A Termination Prediction for Postpaid Mobile Service using Machine learning: the case of ethio telecom

BY: TEWABE KASSAW

ADVISOR: EPHREM TESHALE (PhD)

A Thesis submitted to

School of Electrical and Computer Engineering

Addis Ababa Institute of Technology

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

(Telecommunication Engineering)

Addis Ababa University

Addis Ababa, Ethiopia

February 4, 2022

Declaration

I, the undersigned, declare that the thesis comprises my own work in compliance with internationally accepted practices; I have fully acknowledged and referred all materials used in this thesis work.

Tewabe Kassaw

_____________________ _____________________

Name Signature

Addis Ababa University

Addis Ababa Institute of Technology

School of Electrical and Computer Engineering

This is to certify that the thesis prepared by **Tewabe Kassaw**, entitled *A Termination Prediction for Postpaid Mobile Service using Machine learning: the case of ethio telecom* and submitted in partial fulfillment of the requirements for the degree of Master of Science (Telecommunication Engineering) complies with the regulations of the University and meets the accepted standards with respect to originality and quality.

Signed by the Examining Committee:

Examiner 1	Surafel Lemma (PhD)	Signature	________	Date	______
Examiner 2	Sosina Mengistu (PhD)	Signature	________	Date	______
Advisor	Ephrem Teshale (PhD)	Signature	________	Date	______

____

Dean, School of Electrical and Computer Engineering

ABSTRACT

Customer churn is a major issue for operators that may greatly affect their revenue. Even if in Ethiopia's monopolistic market, there is still a challenge in the form of subscriber service termination. To solve this issue, operators must first identify potential service termination in advance and then take proactive measures to reduce the number of terminations. For this purpose, telecom operators need a prediction model to predict correctly and timely potential service termination subscribers based on collected service usage-related data. To anticipate subscriber behavior, the performance of several prediction methods has been explored. However, the performance of such algorithms is not examined in the context of Ethiopia postpaid mobile service and in the case of multi-class, where ethio telecom possesses multi-class data that enable develop a multi-class prediction model for mobile subscriber service termination.

The goal of this study is to investigate the performance of prediction learning algorithms with multi-class scenarios for predicting service termination in the context of Ethiopian postpaid mobile subscribers. The algorithms investigated include J48 Decision tree, Random Forest (RF), and Multilayer Perceptron (MLP), and their performance is measured in terms of prediction accuracy, precision, recall, and F-Measure. Cross-validation (k=10) techniques and a multi-class dataset are used to test the performance of algorithms. WEKA 3.9.4 tool algorithm implementation was utilized for performance evaluation. As a result, the J48 and RF prediction algorithms almost have the same performance on all performance parameters result. However, MLP algorithm achieved a low-performance score compared to J48 and RF, and the accuracy of J48, RF, and MLP are 94.9%, 95.1%, and 93.3% respectively.

Keywords: Postpaid mobile service, Termination prediction, Machine learning, Multi-class

ACKNOWLEDGMENTS

I would like to express my heartfelt gratitude to everyone who helped me complete my thesis work.

First of all, I would like to express my thank the Almighty God, the source of all good things, for granting me the capacity to complete my thesis.

I am profoundly grateful to thank my advisor, Dr. Ephrem Teshale, for his time, continued support, and excellent comments. He always listened and encouraged me.

It is also my pleasure to heartfully thank Dr. Surafel Lemma and Dr. Sosina Mengistu for their genuine, constructive feedback during the seminar presentation phases from which I got very important ideas and useful perspectives.

I am grateful to Werksew Motbayhon, my wife, for her unreserved support throughout the course work and during this thesis. Thanks also to all my friends who deserve my appreciation for their frequent encouragement and invaluable feedback.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACRONYMS

ANN	Artificial neral network
ARFF	Attribute relation file format
B2C	Bussiness-to-Customer
C2C	Consumer-to-Consumer
CBS	Convergent Billing System
CDR	Call Detail Record
CRM	Customer Relationship Management
CSV	Comma Separated Values
EM	Expectation maximization
FN	Falls negative
FP	Falls positive
GPS	Global Positioning System
IQR	Interquartile range
ML	Machine learning
MLP	Multilayer perceptron
M-service	Mobile service
PDAs	Personal Digital Assistants
RF	Random forest
SCUT	SMOTE+ Cluster-based undersampling
SMOTE	Synthetic Minority Oversampling Technique
SMS	Short Message Service
TN	True negative
TP	True positive
WEKA	Waikato Environment for Knowledge Analysis

1 INTRODUCTION

Mobile service is a service that can be accessible at any time and place. It is one of the primary services provided by telecom service providers, to which consumers may subscribe in two ways: prepaid subscription and postpaid subscription. A postpaid subscription is a service contract in mobile communications in which a client gets billed at the end of each month for the mobile services they have used during that month. Prepaid, on the other hand, refers to a subscription in which customers pay in advance for mobile services by purchasing the service allocation as credit [1, 2]. Postpaid mobile plans are appropriate for customers who are frequent mobile users looking for a more stable contract without having to constantly top-up (recharge). For many mobile service providers, postpaid subscribers are generally the preferred option because they can expect a consistent income from these subscribers [2, 3].

Service termination is a serious difficulty for operators that may considerably impact their income, even in a monopoly market like ethio telecom in Ethiopia. Since according to author [4], the definition of service termination is simply quite a service from a service provider without moving to another competitor even if subscribers do not have other alternative service providers. Service termination is prevalent in many sectors including telecommunications and it can be voluntary or involuntary from the customer's or subscriber's perspective.

Service termination can harm the company's reputation as well as its retention and acquisition rates, particularly if it is done involuntarily. However, service termination initiated by service providers, including telecom operators such as ethio telecom, is rising and becoming more widespread [5]. Ethio telecom is Ethiopia's sole telecom service provider, offering a variety of telecom services such as fixed-line, mobile, broadband, and so on. Ethio telecom has over 54 million mobile subscribers and a considerable quantity of infrastructure installed throughout Ethiopia [6].

In Ethiopia, subscriber service termination occurs as a result of any telecom service provider company's implementation, which can be voluntary or involuntary. According to [4], ethio telecom lost around 589,338 of its mobile service subscribers within a quarter (November 2018-January, 2019) due to service termination, regardless of whether the implementation approach is voluntary or involuntary. Furthermore, according to ethio telecom semi-annual report 2021(January2021-May 2021) and showed from customer relationship management (CRM) system which deployed in ethio telecom, there is a postpaid mobile service subscribers termination with a significant number, and more of those subscribers terminated involuntary, implying that ethio telecom initiated the termination. In contrast, involuntary service termination implemented without superior information and an adequate management approach is deemed to advocate service termination as a good business strategy [5].

One of the primary business strategies of telecom service providers identify subscribers who are about to quit telecom services and retain them with enticing offers and discounts [4]. Because recruiting new clients is five to six times more expensive than retaining existing customers, in addition increasing retention rate by 5% increases earnings from 25% to 95% [4, 7, 8]. When a telecom subscriber leaves service, the operator loses not just the customer's money but also the cost of material, labor, and time resources used to attract those subscribers.

Customer data is critical for sustaining company income and customer satisfaction through the development of an appropriate strategic plan to provide proactive attractive customer service, establish a better marketing strategy, and develop a targeted retention plan. As a result, identifying terminating customers early is crucial by analyzing large amounts of data at the service provider's hands using a machine learning approach that may be easily implemented to enable develop different models [4].

The data-driven prediction model is necessary to identify mobile subscribers will be terminated near the future in advance to prepare a strategic plan and retain customers.

Telecom operators generate a huge amount of data due to various reasons such as the increasing number of subscribers, rapidly renewable technologies; data-based applications, and another value-added service [9]. This data is used to build data-driven prediction models with different machine learning algorithms that enable them to identify their customers who are likely to terminate service (churn) nearly in advance and offers a firm with important knowledge to retain and improve its customer base [4].

A good prediction model should not only correctly identify prospective churners (terminated service), but also offer causes for their churn and anticipate such results over a sufficiently extended time horizon [10]. Quality and adequate data are required to develop a good prediction model since machine learning algorithms cannot learn successfully without enough and quality data and results may be garbage in garbage out. Quality of data will be get through data preprocessing including data cleaning, integration, aggregation, and feature selection or data reduction [11].

The prediction may be characterized as a classification. According to the [12] definition in machine learning, classification is the separation of data into various parts/classes/groups and is used to predict which dataset the incoming data originates from. From types of classifications: Binary classification is the process or activity of categorizing a given set of data into two categories, whereas multi-class classification is the process of categorizing items into several classes. It is not limited to a certain number of classes, unlike binary.

1.1 STATEMENT OF THE PROBLEM

Telecom service providers face several issues, one of which is service termination or customer churn, even when customers have no option and deals with monopoly telecom operators [4]. It has revenue consequences for operators, particularly lost postpaid mobile users, because operators expected to collect more money from postpaid mobile service subscribers than from prepaid subscribers, and postpaid subscribers are limited in number. Thanks to researchers different customer churn or service termination prediction models proposed to tackle the challenge. However, existing developed prediction models are more focused on:

- Prepaid mobile service: Most mobile users are prepaid subscribers and focused on them, but at the same time, emphasizing postpaid subscribers is vital, since a loss of postpaid customers means a loss of more money [9, 13].
- Voluntary churn prediction model: Focusing only identifying a terminated voluntary subscriber from an active subscriber. As a result, focusing solely on voluntary churners may reduce the cost of offering benefits and allow for the retention of all churners (voluntary and in-

voluntary) [14]. Since involuntary service termination has grown and become popular [5].

