER&TRANSMISSION classes.

Rule 1: If number of customers = LOW and DGStatus = DG and season = BEGA and Device Location = EAAZ, then the reason for mobile drop is EEU POWER.

If number of customers is LOW and if there is standby generator for the device and if the season is BEGA and if the device is found in East Addis Ababa zone, then the reason for mobile drop is EEU POWER.

Rule 2: If number of customers = LOW and DGStatus = DG and season = KIREMT and Resource = EXCELLENT, then the reason for mobile drop is FIBER&TRANSMISSION.

If number of customers is LOW and if there is standby generator for the device and if the season is KIREMT and if there is an EXCELLENT amount of resources, then the reason for mobile drop is FIBER&TRANSMISSION.

Rule 3: If number of customers = LOW and DGStatus = NO DG and Technology = 2G, then the reason for mobile drop is ETHIO POWER.

If number of customers is LOW and if there is NO standby generator for the device and if the Technology is 2G, then the reason for mobile drop is EEU POWER.

Rule 4: If number of customers = MEDIUM and DGStatus = NO DG and Device Location = WAAZ, then the reason for mobile drop is FIBER&TRANSMISSION.

If number of customers is MEDIUM and if there is NO standby generator for the device and if the device is found in West Addis Ababa zone, then the reason for mobile drop is FIBER & TRANSMISSION.

Rule 5: If number of customers = MEDIUM and DGStatus = DG and Resource = POOR, then the reason for mobile drop is EEU POWER.

If number of customers is MEDIUM and if there is standby generator for the device and if the amount of resources is POOR, then the reason for mobile drop is EEU POWER.

Rule 6: If number of customers = MEDIUM and DGStatus = NO DG and Device Location = EAAZ, then the reason for mobile drop is EEU POWER.

If number of customers is MEDIUM and if there is NO standby generator for the device and if the device is found in East Addis Ababa zone, then the reason for mobile drop is EEU POWER.

Rule 7: If number of customers = HIGH and Resource = POOR, then the reason for mobile drop is ETHIO POWER.

If number of customers is HIGH and if there is POOR amount of resources, then the reason for mobile drop is ETHIO POWER.

Rule 8: If number of customers = HIGH and Resource = GOOD and DGStatus= DG, then the reason for mobile drop is EEU POWER.

If number of customers is HIGH and if there is GOOD amount of resources and if there is standby generator, then the reason for mobile drop is EEU POWER.

Rule 9: If number of customers = HIGH and Resource = EXCELLENT, then the reason for mobile drop is FIBER&TRANSMISSION.

If number of customers is HIGH and if there is EXCELLENT amount of resources, then the reason for mobile drop is FIBER&TRANSMISSION.

Rule 10: If number of customers = LOW and DGStatus = DG and Season = BEGA and Device Location = WAAZ, then the reason for mobile drop is ETHIO POWER.

If number of customers is LOW and if there is standby generator and if the season is BEGA and if the device is found in West Addis Ababa zone, then the reason for mobile drop is ETHIO POWER.

Rule 11: If number of customers = MEDIUM and DgStatus = NO DG and Device Location = EAAZ, then the reason for mobile drop is EEU POWER.

If number of customers is MEDIUM and if there is NO standby generator and if the device is found in East Addis Ababa zone, then the reason for mobile drop is EEU POWER.

Rule 12: If number of customers = HIGH and Resource = GOOD and DGStatus= NO DG, then the reason for mobile drop is ETHIO POWER.

If number of customers is HIGH and if there is GOOD amount of resources and if there is NO standby generator, then the reason for mobile drop is ETHIO POWER.

The overall decision tree of J48 algorithm with 10-fold cross validation generated tree is shown in annex 6.

4.9. Discussion of the results with domain experts

The experts explained that more attention should be given to locations where there is high number of customers and kiremt seasons because when the number of users in a specific area increases the resources like switching antennas becomes crowded and more congested; due to this the calls become dropped. The season status has also direct impacts on call drops in that, in Ethiopian condition there are seasons when there is high amount of natural effects like thunderstorm and heavy rains. So, because of these natural effects the mobile transmitting resource become down and unable to transmit calls.

From the generated rules it is observed that most determinant factors are number of customers, followed by DG Status and Resources. The domain experts argued that, the discovered rules are acceptable, and amount of resources has a direct relationship and great impact on call drop reasons.

4.10. Use of Knowledge

After evaluating the discovered knowledge, the last step is using this knowledge for the industrial purposes. In this step the knowledge discovered is incorporated in to performance system and take this action based on the discovered knowledge.

In this research the discovered knowledge is used by integrating the user interface which is designed by java programming with a Weka system in order to show the prediction of call drop reasons.

In order to predict the future mobile call drop reasons, we analyze the current call dropped reasons based on the available data by generating rules with selected algorithm. Then we use the generated rules by implementing it with a java programming language to predict the future call drop reasons.