- Model of binary (terminated/not-terminated) prediction: This prediction indicates that two forms of termination (voluntary and involuntary) are considered as one, each with a separate termination initiation.
- One type of feature selection evaluation approach: Feature selections are key tasks in model development, and utilizing more than one technique to choose the best features will be better. Because feature selection not only reduces space and time complexity but also improves prediction accuracy [11].

The questions that will be going to be answered in this research are:

1. What features or attributes are potentially used to predict postpaid mobile service termination?
2. Which classification technique is more appropriate to predict possible service termination of postpaid mobile subscribers?

1.2 OBJECTIVES

1.2.1 *General Objective*

The general objective of this research is to develop a multi-class prediction model of postpaid mobile subscribers service termination with a machine-learning algorithm and propose best features.

1.2.2 *Specific Objectives*

The specific objectives of the thesis are as follows:

- To identify the best attribute or features for postpaid mobile service termination prediction.
- To demonstrate the impact of feature extraction on the performance of postpaid mobile service termination prediction.
- To select the best algorithm for postpaid mobile service termination prediction.
- Build prediction model and evaluate the performance of the models.

1.3 SCOPE AND LIMITATION OF THE THESIS

1.3.1 *Scope of the thesis*

The scope of this thesis focused on developed a termination prediction model of postpaid mobile service termination in telecom operators, and features extracted from CDR and customer profile only, implying that features from the network side not be included in this model building. It is also focused on developing a multi-class service termination prediction model. In this investigation, only outgoing voice calls, SMS, and data upload and download CDR data were used.

1.3.2 *Limitation of the thesis*

This thesis CDR data were collected from ethio telecom CBS system of three months to each sampling service number and it is obvious that the three months data may or may not have the same information to investigate the customer usage behavior in a specific month. Therefore change this three months CDR data to a month on average which may lose high and low usage customer behavior for a specific month if have any within the three months.

1.4 CONTRIBUTIONS OF THE RESEARCH

This study offers value with developing a better postpaid mobile service termination multi-class prediction model by including one new feature which is not included previously that can effectively and efficiently predict to enable identifying two types of possible postpaid mobile subscriber service termination separately based on CDR usage and selected customer profiles by using machine learning algorithms. As a result, telecom operators, especially ethio telecom, can profit from the following:

- Identifying postpaid mobile services that will be terminated shortly and taking proactive measures to keep them as much as possible.
- Identifying postpaid mobile service will be a potential terminated and developing a strategic plan based on targeted consumer data.
- To develop an appropriate strategic plan based on a reasonable customer base.

1.5 LITERATURE REVIEW

Churn in literature is defined as defection, customer loss, attrition, or turnover. Churn is simply used as the act of a subscriber switching from one service provider to another [4, 7, 8, 10, 15]. Termination of a subscriber is defined as only quitting a service from a mobile operator and considered as churn when there is no other competitor to leave for in the case of a monopoly market like ethio telecom [4]. According to [4], the main feature that distinguished churn prediction analysis from service termination prediction is Off-net calls. Off-net call means a call, which makes between different network operators within inland.

Different researchers provide customer churn prediction models with support machine learning more focus on the competitive environment and prepaid mobile service. Author[4], conducted prepaid mobile subscriber termination prediction in a monopoly telecom service provider by comparing three supervised machine learning algorithms those were random forest, J48 decision tree, and Naive Bayes with combined the most 7 prediction features selected by information gain and correlation coefficient. Consequently, results showed that RF scores the highest prediction accuracy while Naive Bayes scores the least by using confusion matrix-based validation such as accuracy, precision, recall, and F-measures.

All churning customers may not have the same churning factors and attribute ranking is used to identify the factors and hidden patterns in data that are the main reasons for churning. Author [7] developed a churn prediction model by classifying churn and non-churn customers with different classification algorithms such as random forest, J48, random tree decision stump, and so on. This study utilized information gain and correlation attribute ranking filter techniques in the WEKA tool to select features in result used 17 high-ranking value attributes out of 29 to the developed churn prediction model. The author evaluated constructed model with cross-validation technique (K=10) included accuracy, precision, recall, f-measure, and receiving operating characteristics (ROC) area performance metrics. This study also provided factors of churn by using attribute selected classifier enables to generate rule, which is important to prepare specified retention plan. This research, results showed that RF and J48 algorithm performance have better than the others have, and then performed customer-profiling (clustering) for churn customers who identified by random forest before using a k-means clustering algorithm. The k-means algorithm segmented the data into three groups those are low, medium, and risky due to the nature of the data. Furthermore, it also provides retention guidelines for decision-makers based on factors behind the churning of churn customers through the rules generated by using the attribute-selected classifier algorithm.

Features used to build a churn prediction model can be extracted from the application of social network analysis. According to [15], the two quantities that can be calculated for each customer in the network are centrality degree and PageRank. Author [15], developed a churn prediction model that can output the probabilities that customers will churn soon with including features extracted from six sources those are demographics, billing data, refill history, calling pattern, CDR billed and network. This study compared the C4.5 decision tree, alternating decision tree, Naive Bayes, and logistic regression to each other and discovered that classifiers achieved above 60% overall accuracy for postpaid and above 70% overall accuracy for prepaid on a testing dataset of 30 and 40 thousand instances, respectively.

In customer churn model-building, most of the feature sets are extracted from call detail records (CDR) and customer profiles. Author [16], Derived a new subset of features to improve the accuracy of customer churn prediction in the wireless telephony industry and categorized them as contract-related, call pattern description, and call pattern change description features. All stated existing prediction models build on a binary classification basis and all have considered two types of customer churn or service termination as one. In this study, developed a multi-class service termination prediction model.

1.6 METHODOLOGY OF THIS RESEARCH

To reach the objective of this research the following processes will be performed:

Data collection: Data collected from various ethio telecom systems. CDR data have many features like voice calls, data consumption, sent SMS, and so on, and collected from ethio telecom convergent billing system (CBS). Customer profile contains customer's service status, customer bill cycle ID, customer demographics, and so on data are obtained from customer relationship management (CRM) system.

Data prepossessing: Main activities performed at this stage were integrating data gathered in.csv format from systems, cleaning missing values, detecting and replacing outliers, deleting unnecessary characteristics, aggregating three months of CDR data as one month per average, data balancing, and feature selection. Before using ML algorithms, all data sets were eventually translated to a suitable format.

Model building and testing: we conducted model training and testing with supervised machine learning on three algorithms: J48, RF, and ANN(MLP), using WEKA 3.9.4 open source tools and cross-validation (k=10) evaluation techniques.

Analysis result and report finding: In this section, compared and contrasted the results are obtained from the experiment of three prediction algorithms based on different performance measurement parameters like accuracy, precision, recall, and F-measure. Then select the best appropriate one based on result for postpaid mobile subscriber service termination prediction.

1.7 THESIS ORGANIZATION

The remainder of this paper is arranged as follows: The chapter 2, provides background information on mobile services, types of mobile services, mobile service termination, and types of mobile subscriber service termination, the difference between service termination and customer churn, and the benefits of predicting customer churn or service termination. In the chapter 3, The machine learning concept and selected classification algorithms for this study are thoroughly explained in this section. Chapter 4 points out all the data preparation and experimental procedures. It starts from the data collection, preprocessing and prepaired expermental setup. Chapter 5 illustrates the main findings based on the results from the experimental analysis and performance of the three supervised machine learning algorithms towards predict postpaid mobile service termination . Finally, chapter 6 holds conclusions and recommendations based on the results from the experimental works. It also tries to point out some future works.

2

MOBILE SERVICE

Mobile services (m-service) or electronic services (e-service) using mobile devices (such as mobile phones, handheld computers, or vehicle interfaces) and wireless telecommunication networks have become hot topics and market research in the information systems community. Recently, the potential of mobile service applications has led many organizations to spend heavily on these technologies. M-service provides values that traditional wired e-commerce does not possess, such as ubiquity, personalization, flexibility, and dissemination.

Providing customers with value-added, interactive, and/or location-based mobile services (such as banking, content downloads, emergency/roadside assistance, and wireless coupons) seems to be increasingly important gains in the mobile market by strengthening the relationship with key customers [17].

The term mobile service is used to describe all services that can be used regardless of time and space restrictions and can be accessed via mobile phones (mobile phones, PDAs, smartphones, GPS, etc. Mobile services are extremely efficient in reaching out to specific clients.

The customer relationship can be divided into discrete transactions and ongoing connections. Discrete transactions may be viewed as episodes that reflect a series of various interrelated acts, whereas a continuous connection is built on several interconnected sequences. Episodes represent a client's transactions, a continuous connection frequently incorporates some type of agreement between the consumer and service provider. For mobile services, this indicates that ongoing partnerships are mostly associated with contract-based services [18].

The factors that have an impact on the evolution of mobile services are rapid internet growth, increasingly data-hungry applications, data overtaking voice traffic, users want mobility, voice prices are decreasing, mobile internet has a huge potential, voice revenues have a bleak future, and access to desktop applications is required when mobile [19].

Mobile services vary from traditional services in that they may deliver service offers independent of time and space restrictions. They are also distinct from conventional face-to-face interpersonal services or other forms of e-services, such as wireless internet services, in which service delivery is tied to a specific fixed local area network or a specific place. Mobile services may be accessible while on the go, wherever and whenever they are required [18].

Mobile services will allow customers to make purchases, request services, access news and information, and pay bills via mobile communication devices such as PDAs, laptops, and cellular phones. Some of the key drivers for mobile services are [20]:

Mobility: Mobility is the key benefit of mobile services. Through internet-enabled mobile devices, users may access whatever information they want, anytime they want, regardless of their location. Mobile services meet the demand for real-time information and communication at all times. Mobile users may be engaged in activities such as meeting people or traveling while making purchases or getting information via their internet-enabled mobile devices.

Reachability: Customers/constituents can be reached by business/government organizations via mobile devices at any time and from any location.

A mobile terminal allows a user to communicate with and be accessible to others at any time and from any location. The user can also limit his or her reachability to specific people or times.

Localization: Knowing a user's physical location at any one time offers tremendous value to mobile services. Many location-based services may be given using available location information. For example, knowing a user's location allows the mobile application to immediately inform him or her when a friend or colleague is nearby.

Personalization: Although the internet now contains vast volumes of information, services, and applications, not all of it is useful to all users. Mobile services can be customized to filter information or deliver services in ways that are more relevant to a specific user.

2.1 TYPES OF MOBILE SERVICE

The number of individuals who use cell phones is increasing all the time and mobile services may be divided into several categories from a different perspective like mobile infrastructure, mobile technology, subscription, etc. According to the author [20], mobile services are categorized into four major areas based on mobile infrastructure and mobile technology, those are

Mobile business-to-business (B2B): Mobile business-to-business transactions encompass both intra- and inter-organizational transactions. Mobility incorporated into organizational operations can improve information flows, coordinate operations throughout the company, and therefore improve supply chain management.

Mobile business-to-consumer (B2C): Mobile B2C transactions are retailing transactions with individual shoppers, often known as customers or consumers. Mobile services can provide clients with automated, unassisted operations straight from mobile terminals.

Mobile consumer-to-consumer (C2C): Customers sell directly to other consumers using mobile devices in the mobile C2C sector. Individuals will soon be able to use their mobile phones to find possible purchasers for their houses, automobiles, and other assets, engage and bargain with them, and complete deals.

Mobile government: As the internet goes wireless, the mobile government is an extension of e-government and allows people to access government agencies and organizations via websites on the internet, regardless of their physical location or time of day, using their mobile phones.

Based on subscription prepaid and postpaid mobile phone subscriptions are the two types of contracts available for purchasing mobile services and in this section see one by one both types of mobile service subscription.

2.1.1 *Prepaid Mobile Service*

A prepaid, prepay or Pay As You Go (PAYG) subscription is one in which clients pay for mobile services in advance by obtaining credit for the service allowance. This subscription usually includes a set allowance or rate for mobile services such as voice calls, text messaging, and mobile internet data consumption and charges. The prices for these services are deducted from the consumers' prepaid credit at set rates when they utilize them. Once a customer's credit has been used up, the network will not allow utilizing any more services unless they recharge or top up their credit.

Prepaid service decreases the service's operational costs significantly, according to the service provider may avoid the extra costs of credit checks and fees since no services will be given if the user does not deposit enough money into the account. Prepaid services, from the customer's perspective, give quick services without the need to sign long-term commitment contracts and allow for better budget control. According to [1, 2], the new electronic payment scheme will most likely allow for short-term contracts for postpaid services, but such solutions are not commonly employed, and most prepaid consumers do not want any contracts.

If the balance is depleted, the customer will not be able to make calls but will be allowed to receive calls during the period (for example, six months) so that to resume prepaid services, customers must purchase a rechargeable card and top up a balance [1, 2].

In general, a prepaid subscription requires us to pay in advance for the services we expect to use, whereas a postpaid subscription requires us to pay monthly for the services we use throughout the month and telecom operators often earn less income from prepaid subscribers than postpaid subscribers do; as a result, those operators focus primarily on postpaid customers.

2.1.2 *Postpaid Mobile Service*

In mobile communications, a postpaid, post-pay, or pay-monthly subscription refers to a service contract in which a user is billed for mobile services at the end of each month after they have consumed the services to which they are entitled. Postpaid subscriptions are packaged and sold to a variety of customer sectors, including individuals and businesses. Customers are given a set number of fixed or unlimited mobile voice minutes, fixed or unlimited text messages, and often a set quantity of mobile data as part of their subscriptions. If a customer uses more data than their allocation, they will be charged at a rate set by the mobile service provider. When a customer wants to acquire a postpaid subscription, the mobile operator or service provider will generally perform a credit check to guarantee that customers who buy postpaid services can pay their bills on time and have a worthy record of accomplishment.

Postpaid subscribers are fewer in number, but they make a significant contribution to an operator's revenue. They generate more consistent income than prepaid subscribers, who only top up their phones when necessary. The loss of postpaid customers is a significant challenge for operators, and telecoms work hard to retain these customers by offering them tailored plans. Marketing campaigns aimed at the right customers can help to reduce operational costs while also improving overall business performance [2, 3].

In the telecoms sector, mobile carriers want to attract postpaid users and grow their incoming income based on postpaid line usage. When compared to postpaid lines, customers choose to acquire and utilize prepaid mobile lines owing to the simplicity of use and greater control over the cost of the line.

2.2 MOBILE SERVICE TERMINATION

Service terminations are common and occur in a wide range of industries, including finance, telecommunications, healthcare, and professional services and one of the primary reasons for the service's termination is a lack of profitability. Service provider-initiated service termination has grown in popularity may have an impact on company's reputation as well as its retention and acquisition rates. On the other way, it may look premature to advocate service termination as a good business strategy without greater knowledge of terminated customers' reactions to service termination and whether the process might be managed more effectively [5].

As stated [21], on average, when a serviced customer experiences a low level of quality, they are more likely to complete the service than a customer who experiences a higher average quality. This may be due to a variety of reasons, including disappointment and/or inability to use the service (for example, the use of mobile phones by street vendors depends on the quality of service at the travel location).

2.3 DIFFERENCE BETWEEN SERVICE TERMINATION AND CUSTOMER CHURN

The difference between customer churn and service termination may be raised from two perspectives which are definition and features. Customer churn is referred to as attrition, defection, or turnover. According to [4], in the telecom industry, most works of literature agreed that churn is the loss of existing subscribers to another service provider, and also churn is sometimes defined simply as customer loss whereas service termination, as opposed to churning, is defined as quitting a service without leaving it to competitors, Since in a monopoly market, customers (subscribers) have no choice but to quit. Also, in this study, the prediction of subscribers' service termination and churn have a common benefit and how subscribers terminate or churn is strikingly similar as stated by the author [4].

Any user of operator services is referred to as a customer, whereas a customer with a subscription is referred to as a subscriber. From a feature perspective, In the event of a monopoly market, there are no other competitors to choose from the current service provider. On-net calls are the primary characteristic that separates churn prediction analysis from service termination prediction analysis. On-net calls are calls between the same network. It includes the duration of calls, the number of calls made, and the number of SMS messages sent or received. Off-net calls are distinguished by the fact that they are made with another network (service provider). They are mostly utilized in the presence of other networks or rivals. Otherwise, all of the features and methodologies used in churn prediction may be utilized to predict subscriber service termination [4].

2.4 TYPES OF CUSTOMER CHURN OR SUBSCRIBER SERVICE TERMINATION

A subscriber may quit the operator's service voluntarily or involuntarily; leaving voluntarily is further subdivided into two sub-categories: intentional and incidental. According to [22], Churns can be categorized into various types depending on two aspects those are revocation methods and business aspects. Churns are categorized as voluntary or involuntary based on the first aspect (revocation process):

Voluntary Churn: This form of churn occurs when the user decides to cancel his service contract subscription and is further subdivided into two categories: deliberate and incidental. Deliberate leaving might occur because of issues linked to subscriber sensitivity contrasted to subscriber expectations, such as excessive service price, poor service quality, poor customer service, and other associated inconvenience factors and it is the problem that most company strives to address. Incidental churn occurs when changes in situations prevent the subscriber from continuing to subscribe to the offered services. For example, relocate to a different area where the company's services are not available. Incidental churn accounts for a small portion of an organization's voluntary turnover [4, 22].

Involuntary Churn: When a company decides to discontinue a customer's subscription service referred to as involuntary churn. Misuse of the service, nonpayment, and credit card expiration are some of the reasons for termination [4, 22].

In the second aspect (business setting), churn is broadly classified as contractual churn and non-contractual churn which means customers have the option to buy or not buy at any anytime.

Contractual Churn:Customers would execute an activity at periodic times in a contractual churn and revocations are expressly noticed in this form of churn [22].

Non-contractual Churn: Noncontractual churn occurs when customers cease to engage in certain behaviors over time. This describes situations in which churn cannot be detected quickly on a specific action taken by a customer. As a result, analyzing this nature of churn is difficult. An example of this churn is an online fashion shop, where clients can buy or not buy at any moment [22].