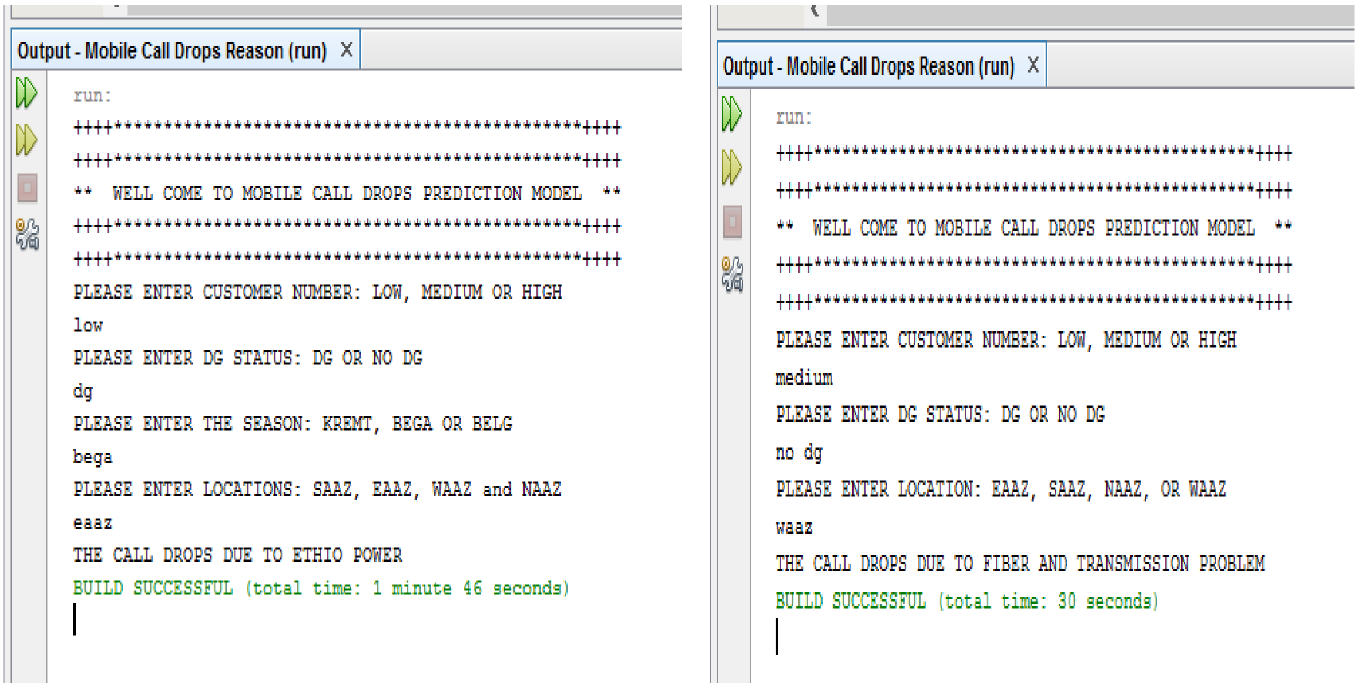


Figure 4.3 Mobile call drops prediction model sample prediction outputs

4.11. User Acceptance Testing

According to Luo [29], User Acceptance Testing is used to conduct operational readiness of a product, service or system as part of a quality management system. It is a common type of non-functional software testing, used mainly in software development and software maintenance projects. This type of testing focuses on the operational readiness of the system to be supported. It is done when the completed system is handed over from the developers to the customers or users. The purpose of user acceptance testing is rather to give confidence that the system is working than to find errors.

User Acceptance Testing verify the system's behavior is consistent with the requirements. These tests will reveal defects within the system. The work associated with it begins after requirements are written and continues through the final stage of testing before the user accepts the new system [29].

The goal of User Acceptance Testing is to assess if the system can support day-to-day business and user processes and ensure the system is sufficient and correct for business usage. The primary objective of User Acceptance Testing is to demonstrate that you can run your business using the system if it is fit for the purpose [29].

In this research we perform User Acceptance Testing by presenting and discussing with the organization's domain experts. The following areas are discussed in detailed with domain experts and discussions are presented below.

4.11.1. Efficiency

Efficiency is the ability to avoid wasting materials, energy, efforts, money, and time in doing something or in producing a desired result. In a more general sense, it is the ability to do things well, successfully, and without waste. In scientific terms, it is a measure of the extent to which input is well used for an intended task or function (output). It often specifically comprises the capability of a specific application of effort to produce a specific outcome with a minimum amount or quantity of waste, expense, or unnecessary effort.

In this research efficiency is considered as a time taken to predict the call drop reasons by taking inputs from the user. As discussed in chapter one, currently in ethio telecom mobile network data analysis is done by traditional simple statistical methods which need more time to operate because every operation is done manually.

During our presentation and discussion with domain experts, we conduct sample experiments and compare the efficiency between the current statistical method and our new prediction method. From these sample experiments, our new call drops reason prediction method become more efficient and every domain expert agreed up on this. They argued that due to this efficiency improvement they can minimize more than half a time taken before.

4.11.2. Effectiveness

Effectiveness is the capability of producing a desired result or the ability to produce desired output. When something is deemed effective, it means that it has an intended or expected outcome or produces a deep impression.

In this research effectiveness is considered as the accuracy to predict right call drops reason. As discussed in chapter four, we perform experiments with J48 and Random Forest from decision tree algorithm and PART and JRIP from rule induction algorithm. As a result, J48 algorithm registers better performance of 95.43% accuracy.

During our presentation and discussion with domain experts, our experimental results are discussed, and they give a recommendation to improve this performance.

4.11.3. Error Tolerance

Error Tolerance is concerned about management of faults originating from defects in design or implementation. In this research error tolerance is considered as making the experiments error free or making it intelligence. As discussed before this experiment registers a better performance of 95.43% accuracy with J48 decision tree algorithm. During our presentation and discussion with domain experts, we discuss on how to make this experiments intelligence by integrating this experiment with knowledge-based systems and we agreed that some improvements also need.