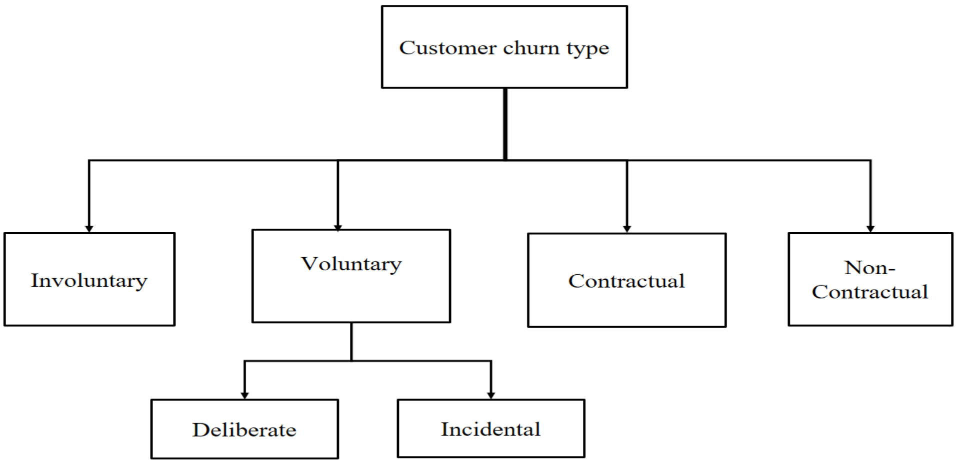

Figure 2.1: Customer churn type (22)

2.5 BENEFITS OF PREDICTING CUSTOMER CHURN OR SERVICE TERMINATION

Churn is defined in the telecommunications paradigm as the activity of customers leaving the company and discarding the services offered by it when customers are unhappy with the services and/or better offerings from other service providers within the customer's affordable price tag, which leads to a potential loss of revenue or profit to the company. Customer churn refers to customers moving to a rival or canceling their membership to a service. Whereas service termination simply quite services from a service provider and the churn rate is an essential metric that all companies want to reduce, as a result, churn prediction is an essential component of a proactive customer retention strategy since attracting new customers into a business is a more time-consuming and expensive work. As a result, rather than focusing on new customer acquisition, a firm should prioritize retaining current customers [23, 24].

Churn prediction and control have become top priorities for telecom companies. Customer churn prediction enables telecom operators to make customized offers to interested potential customers to increase their use, and the forecast may be used as an input to taking proactive action to keep clients will reduce churn in the near future, to manage overall income expectations, and to develop a proper marketing plan and retention initiatives. Prediction of service termination in a monopolistic market provides similar advantages as churn prediction of the market have competitors [4].

3

MACHINE LEARNING

Machine learning is a subfield of artificial intelligence that attempts to enable computers to do skilled tasks by utilizing intelligent software. Statistical learning methods are the foundation of intelligent software used to build machine intelligence. Because machine-learning algorithms require data to learn and must be linked to the database discipline. Machine Learning is a natural progression from the combination of Computer Science and Statistics [25].

As stated by authors [25, 26], A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E. The goals of machine learning are to enable machines to generate predictions, cluster data, extract association rules, and make decisions based on a given dataset.

3.1 TYPES OF MACHINE LEARNING

Machine learning can easily be used to assist in the detection, classification, and prediction of subscriber service termination. It is possible to identify terminated subscribers early by analyzing massive amounts of data at the hand of a service provider. Many prediction techniques can be used to find patterns in actual data in a system to support predictive analysis and classification. As a result, the operator can use a proactive retention operation or other approaches to prepare for the consequences of subscriber loss [4]. Machine learning refers to the ability of machines to learn without being explicitly programmed and machine-learning scenarios differ from each other based on the types of training data available to the learner. There are four general machine learning methods: supervised, unsupervised, semi-supervised, and reinforcement learning methods [25, 27, 28] described as follows on figure 3.1.

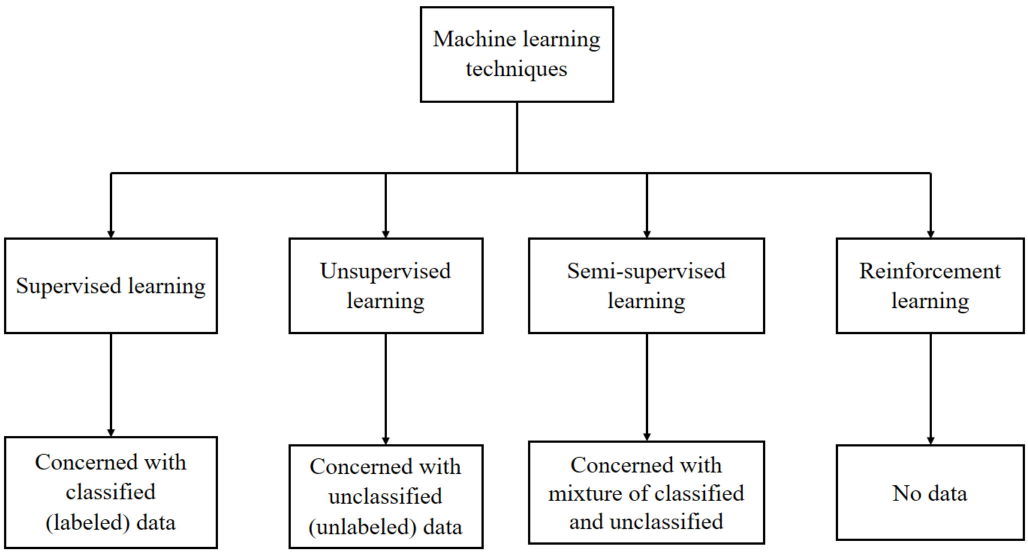

Figure 3.1: Machine learning techniques and their required data (25)

Supervised learning: The computer is given sample inputs and their intended outputs by a "teacher," and the objective is to learn a general rule that maps inputs to outputs.Two groups or categories of algorithms come under the umbrella of supervised learning are regression and classification.

Unsupervised learning: Non-labeled are assigned to the learning algorithm, leaving it to identify structure in the data on its own. Unsupervised learning can be a goal by itself (discovering hidden patterns in data) or a means to an outcome.

Semi-Supervised Learning: The supplied data in semi-supervised learning are a combination of classified and unclassified data. This mix of labeled and unlabeled data is utilized to create a suitable model for data categorization. The goal of semi-supervised classification is to build a model that can predict classes of future test data better than a model built only from labeled data.

Reinforcement Learning: The reinforcement learning approach attempts to perform activities that maximize the reward or minimize the risk based on observations obtained through interactions with the environment.

3.2 TYPES OF CLASSIFICATION

Classification is a process that may be accomplished by utilizing machine learning algorithms that learn how to assign a class label to instances originating from the issue areas, such as spam detection, churn prediction, and so on. Binary classification, multi-class classification, multi-label classification, and unbalanced classification are the four primary types of classification problems [29].

Binary classification: Classification problems that contain two class labels are called binary classifications. Typically, binary classification tasks include one class representing the normal state and another representing the abnormal state. Logistic Regression, k-Nearest Neighbors, Decision Trees, Support Vector Machines, and Naive Bayes are some popular binary classification techniques.

Multi-Class Classification: Multi-class classification problems include more than two class labels, and some of the algorithms that may be used to do multi-class classifications are k-Nearest Neighbors, Decision Trees, Naive Bayes, Random Forest, and Gradient Boosting. In contrast to binary classification, Multi-class classification does not contain the idea of normal and abnormal results whereas instances are classified as belonging to one of many recognized classes.

Multi-Label Classification: Multi-label classification refers to classification tasks with two or more class labels, where one or more class labels can be predicted for each example. Multi-label classifiers can able to output multiple tags at the same time. For example, we may categorize movies according to their type: comedy, drama, romance, horror, and so on. In this situation, one movie might be classified as both a comedy and a romance [30]. Classification algorithms designed for binary or multi-class classification cannot be used directly for multi-label classification; instead, specialized versions of standard classification algorithms such as Multi-label Decision Trees, Multi-label Random Forests, and Multi-label Gradient Boosting must be used.

Imbalanced Classification: Imbalanced categorization tasks have an unbalanced distribution of examples in each class. Unbalanced classification issues are often binary classification tasks in which the majority of the instances in the training dataset belong to the normal class and a minority of examples belong to the abnormal class. These tasks are classified as binary classification tasks, however, they may necessitate the use of special techniques to change the sample composition in the training dataset by undersampling the majority class or oversampling the minority class. One of the examples provided in this area is fraud detection.

3.3 CLASSIFIERS

The process of allocating an instance to one of the predefined classes based on its characteristics is referred to as classification and it is one of the most significant data mining challenges. The input of a classifier is a series of training records, each with a different set of attributes. The numerical characteristics or attributes have numerical domains that mean that contain numerical data types and categorical attributes contain non-numerical data types. There is also a distinguishing feature called class label. The purpose of classification is to create a console model that can predict the class label of future unlabeled records [31]. Prediction models can be developed by using a classifier algorithm and different algorithms proposed for this purpose three of them are J48, Random forest, and Multilayer Perceptron which is a type of artificial neural network that are utilized in this research, and are found under supervised machine learning methods described as follows.

3.3.1 *J48 Decision Tree*

Classification is the process of constructing a class model from a set of data that include class labels. The decision tree algorithm is used to determine how the attributes-vector acts in a variety of situations. A decision tree is a flowchart-like tree structure in which each internal node represents a test happening an attribute, each branch represents a test ends, and each leaf node or terminal node represents a class label. The attribute value of each tuple is checked adjacent to the decision tree for each tuple. A route is traced from the root to a leaf node that contains the class prediction for the tuple. The process of converting decision trees into categorization rules is straightforward. A decision tree is used in decision tree learning as a predictive model that translates observations about an item to inferences about the object value of the item. It is a predictive modeling method used in statistics, data mining, and machine learning. Classification trees are tree models in which the object variable can take a limited set of values. Within this tree structure, leaves correspond to class labels and branches reflect the combination of characteristics that lead to specific class labels. When compared to other classifier algorithms, decision trees may be can built models very quickly [31].

In the WEKA data-mining tool, J48 is an open-source Java implementation of the C4.5 algorithm. C4.5 is a software that generates a decision tree from a collection of labeled input data, and the decision trees it generates can be used for classification [31]. This algorithm produces understandable results and can handle both continuous and categorical target variables. J48 decision tree algorithms need less data cleaning and perform well on huge datasets. One disadvantage of J48 is that it is prone to overfitting and may have an impact on tree visualization, which may be addressed using the pruning approach [32].

3.3.2 *Random Forest*

Random forest (RF) is a decision tree classification approach composed of a set of decision trees. Each tree is reliant on the values of a vector separately, with the same distribution across all trees in the forest. The error associated with a model of this classifier is mostly determined by the strength of the individual trees in the forest as well as their connection between trees [33]. Random forest is a powerful technique that may be used for regression as well as classification and easy-to-use and is also the most popular machine learning algorithm that can consistently give high results even if without hyperparameter adjustment. The random forest has both advantages and disadvantages. One of the advantages is that it can be used for both classification and regression problems, which are common in machine learning systems nowadays. The random forest also prevents overfitting since it contains a group of trees, and if there are enough trees in the forest, overfitting will not occur on the model system. It also can handle many input features without feature deletion, it is efficient on big datasets, and feature selection is available. Its main disadvantages are interpretation complexity, slow and ineffective in real-time predictions since the random forest is built from a large number of trees [32, 34, 35].

3.3.3 *Artifical neural network (Multilayer Perceptron)*

Multilayer perceptron (MLP) is a classifier in which the weights of the network are found by solving a quadratic programming problem with linear constraints, rather than by solving a non-standard neural network training [36]. This algorithm is an artificial neural network and is made up of three layers: input, hidden, and output. Except for the input layer, all layers employ a non-linear activation function. The activation function connects the input and output in a rounded way. This method allows for learning by adjusting the weights of the neurons it assigns to itself. Multilayer perceptron can be applied to complex non-linear problems; works well with large and small input datasets, and can provide quick predictions after training. The disadvantages of multilayer perceptron are computations are difficult and time-consuming, and the functions of the model depend on the training data quality [35].

4

DATA PREPARATION AND EXPERIMENTAL SETUP

This chapter describes data preparation and experimental setup which is the most important and a critical task to get the desired output. The overall process to conduct this research is depicted as follows in figure 4.1.

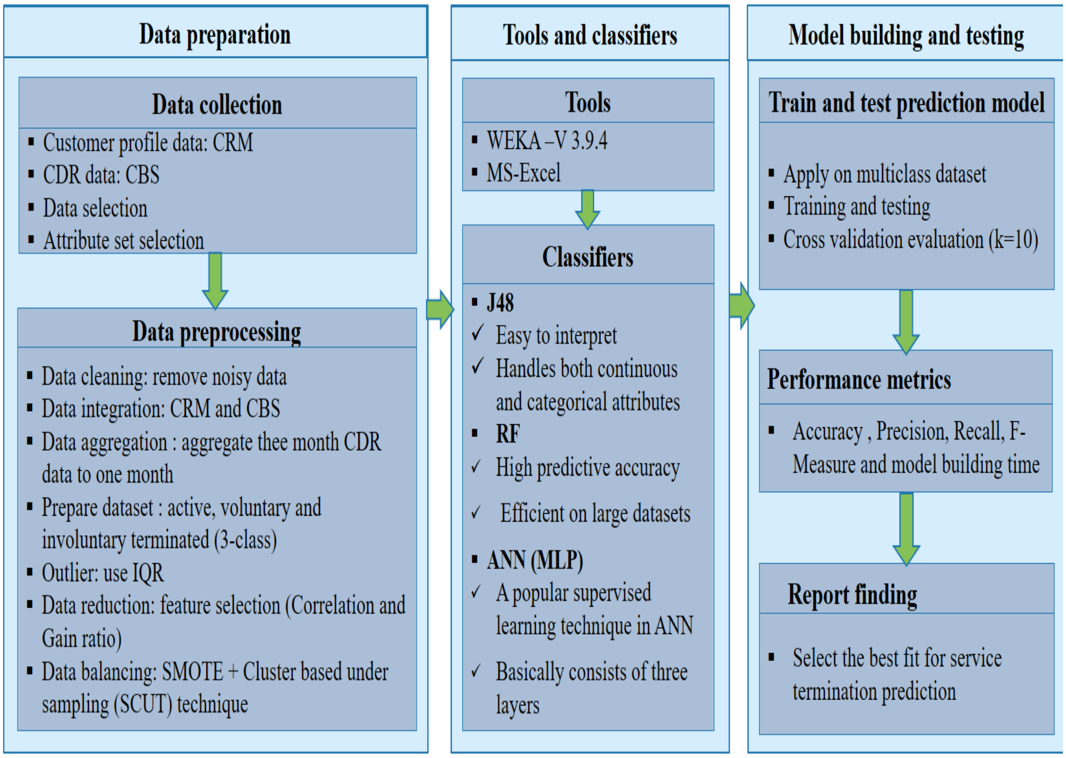

Figure 4.1: Overall research methodology

4.1 DATA PREPARATION

Data is the most important component of data analytics, machine learning, and artificial intelligence because without it no model can be trained. In machine learning, data is vital, and the learning algorithm is used to discover and gain information or characteristics from the data. The quality and amount of the dataset impact learning and prediction performance, which means machine-learning algorithms require adequate and clean data to train and test for model construction [28].

4.1.1 *Data Collection*

CDR data collection began following the approval of data access requests from the company. The ethio telecom data center is used to collect all training and test data. Three months of total summary CDR data and profile information for mobile postpaid subscribers are extracted from the CBS and CRM system respectively.

Raw data, which are outgoing voice call, Send SMS, Upload and download data usage CDR data collected from ethio telecom convergent billing system (CBS) and customer profile get from customer relationship management (CRM) system. This data contains active and terminated service number with a three-month transaction CDR have been collected as follows.

For active service-number, we take 17,050 service numbers by using simple random sampling from active postpaid mobile subscribers up-to-date, June 2021, then three months (February 2021 to April 2021) of total number CDR data for individual numbers extracted from the CBS system.

For terminated service, number collect in two ways by using service-terminated reason which is called as per customer request and the system (failed payment to the bill) has a direct relation to voluntary service termination and involuntary service termination respectively. In this research, those two types of service termination are considering as individual classes or labels separately and the result is three classes including active service. For terminated service voluntary, numbers were collected from five months (January to May 2021) and we got 1007 service numbers, then the total number CDR data for each service number was collected for the last three months before the termination of those numbers. For terminated service involuntary, numbers taken from one month (May 2021) in number 3967 and three-months (September, October, and November 2020) CDR data collected for these numbers which are before termination with the same procedure as voluntary terminated service. In ethio telecom, CDR history data for terminated service numbers are only available on the CBS for a year, and the life cycle of a postpaid mobile takes 6 months from start the barring state to terminate, that is the reason only a month of involuntary termination service number data is collected.

4.1.2 *Field Selection*

In ethio telecom customers' CDR and profile data have a number of fields that may all have no relevancy in this study, so 15 fields were extracted from the CBS and CRM system including target variables that are considered the most important to this research. Bill cycle ID attribute is an attribute that is used to know subscriber billing collection period and it hasn't been included previously on the postpaid mobile service termination prediction model. The extracted fields with descriptions are shown in the table 4.1.

Table 4.1: Extracted fields from CBS and CRM systems

No. list	Attribute name	Descriptions
1	Service Number	To identify the subscriber
2	Customer Category	To identify the subscriber level type
3	Service activation date	To know subscribers activation date
4	Service termination date	To know subscribers termination date
5	Vice usage in Minutes	Subscriber's interest to spend time on voice usage
6	Number of Calls	To measure subscribers how often to make a voice call
7	Voice Fee ETB	Subscribers sensitivity in spending money on voice call
8	Data usage in MB	Subscriber's interest to spend time on data usage
9	Data usage freq.	Subscriber's interest to use data
10	Data Fee ETB	Subscriber's sensitivity in spending money to use data
11	Number of SMSes	Subscriber's interest how often to send SMS at all
12	SMS Fee ETB	Subscribers sensitivity in spending money on SMS
13	Total Fee ETB	Subscribers sensitivity in spending money on mobile ser.
14	Bill cycle ID	To know subscriber billing collection period
15	SIM-Status	To rank or label subscribers' stage

4.1.3 *Data Preprocessing*

Data quality and representation are some of the most important factors determining the performance of machine learning (ML) on a given assignment. When there is a lot of irrelevant and redundant information or noisy and erroneous data, knowledge discovery becomes more difficult during the training phase. Data preparation frequently influences the generalization performance of a supervised ML algorithms and minimization of noise occurrences is one of the most difficult issues in inductive ML. Data preprocessing covers operations such as data cleansing, normalization, transformation, feature extraction and selection, and so on. It is generally known that data preparation and filtering procedures consume a large amount of processing time in ML issues, and the final training set is the outcome of data pre-processing [37].

4.1.3.1 *Data Cleaning*

The process of discovering and eliminating problems from data is known as data cleaning, and inconsistent data conventions between different sources, such as different abbreviations or synonyms, data entry errors such as spelling mistakes, inconsistent data formats, missing, incomplete, outdated or otherwise incorrect attribute values, data duplication, and irrelevant objects or data are common anomalies [38]. In our data instances with no values connected to voice service, which are numbers given service data alone are removed from the collected data by utilizing the main service type.