4.11.4. Easy to Use (easy to learn & easy to remember)

Making a product easy to use is one of the non-functional requirements for any product. In order to make this research outputs easy to use, we prepare sample screen shots on the document.

In this study, we perform user acceptance testing to evaluate systems efficiency, effectiveness, easy to learn and easy to remember point of view.

In this study, a total of 8 domain experts from ethio telecom specifically from mobile performance and mobile quality analysis sections are participated to evaluate the systems acceptance. Each of the study participants are asked to give feedback on the acceptability of the prediction and to rate it on a scale of 1 (Strongly Disagree) to 5 (Strongly Agree). Summary of the result is presented in table 4.12 below.

Questionnaires	Strongly Agree (5)	Agree (4)	Undecided (3)	Disagree (2)	Strongly Disagree (1)
Efficiency:					
➢ The prediction response is fast.	90%	10%	-	-	-
➢ The prediction saves energy & materials.	95%	5%	-	-	-
Effectiveness:					
➢ The prediction is Reliable.	80%	10%	10%	-	-
➢ The prediction produces a desired result.	70%	20%	5%	5%	-
➢ The prediction system tolerates errors.	60%	20%	10%	5%	5%
Easy to Learn:					
➢ The prediction system is Easy to learn.	80%	10%	10%	-	-
➢ The prediction system is User friendly.	70%	20%	10%	-	-
Easy to Remember:					
➢ The prediction system is easy to remember.	75%	20%	5%	-	-
➢ The prediction model is explicit.	65%	25%	10%	-	-

Table 4.12 Experts response summary on the proposed call drops reason prediction model

This study revealed that from 8 domain experts 7 of them confirm that this call drops reason prediction model was much efficient, and it saves their energy and materials while comparing with the way they perform currently which is the simple statistical method. But one domain experts do not satisfy with the prediction system specially in error tolerance criteria.

In the case of effectiveness, the domain experts revealed that this prediction model produces a desired result. In order to make the prediction near perfect, there is a need to enhance the performance of the model near 100%. Some (5%) of the domain experts were disagree with the effectiveness result because they stated that in case of call drops reason analysis the reason must have to be perfect because it has a significant impact on the organizations revenue.

According to this study, the performance result of the prediction model scored an accuracy of 95.43% which is good. Most of the domain experts satisfied with the prediction results but some of them are strongly disagree with the prediction because in some cases the model doesn't tolerate errors. In order to make the prediction more accurate and error tolerable, we advise integration of the discovered classification rules with knowledge-based system.

Concerning the extent to which the prediction model was easy to learn and easy to remember, the respondents reply shows that the prediction model was user friendly and it is more explicit than before. However, there are up to 10% respondents who are undecided because they are agree to some extent on the systems user friendless and to some extent they are not agree.

In the overall user acceptance criteria, 90% of the domain experts agreed that this call drop reasons prediction system is much efficient, it saves energy and materials, the prediction is reliable, the prediction produces a desirable result, the prediction system is easy to use and remember, it is user friendly and the prediction model is more explicit than before. The domain experts also suggested that there need to enhance the performance of the model to make the prediction near to 100%.

CHAPTER FIVE: CONCLUSION AND RECOMMENDATIONS

5.1. Conclusion

The traditional method of turning data into knowledge relies on manual analysis and interpretation. In telecommunications sector, different analysis was made manually, and different reports are generated. The report becomes the basis for future decision making and planning for telecommunications network expansion, performance and quality of service (QoS) evaluation. Data analysis using traditional simple statistical tools is slow, expensive, and highly subjective. Huge amount of data is generated from mobile network elements and stored in telecommunication databases for different purposes. Hence, manual data analysis and interpretation is impractical.

The objective of this research was to develop a predictive model that can determine mobile call drops from ethio telecom mobile network data using data mining techniques. To achieve this objective, we use Fault Management System (FMS) data because the FMS data includes enough information about call drop reasons. Due to the fact that the FMS data is very huge and requires more space, the ethio telecom servers stored no more than 2 years data. Therefore, we took 1 year and six-month data of the telecommunication company.

This research proposes data mining to overcome the problem of manual data analysis. Hybrid process model was used while undertaking the experimentation. The study was conducted using WEKA software version 3.8 and four data mining algorithms of classification techniques was used, namely J48, PART, Random Forest and JRIP.

We use the purposive sampling techniques to extract the data. In ordered to extract the huge damp file from the ethio-telecom database, we use oracle database software. In ordered to manage the data in application software (in MS-Access and MS-Excel), file splinter software is used. After eliminating irrelevant and unnecessary data, a total of 16996 datasets are selected from FMS and used for the purpose of conducting this study. Two derived attributes and out of 26 attributes six relevant attributes from FMS network database server are selected to conduct this research. It has been preprocessed and prepared in ARFF format which is suitable for the DM tasks.

The J48 decision tree algorithm registered better performance of 95.43% accuracy and processing speed of 0.06 sec running with 10-fold cross validation using 8 attributes than any experimentation done for this research.

One of the basic targets of data mining is to compare different models and to select the best classification accuracy accordingly. Therefore, detailed experimentation for different models has been conducted. Among the four models, the J48 decision tree algorithm with 10-fold cross validation registers better performance and processing time than other experimentations done in this research. It registers an accuracy of 95.43% and processing time of 0.06 sec.