4.1.3.2 *Data Integration*

Data integration is the activity of integrating data from several sources to give the user a unified view of that data [39]. Collected data related to voice service, data service, and SMS service from CBS was integrate into one table, and another integration was performed between this data and customer profile data which was collected from CRM by using service number as a primary key for this study.

4.1.3.3 *Data Aggregation*

Aggregate is any process in which information is gathered and summarized in a summary form to obtain more information about certain groups based on specified variables, which is a typical objective of aggregation. The three months collected usage CDR data aggregate to one month in average per service number basis.

4.1.3.4 *Data Labeling*

Data labeling is the process of recognizing and tagging data samples in the context of machine learning. Labeling data instances is a key activity in machine learning and visual analytics, and it is the foundation of supervised machine learning [40]. According to stated by author [41], current research estimate that data preparation and labeling accounts for more than 80% of engineering activities in a machine-learning project. The obtained data were labeled as follows in this study: active subscribers were labeled as active service, terminated subscribers with marked "per customer request" reason labeled as voluntary termination service, and terminated subscribers with marked "failed to pay the bill (by the system)" reason labeled as involuntary termination service, with the number of sample instances for each class described in the table 4.2.

Table 4.2: Labeled data

Active service no. samples	**Voluntary terminated service no. samples**	**Involuntary terminated service no. samples**
17,004	1,007	3967

4.1.3.5 *Dataset Preparation and Formatting*

A dataset (data set) is a collection of data that most frequently corresponds to the contents of a single database table or a single statistical data matrix, where each column of the table represents a specific variable and each row relates to a specific member of the data set under consideration. In addition, datasets are not limited to just numbers and text and may include collections of images or videos.

To get valuable information from data it must be structured. Dataset preparation means labeling the data and being ready for feature selection. In this research, the prepared dataset has three classes, which is a part of multi-class and the name of class are active service, voluntarily terminated service, and involuntarily terminated service. This data set contains 17,004, 3967, 1,007 instances for active, involuntary, and voluntary termination respectively. The ratio of active, involuntary, and voluntary terminated 77.4%:18.0%: 4.6%. The format of a dataset is in comma-separated value (CSV) and attribute relation file format (ARFF) that are both acceptable in the WEKA tool.

4.1.3.6 *Outlier Detection and Handling*

Outliers are defined as observations that deviate from the other observations in the data collection, and outlier detection is the process of discovering those observations using various techniques [42]. Outliers can be any novel, abnormal, inconsistent, irrelevant, incorrect, or noisy information, and detecting outliers and analyzing big data sets can lead to the discovery of unexpected knowledge in areas such as fraud detection, telecommunications, weblogs, and online documents, among others [43]. In general, outlier identification is required for the processing of valuable, significant, and relevant data. As a result, outlier detection and handling become significant in data processing and result validation, and different techniques for managing outliers have been developed in data mining and machine learning [42].

To find and replace outliers, this study used interquartile range (IQR) approaches, which is the difference between the first and third quartiles. The first quartile, designated Q1 is the value in the data set that contains 25% of the values below it whereas the third quartile marked Q3 contains 25% of the values above it.

$$\text{The formula of IQR} = Q3 - Q1 \tag{4.1}$$

$$\text{Lower inner fence} = Q1 - (1.5 * IQR) \tag{4.2}$$

$$\text{Upper inner fence} = Q3 + (1.5 * IQR) \tag{4.3}$$

$$\text{Lower outer fence} = Q1 - (3 * IQR) \tag{4.4}$$

$$\text{Upper outer fence} = Q3 + (3 * IQR) \tag{4.5}$$

This fence is used to demonstrate minor and severe outliers. Points beyond the inner barrier in either direction are considered minor outliers, but points beyond the outer fence in either direction are considered extreme outliers [44]. This study used inner lower and upper fence values for replacing outliers to preserve the number of data samples and outlier data values around their source.

4.1.3.7 *Data Balancing*

The majority of learning systems assume that datasets used for learning is balanced [45]. However, more data gets for the model building have an imbalance between classes because it is natural positive cases happen in rare than negative in most of the application areas like investigation disease, customer churn, telecom fraud, etc. If there is data imbalance in the process of building a classification model the minority class may be biased by the majority and happen abnormal model training and generate inaccurate prediction performance results on testing [45].

The existing research proposed different solutions to solve data imbalance. The solutions can apply on two approaches are on data level and algorithm level. The data level approach applies before data feed to a classifier train and the algorithm levels approach can apply to strengthen the existing classifier to recognize the small class by adjusting the applied algorithm. The data level solution is implemented through data resampling which means resizing the data.

The resampling of sample data can be performed by using undersampling the majority class, oversampling the minority class, and combining them [46]. Data undersampling approaches include such as random undersampling, cluster-based under-sampling, One-Sided Selection, Condensed Nearest Neighbor Rule, Tomek Links, and oversampling approaches may use random oversampling, Synthetic Minority Oversampling Technique (SMOTE) which is used to generate synthetic or artificial data to minority class by using interpolation between minority class instance that are close together, and Borderline SMOTE [47]. According to [48], both undersampling and oversampling have advantages and disadvantages, thus combining them by utilizing superior methodologies from the two, such as cluster-based resampling from under-sampling and SMOTE from oversampling (SCUT=SMOTE+ Cluster-based undersampling) is better.

In this research, we have a severely imbalanced multiclass dataset in ratio 77.4%:18.0%: 4.6% to active service, involuntary termination, and voluntary termination service. Therefore to balance this dataset we have used the Cluster-based undersampling and SMOTE oversampling technique to resample and balance our dataset. To balance the multiclass dataset with the selected methods the following tasks are performed:

- First, calculate the mean (m) of the number of all instances of all classes divided by the number of classes.
- Used SMOTE oversampling technique to all classes that have numbers of instances less than mean(m).
- Cluster-based under-sampling is performed for all classes that have a greater number of samples than the mean (m).
- Used expectation-maximization (EM) techniques to conduct clustering which is used to cluster instances within each class have an advantage in performed clustering even number of clusters not provided by users, then instances randomly selected from each cluster such that the total number of instances from all the clusters is equal to m.
- At last, merge all the datasets generated by using both techniques and got final balanced dataset.

  The following tables (4.3, 4.4, 4.5) are describe undersampling, oversampling and final balanced data set respectively.

Table 4.3: Undersampling result

No. Cluster	No.Instance per cluster	No.Samples (43.08%)
Cluster 1	3395	1463
Cluster 2	7753	3340
Cluster 3	3978	1714
Cluster 4	1878	809
Total	**17004**	**7326**

Table 4.4: Oversampling result

Minority class	**No. Instance before SMOTE**	**No. Instance after SMOTE**
Voluntary terminated	1,007	7,326
Involuntary terminated	3,967	7,326

Table 4.5: Balanced dataset

# Active service	**# Voluntary terminated**	**# Involuntary terminated**	**# Total subscriber**
7326 (33.33%)	7326 (33.33%)	7326 (33.33%)	21978 (100%)

4.1.3.8 *Feature Selection*

Feature selection is an important process in the development of a prediction model, since chosen features can improve prediction model performance by decreasing time and space complexity. The technique of eliminating irrelevant and redundant attributes from a dataset is known as feature selection.

Nowadays, there are several feature selection evaluation techniques based on various functions such as distance, information, dependence, consistency, and classifier error rate measure [15]. In this study, we used two prominent and included in the filter method features selection evaluation techniques (Correlation and Gain ratio). The correlation and gain ratio features selection evaluation technique ranks the features based on their values in respect to the target variable or class.

Table 4.6: Feature selection result with correlation and gain ratio

Attirbute Name	Correlation	Rank	Attribute Name	GainRatio	Rank
Tenure_month	0.3597	1	Tenure_month	0.1411	1
Voice_usage_minutes_per_call	0.343	2	Data_connection_frequency	0.1136	2
Data_connection_frequency	0.2575	3	Voice_usage_minutes_per_call	0.0954	3
SMS_sent_frequency	0.2206	4	SMS_fee_ETB	0.0718	4
SMS_fee_ETB	0.2055	5	Data_fee_ETB	0.0525	5
Bill_cycle_ID	0.1889	6	Voice_fee_ETB	0.0447	6
Total_fee	0.1684	7	Bill_cycle_ID	0.0437	7
Voice_usage_inminutes	0.1567	8	Data_usage_MB	0.0432	8
Voice_fee_ETB	0.1292	9	Voice_usage_inminutes	0.0338	9
Data_fee_ETB	0.1214	10	Voice_call_frequency	0.0337	10
Voice_call_frequency	0.1116	11	SMS_sent_frequency	0.0336	11
Customer_category	0.0566	12	Total_fee	0.0269	12
Data_usage_MB	0.0433	13	Customer_category	0.0228	13

As we can see from the table above voice call frequency, customer category and data usage in MB attributes are which are found below top ten(10) based on correlation and SMS sent frequency, total fee, and customer category are based on a Gain ratio. Customer category is found under top ten(10) on both techniques. The newly included attribute, which is bill cycle ID is ranked on 6 and 7 with correlation and gain ratio respectively and this feature is used in this study to demonstrate performance change of algorithms with including and excluding on postpaid mobile service termination prediction model.