The finding of this study shows that the causes for mobile call drops are different. The mining result identified that number of customers is the major factor, followed by generator status and availability of resources. Seasonal status, device location and technology used also have a great contribution. The analysis which was closely undertaken with domain experts are achieved a good result.

The result of this research shows that applying data mining to analyze mobile network data helps ethio telecom to improve QoS and make decision based on the information discovered from the analysis. Providing QoS to its customers will lead the organization to satisfy its customers and revenue augmentation for ethio telecom.

We face different challenges in conducting this study. Unavailability of related works on telecommunication mobile call drop area was one of the major challenges during the study. To extract and manage the huge call drops record data from the telecommunication server and to integrate process of the data have been also another challenged times for us.

One of the challenges that the algorithm encountered is inability to classify the EEU POWER and ETHIO POWER classes. This misclassification between these two classes is happened due to the reason that there are unbalanced data instances between these classes and the third class and due to this unbalanced data instances, the prediction model was highly influenced by these two classes. Then in order to avoid this variation between the data, we use resampling techniques.

As discussed with domain experts, the other reason for misclassification between these two classes was there is a relationship between these classes in that if EEU POWER occur, there is also a possibility that ETHIO POWER to be occurred.

5.2. Recommendations

This research is mainly conducted for an academic purpose. This research has proven the applicability of different DM classification techniques namely, J48, Random Forest, JRIP and PART algorithms which automatically discover hidden knowledge that are interesting and accepted by domain experts. Based on the investigations of the study, the following areas are given as a recommendation for the future.

- The techniques employed in this study were decision tree and rule induction algorithms. Even though an encouraging result was obtained, using other types of techniques with a different parameter might perform better in solving the misclassification observed between ETHIO POWER and EEU POWER. Therefore, it is recommended other researchers to test with other types of techniques like artificial neural network and support vector machine.
- In this research, we use only Addis Ababa mobile network data however further investigation is needed by including other regional mobile network data so as to comprehensively see the cause for call drop reasons.
- In this research, we use one of data mining technique, classification. The association rule-based data mining techniques can be used in future:
    - To study the relationship or extract interesting correlations between attributes
    - To study the cause and effect or associations among sets of items
- In order to make the research error tolerance, Integration of the discovered classification rules with knowledge-based system is need for the future.
- This study has attempted to apply DM techniques on Fault Management System data, but it could also be applied in other telecommunication data like network fault isolation, customer profiling and event log analysis for decision making and other purposes.

REFERENCES

[1] E. Frank, Data Mining Concepts and Techniques, Elsevier Inc: Morgan-Kaufmann, 2012.

[2] S. Deshpande, "Data Mining System and Applications: A Review," 2010.

[3] U. Fayyad, "Knowledge Discovery and Data Mining: Towards a Unifying Framework", Gregory , 1996.

[4] W. Bogale, "A Background Paper on Telecom & Telecom Statistics in Ethiopia," ethio telecom, Ethiopia, February 2005.

[5] J. Vitor, "Telecommunication Fraud Detection Using Data Mining techniques," June 30, 2014.

[6] J. Jackson, "Data Mining: A Conceptual Overview," 2002.

[7] N. Pradeep, "Analysis of Telecommunication Data: Call Drop," 2010.

[8] E. Taye, Telecommunication in Ethiopia, Geneva : UNCTAD, 2010.

[9] R. Vaarandi, "Tools and Techniques for Event Log Analysis", 2012.

[10] Methodology, "Methodology," 27 June 2018. [Online]. Available: https://en.methodology.org/Methodology. [Accessed 2018].

[11] K. Cios, Data Mining: A Knowledge Discovery Approach, New York, USA: Springer, 2008.

[12] K. Cios, "The Knowledge Discovery Process In Data Mining," Springer US, 2007, pp. 9-24.

[13] G. Williams, Data Mining with Rattle and The art of excavating data for knowledge discovery, Springer, 2011.

[14] I. Witten, Data Mining: Practical machine learning tools and techniques, Elsevier: Morgan Kaufmann , 2005.

[15] L. Breiman, "Random Forests Machine learning," 2001, pp. 5-32.

[16] C. Hilas, "Data Mining Approaches to Fraud Detection in Telecommunications," *2nd PanHellenic Conference on Electronics and Telecommunications - PACET'12,* March 16, 2012.

[17] G. Jember, "Data mining application in supporting fraud detection on mobile communication: the case of Ethio mobile.," 2005.

[18] A. Mulugeta, "Application of Data Mining Techniques to Customer Classification and Clustering: The Case of Ethiopian telecommunications Corporation.," 2012.

[19] N. Ogwueleka, "Fraud Detection in Mobile Communications Networks Using User Profiling and Classification Techniques.," 2009.

[20] B. Kusaksizoglu, "Fraud Detection in Mobile Communication Networks Using Data Mining.," 2006.

[21] G. Gebremeskel, "Data mining application in supporting fraud detection: on Ethio-mobile services.," 2006.

[22] D. Olson, "Advanced data mining techniques", Springer, 2008.

[23] M. Robertson, "An Analysis of Data Mining: Past, Present and Future," 2012.

[24] M. Venkatadri, "Data Mining for Managing Intrinsic Quality of Service in Digital Mobile Telecommunications.," 2004.

[25] P. Vehviläinen, "Data Mining for Digital Mobile Telecommunication Network's Quality of Service Performance Measurements.," 2012.

[26] N. Hung, "Data Cleaning and Data Preprocessing Techniques," *Academic Press,* 2008.

[27] W. Staling, "Wireless Communications and Networking," 2010.

[28] R. Haryana, "Decision Tree: Data Mining Techniques Department of Computer Science," India, 2010.

[29] L. Luo, "Software Testing Techniques Technology Maturation and Research Strategy," 2012.

[30] J. David, Introduction to Programming Using Java, Hobart and William Smith Colleges, December 2006.

[31] W. Daimler, "CRISP-DM: Towards a Standard Process Model for Data Mining", 2001.

[32] G. Weiss, "Data Mining in Telecommunications Industry," 2009.

[33] R. Osmar, Principles of Knowledge Discovery in Databases CMPUT690, 1999.

[34] J. Seifert, "Data mining: An overview of National security issues," 2004, pp. 201-217.

[35] U. Fayyad, "From data mining to knowledge discovery in databases", 1996.

[36] M. Dunham, Data mining introductory and advanced topics, Upper Saddle River, NJ: Pearson Education, Inc. : , 2003.

[37] D. Crano, "Principles and Methods of Social Research", 2002.

[38] J. Roiger, "Data Mining: A Tutorial-Based Primer", Boston: Addison- Wesley, 2003.

[39] R. Marco, Data Mining Applied to Validation of Agent Based Models, Proceedings of Ninteenth European Conference on Modelling and Simulation, 2005.

[40] L. Rokach, "Top-down induction of decision trees classifiers-a survey," IEEE Transactions on Systems, Man, and Cybernetics, 2005, p. 476–487.

[41] M. Hemalatha, "A perspective analysis of traffic accident using data mining techniques," *International Journal of Computer Applications,* 2011.

[42] G. Dereje, "Discovery of Hidden Knowledge from ethio telecom mobile network data," April 2015.

[43] A. Azevedo, "KDD, SEMMA and CRISP-DM: a parallel overview," IADIS European conference on data mining, 2008, pp. 128-185, ISBN: 978-972-8924-63-8.

[44] C. Nelson, "Neural Network to identify and prevent bad debt in Telephone Companies.," *Niels Brock Business College,* 2005.

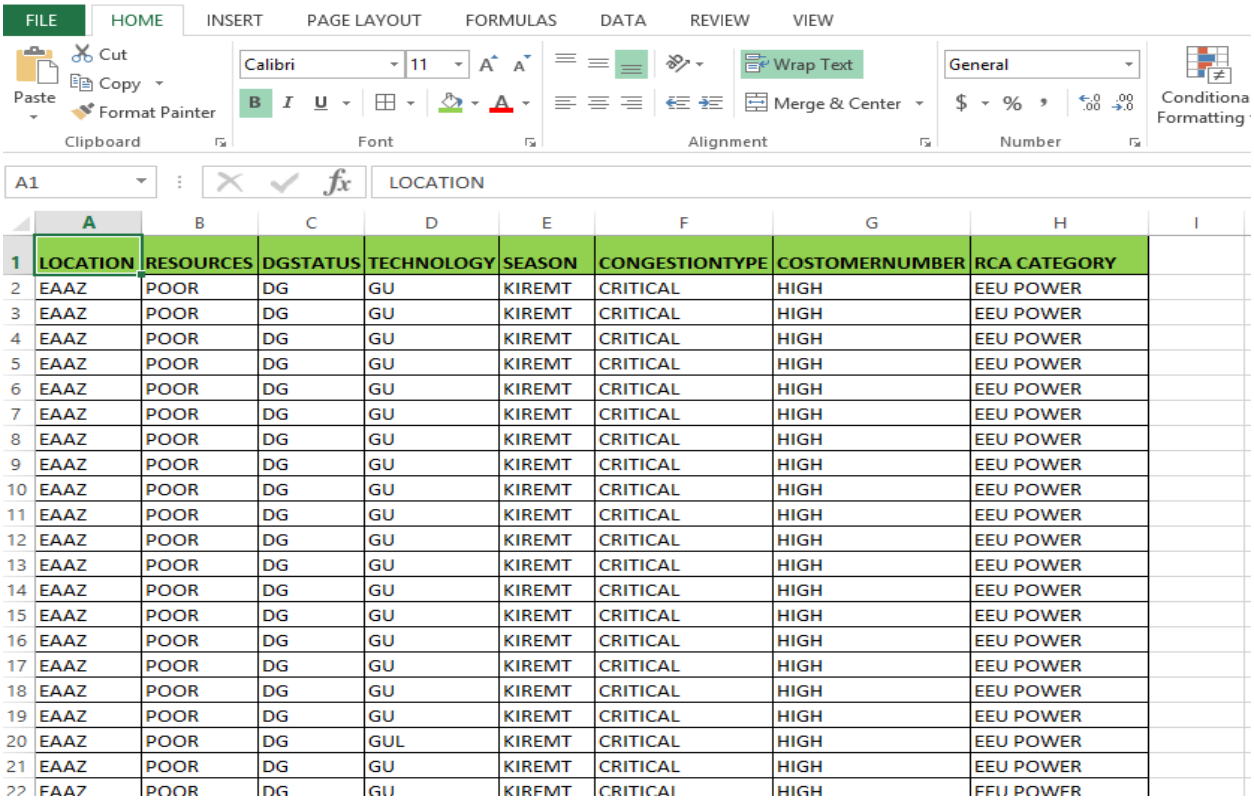
[45] O. Maimon, "Introduction to Data Mining and Knowledge Discovery", Two Crows Corporation, 1999.