To select features appropriately we used two approaches to remove or select the same attributes from the two-feature selection evaluation technique after deciding the ranking threshold. The first approach remove attributes not selected on the top ten(10) in both techniques (Correlation and Gain ratio) and the second approach select attributes that have been selected in common on the top ten(10) in both techniques as if we take the threshold top ten(10). Feature selection evaluation techniques performance can be evaluated by prediction model performance since the best feature selection is considered as a problem and can solve by a feature selection algorithm [11]. Map the above two approaches to this problem and evaluate the approach through prediction model performance. J48 classifier techniques were used to evaluate the two approaches in this research and the performance results as shown in figure 4.2.

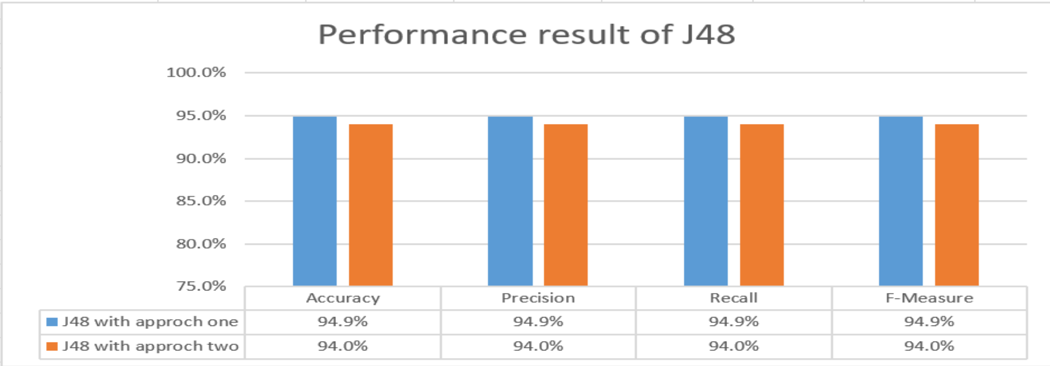

Figure 4.2: Feature selection evaluation approach result

Based on the above figure observation to select more relevant features use the first approach is better which means that removing attributes which not included in the top ten(10) with both (correlation, and Gain ratio) techniques. Therefore customer category features remove in this study and the remaining twelve(12) features are relevant.

4.2 EXPERIMENTAL SETUP

The experiment of this study applied on three selected classifiers and used two approaches across one attribute (including and excluding), which is called bill-cycle –ID is an attribute that has a nominal data type and is used to know subscriber's bill collection period where found in ethio telecom. Dataset used in this thesis is a balanced and multi-class dataset. The proposed algorithms for this study are J48, RF, and ANN(MLP).

4.2.1 *Performance Evaluation Metrics*

Classifier performance can be assessed based on different evaluation metrics which may determine the speed, accuracy rate, error rate of the classifier. Most classifier performance metrics are calculated based on confusion matrix result data. A confusion matrix is a two by two matrix which contains information about a classification system's actual and anticipated classifications [49].

The data obtained in the confusion matrix and used to calculate different performance metrics of the system are TP, FP, TN, and FN numbers when performing the performance validation test. The investigated algorithms confusion matrix result of this study found in appendix A.1.

Table 4.7: Confusion matrix

		Predicted values	
Actual values		A	B
	A	TP	FN
	B	FP	TN

Where;

TP: Predicted positive value matches the actual positive value

TN: Predicted negative value matches the actual negative value

FP: Predicted value is positive and the actual value was negative

FN: Predicted value is negative and the actual value was positive

Performance evaluation metrics can be calculated based on confusion matrix output and have been used in this study are stated and formulated as follows:

Accuracy: is the most often used and easily understood metric since it is a simple ratio of the number of examples successfully predicted to the total number of instances utilized in the observation; in other words, accuracy offers the percentage of correctly anticipated instances [50].

$$\mathrm{Accuracy} = \frac{TP + TN}{TP + FN + TN + FP} \tag{4.6}$$

Precision: is a metric that calculates the ratio of accurately positive predicted instances to the total number of positive examples predicted for each class [50].

$$\mathrm{Precision} = \frac{TP}{TP + FP} \tag{4.7}$$

Recall: is a statistic that offers for each class the ratio of true positive instances anticipated to the total of true positives and false negatives in the observation [50].

$$\mathrm{Recall} = \frac{TP}{TP + FN} \tag{4.8}$$

F1-score: is the weighted average of Precision and Recall, and the highest value is 1.0 when it is perfect, and the lowest possible value is 0.0, therefore a high-value F1 score indicates that we have minimal false positives and false negatives [50].

$$F1 - \text{score} = 2 * \frac{(\text{precision} * \text{recall})}{(\text{precision} + \text{recall})} \tag{4.9}$$

4.2.2 *Performance Evaluation Techniques*

The performance of classifiers can be evaluated with two types of techniques which are percentage split or holdout and cross-validation. When employing percentage split or hold out approaches, the data is divided into two portions, which are referred to as the training dataset and the validation or testing dataset. The model is built with a training dataset and validated or tested with a testing or validation data set. A larger portion of the data is used for model training, and the test metrics of the model are evaluated using a holdout dataset.

When the available dataset is too small for a train or test or train or valid split to generate an appropriate calculation of generalization error, the cross-validation technique is preferred to estimate the generalization error of a learning algorithm. In cross-validation, the original dataset is equally divided into k subparts or folds. For each iteration, one of the k-folds or groups is chosen as validation data, while the remaining (k-1) groups are chosen as training data. Each of these k folds is considered as a holdout dataset, and the model's performance is evaluated using the holdout fold. The overall performance of the model is the average of its performance on all k folds [25, 26]. In this classifier performance investigation research cross-validation with fold k=10 performance evaluation technique has been used, since it is suggested by many researchers.

5

RESULT AND DISCUSSION

This chapter provides a discussion based on different performance results generated from different perspectives. As stated in the previous chapter four performance parameter metrics which are accuracy, precision, recall, and f1-score have been used to measure the performance of the prediction algorithm and compare three machine learning algorithms against each other. Nowadays, building a more accurate prediction model or improvement is the main research area and necessary for telecom operators. Developed models result with each three selected algorithms presented in section 5.1 detail. Results of classifier performance change with including and excluding additional extract features comparison within each above-proposed algorithm described in section 5.2 and the performance comparison between classifiers provided in section 5.3.

5.1 MODEL BUILDING

To build a postpaid mobile service prediction model, this study analyzes the results of three machine-learning classifiers with some customer usage data and customer profiles. After the data preparation and experimentation setup was completed, we used three machine-learning techniques on the balanced data set to generate all the models. Investigated algorithms in this study are J48, Random forest, and multilayer perceptron applied on a multi-class balanced dataset which have twelve(12) independent selected features with 21978 instances prepared in subsection 4.1.3.7. Each of the algorithms performance results based on the experiment is discussed in the following subsection J48 decision tree in 5.1.1 results of RF in 5.1.2 and the third algorithm which is multilayer perceptron performance result describe in 5.1.3.

5.1.1 *J48 Decision Tree*

Based on the experiment result have been measured the accuracy of the J48 decision tree algorithm is 94.9% and The performance of precision, recall, and f1-score are 95.1%, 95.6%, and 95.4% to identify active service from voluntary and involuntary terminated services respectively. The performance of precision, recall, and f1-score are 94.7%, 94.4%, and 94.6% to recognize involuntary termination service from active service and involuntarily terminated service. The performance of precision, recall, and f1-score to identify voluntary termination service from active service and involuntarily terminated service are 94.9% 94.6%, and 94.8% respectively. The J48 prediction algorithm detail performance results based on accuracy, precision, recall, and f1-score metrics is shown in the table 5.1 and a detailed result of the J48 algorithm found in appendix A.2.

Table 5.1: J48 Decision tree performance result

Algorithm	**Accuracy**	**Precision**	**Recall**	**F1-score**	**Class label**
J48	94.9%	95.1%	95.6%	95.4%	Active service
		94.7%	94.4%	94.6%	Involuntary terminated
		94.9%	94.6%	94.8%	Voluntary terminated
Weighted Average		**94.9%**	**94.9%**	**94.9%**	

Results generated with the J48 algorithm is more interpretable compared to others and the disadvantage of J48 is that may happen overfitting and have an impact on tree visualization mean that on-result interpretation and can be addressed by pruning approach one of the pruning approach is change limited number of instance per leaf. This pruning approach may have an impact on accuracy. For example in this study, accuracy dropped from 94.9% to 78.6% when a limited number of instances per leaf change from 2 to 1200, the tradeoff pruning between the limited number of samples and accuracy, between the limited number of samples and number of leaves and tree size described in the figure 5.1 respectively.

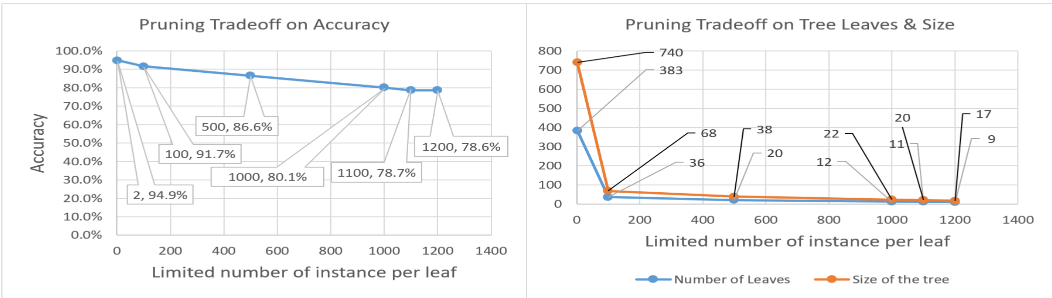

Figure 5.1: Pruning tradeoff

After using the pruning the visualization tree of J48 result for this study depicted as in the figure 5.2.