[46] A. Rajput, "J48, PART and JRIP Rules for E-Governance Data," 2011.

[47] K. Rangra, "Comparative Study of Data Mining Tools," *International Journal of Advanced Research in Computer Science and Software Engineering ,* vol. Volume 4, no. Issue 6, June 2014.

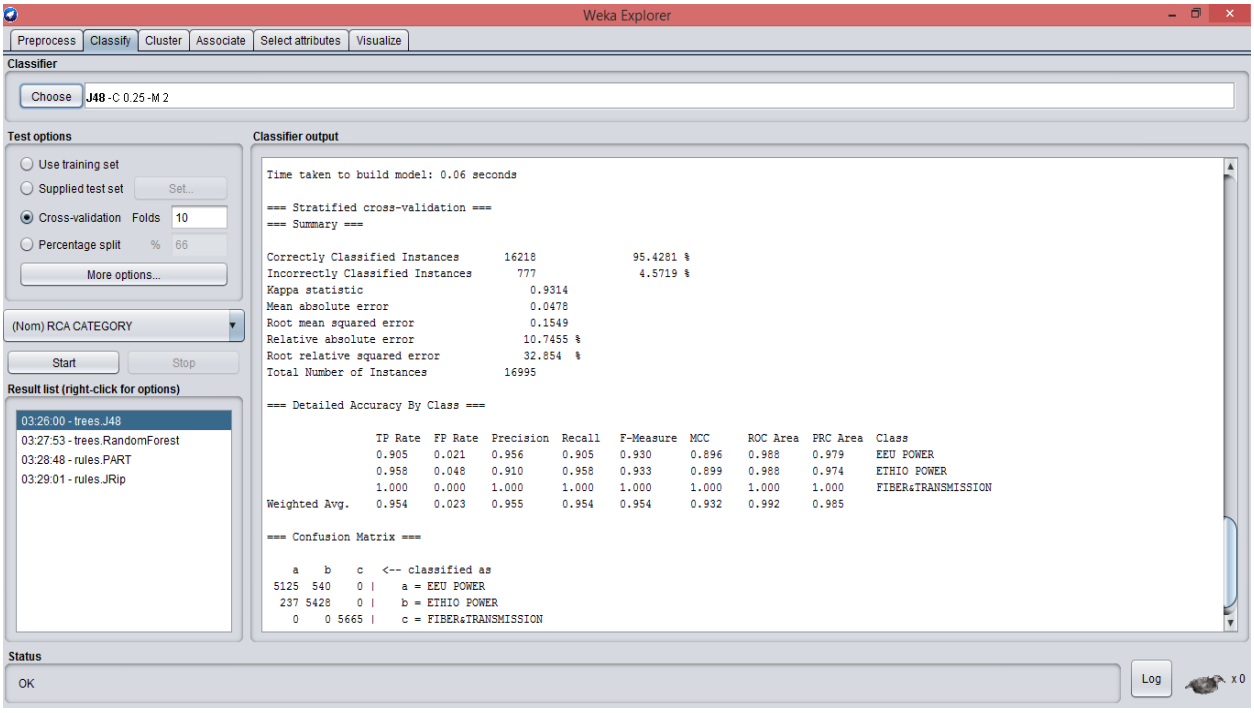
ANNEXES

Annex-1: The original collected sample data

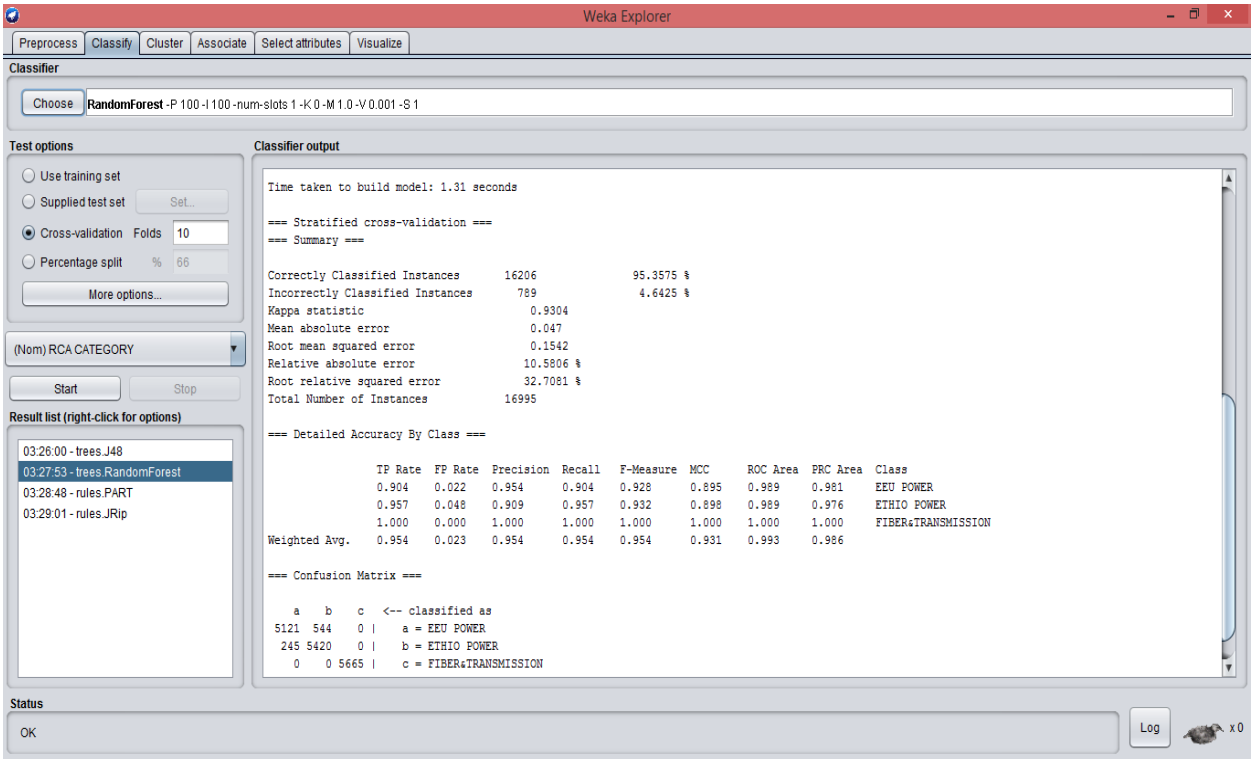


	A	B	C	D	E	F	G	H	I
1	LOCATION	RESOURCES	DGSTATUS	TECHNOLOGY	SEASON	CONGESTIONTYPE	COSTOMERNUMBER	RCA CATEGORY	
2	EAAZ	POOR	DG	GU	KIREMT	CRITICAL	HIGH	EEU POWER	
3	EAAZ	POOR	DG	GU	KIREMT	CRITICAL	HIGH	EEU POWER	
4	EAAZ	POOR	DG	GU	KIREMT	CRITICAL	HIGH	EEU POWER	
5	EAAZ	POOR	DG	GU	KIREMT	CRITICAL	HIGH	EEU POWER	
6	EAAZ	POOR	DG	GU	KIREMT	CRITICAL	HIGH	EEU POWER	
7	EAAZ	POOR	DG	GU	KIREMT	CRITICAL	HIGH	EEU POWER	
8	EAAZ	POOR	DG	GU	KIREMT	CRITICAL	HIGH	EEU POWER	
9	EAAZ	POOR	DG	GU	KIREMT	CRITICAL	HIGH	EEU POWER	
10	EAAZ	POOR	DG	GU	KIREMT	CRITICAL	HIGH	EEU POWER	
11	EAAZ	POOR	DG	GU	KIREMT	CRITICAL	HIGH	EEU POWER	
12	EAAZ	POOR	DG	GU	KIREMT	CRITICAL	HIGH	EEU POWER	
13	EAAZ	POOR	DG	GU	KIREMT	CRITICAL	HIGH	EEU POWER	
14	EAAZ	POOR	DG	GU	KIREMT	CRITICAL	HIGH	EEU POWER	
15	EAAZ	POOR	DG	GU	KIREMT	CRITICAL	HIGH	EEU POWER	
16	EAAZ	POOR	DG	GU	KIREMT	CRITICAL	HIGH	EEU POWER	
17	EAAZ	POOR	DG	GU	KIREMT	CRITICAL	HIGH	EEU POWER	
18	EAAZ	POOR	DG	GU	KIREMT	CRITICAL	HIGH	EEU POWER	
19	EAAZ	POOR	DG	GU	KIREMT	CRITICAL	HIGH	EEU POWER	
20	EAAZ	POOR	DG	GUL	KIREMT	CRITICAL	HIGH	EEU POWER	
21	EAAZ	POOR	DG	GU	KIREMT	CRITICAL	HIGH	EEU POWER	
22	EAAZ	POOR	DG	GU	KIREMT	CRITICAL	HIGH	EEU POWER	

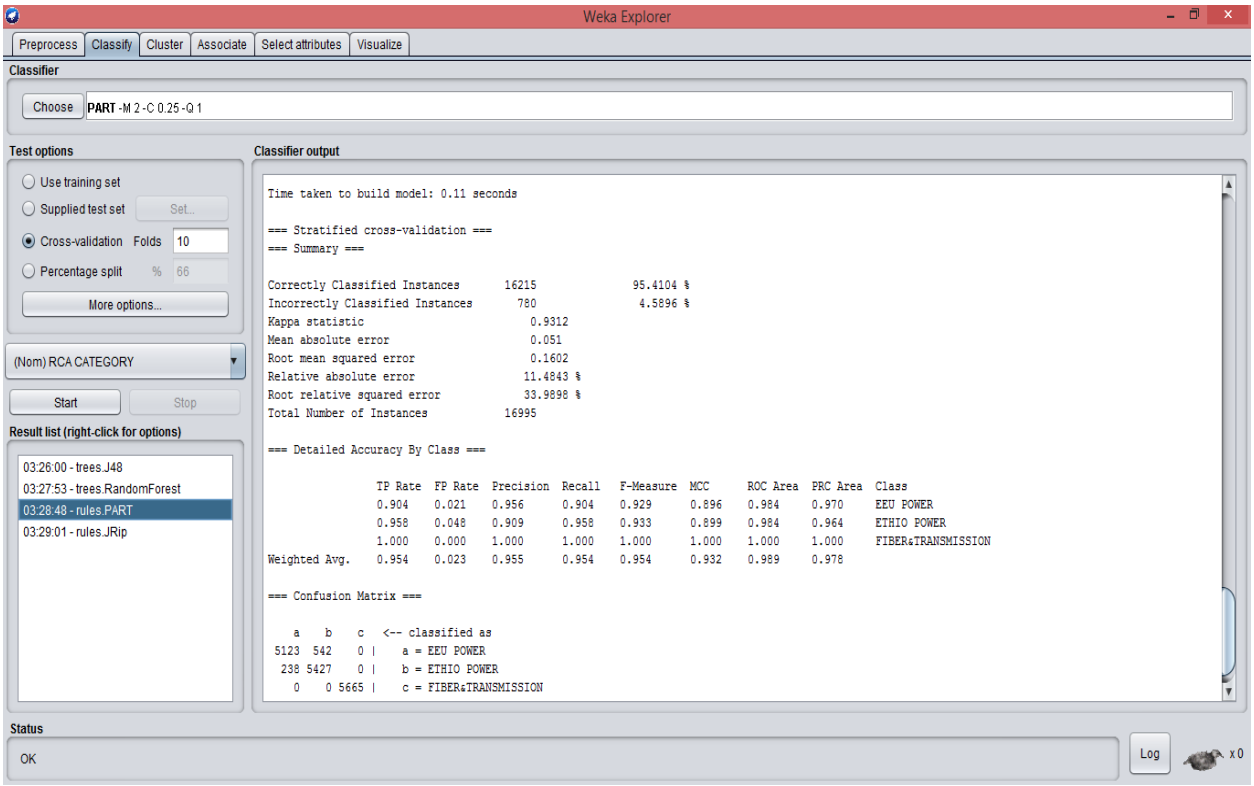
Annex-2: The snapshot running information of J48 with 10-fold validation technique



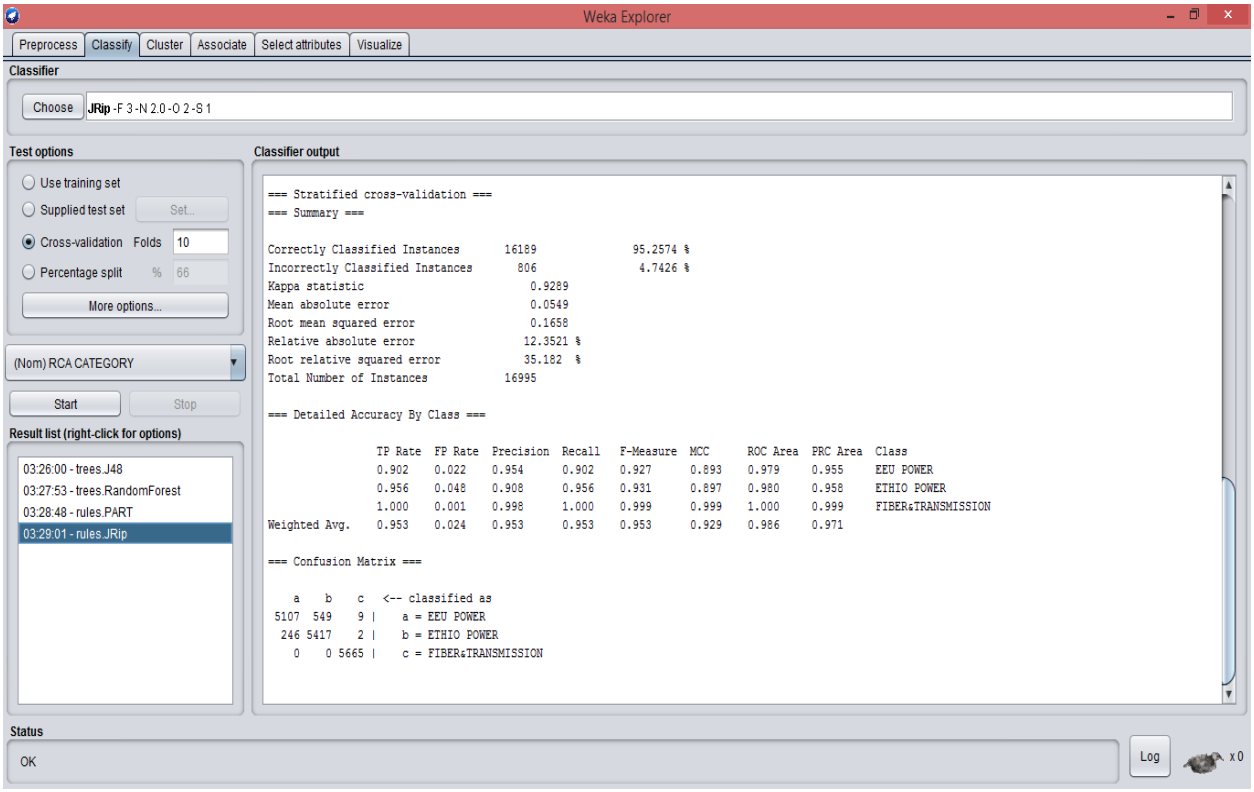
Annex-3: The snapshot of Random Forest with 10-fold validation technique



Annex-4: The snapshot running information of PART with 10-fold validation technique



Annex-5: The snapshot of JRIP with 10-fold validation technique



Annex-6: J48 call drop reasons prediction decision tree