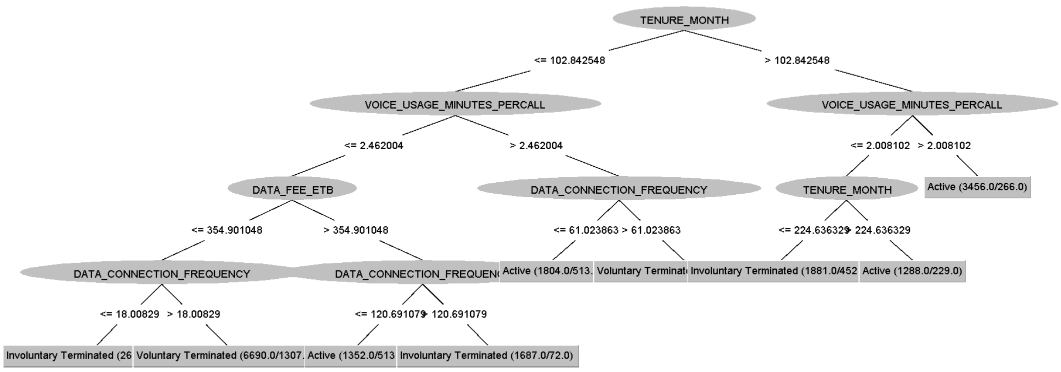

Figure 5.2: J48 pruned tree of postpaid mobile service termination prediction model

5.1.2 *Random Forest*

Random forest scored its better performance at 95.1% of accuracy and precision, recall, and f1-score of RF performance to identify active service from involuntary and voluntary terminated service are 95.0%, 95.8%, and 95.4% respectively. The performance of precision, recall, and f1-score are 95.1% 94.7%, and 94.9% to recognize involuntary termination services from active service and involuntarily terminated service respectively. Recorded performance of precision, recall, and f1-score to identify voluntary termination service from active service and involuntarily terminated service are 95.1% 94.8%, and 94.9% respectively. The accuracy, precision, recall, and f1-sore performance result of RF is depicted in table 5.2 and a detailed result of the RF algorithm found in appendix A.3.

Table 5.2: Random forest performance result

Algorithm	**Accuracy**	**Precision**	**Recall**	**F1-score**	**Class label**
RF	95.1%	95.0%	95.8%	95.4%	Active Service
		95.1%	94.7%	94.9%	Involuntary Terminated
		95.1%	94.8%	94.9%	Voluntary Terminated
Weighted Average		**95.1%**	**95.1%**	**95.1%**	

5.1.3 *Multilayer Perceptron*

Multilayer perceptron classifier can be performed with 93.3% accuracy. Precision, recall, and f1-score performance of MLP to identify active service from voluntary and involuntary terminated service are 93.3%, 94.9% and 94.0% sequentially. The performance of precision, recall, and f1-score are 93.8%, 92.1%, and 92.9% to recognize involuntary termination services from active service and voluntary terminated service respectively. The performance of precision, recall, and f1-score to identify voluntary termination service from active service and involuntarily terminated service are 93.3%, 93.0%, and 93.1% respectively. Detail results were recorded on multilayer perceptron machine learning algorithms between class labels are shown in the table 5.3 and a detailed result of the MLP algorithm found in appendix A.4.

Table 5.3: Multilayer perceptron performance result

Algorithm	Accuracy	Precision	Recall	F1-score	Class label
ANN (MLP)	93.3%	93.0%	94.9%	94.0%	Active service
		93.8%	92.1%	92.9%	Involuntary terminated
		93.3%	93.0%	93.1%	Voluntary terminated
Weighted Average		**93.3%**	**93.3%**	**93.3%**	

5.2 COMPARISON OF CLASSIFIERS WITH INCLUDING AND EXCLUDING ONE ATTRIBUTE

Extract features, features selection techniques, data size, validation method, and other factors may affect the performance of algorithms. One feature, which is called bill cycle ID not have been used to build a prediction model previously investigated and included in this prediction modeling research. As a result, we observed a performance change on classifiers in the case including and excluding it. Measured results on selected above algorithms regarding bill cycle ID feature including and excluding depicted on table 5.4 and a detailed result found in appendix A.5.

Table 5.4: Proposed algorithms performance result with including and excluding bill cycle ID attribute

	Before excluding bill cycle ID attribute				**After excluding bill cycle ID attribute**				
Algorithm	Accuracy	Precision	Recall	F-Measure	Accuracy	Precision	Recall	F-Measure	Class
J48	94.9%	95.1%	95.6%	95.4%	90.8%	93.2%	94.9%	94.1%	Active Service
		94.7%	94.4%	94.6%		91.2%	87.1%	89.1%	Involuntary Terminated
		94.9%	94.6%	94.8%		88.1%	90.4%	89.3%	Voluntary Terminated
Weighted Average		**94.9%**	**94.9%**	**94.9%**		**90.8%**	**90.8%**	**90.8%**	
RF	95.1%	95.0%	95.8%	95.4%	92.5%	94.2%	95.5%	94.9%	Active Service
		95.1%	94.7%	94.9%		93.1%	89.3%	91.2%	Involuntary Terminated
		95.1%	94.8%	94.9%		90.2%	92.7%	91.4%	Voluntary Terminated
Weighted Average		**95.1%**	**95.1%**	**95.1%**		**92.5%**	**92.5%**	**92.5%**	
ANN (MLP)	93.3%	93.0%	94.9%	94.0%	89.5%	92.1%	94.4%	93.2%	Active Service
		93.8%	92.1%	92.9%		91.8%	82.8%	87.1%	Involuntary Terminated
		93.3%	93.0%	93.1%		85.0%	91.2%	88.0%	Voluntary Terminated
Weighted Average		**93.3%**	**93.3%**	**93.3%**		**89.6%**	**89.5%**	**89.4%**	

As we can see from the table 5.4 the accuracy performance difference measured on J48, RF and MLP are 4.1%, 2.6%, and 3.8% respectively. The recorded deviation of accuracy, precision, recall, and f-measure is the same, which is 4.1% on J48 and RF recorded deviations on the accuracy, precision, recall, and f-measure metrics are the same in value is 2.6%. Recorded deviation of accuracy, precision, recall, and F-Measure are 3.8%, 3.7%, 3.8%, and 3.9% respectively on ANN (MLP). The highest difference was recorded on J48 whereas the minimum deviation point scored on RF and the recorded deviation on J48 and MLP is almost the same.

The figure 5.3 is prepared based on the accuracy, weighted average precision, recall, f-measure performance and presented for simple visualization to show the difference between including and excluding bill cycle ID attribute when developing postpaid mobile service termination prediction model.

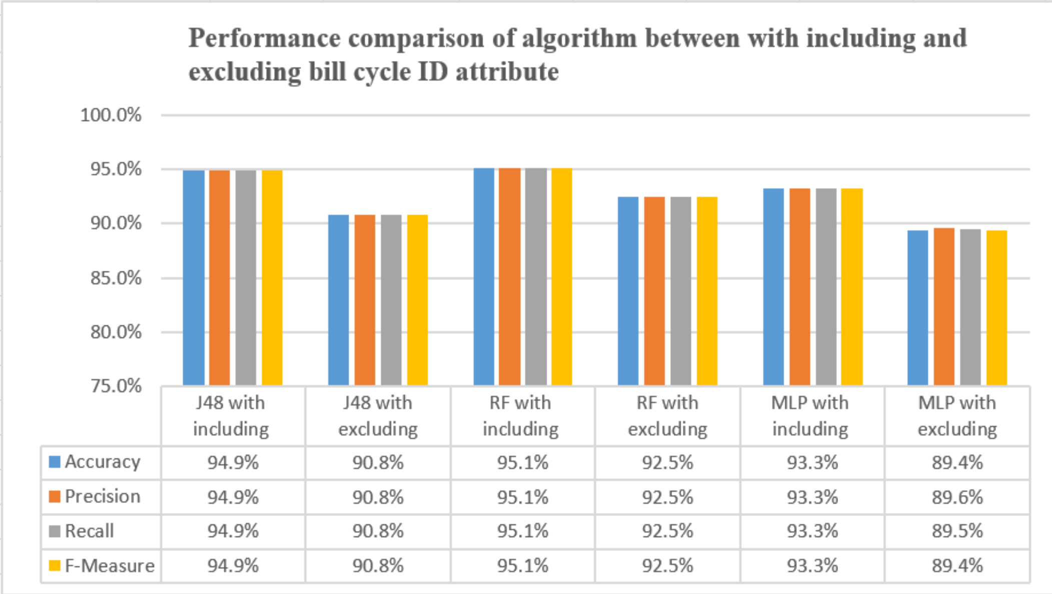

Figure 5.3: Comparison of algorithm between with including and excluding bill cycle ID attribute

Based on the result have been observed from the above bar chart we can conclude that billing cycle ID has a contribution to building a better post-paid mobile service termination classification model.

5.3 COMPARISON OF CLASSIFIERS

The comparison of the proposed three algorithms is discussed based on their accuracy and efficiency which have been measured when building and testing models with each algorithm. The efficiency of algorithms will be measured based on the time taken to build prediction models by using the given dataset. Time taken for proposed algorithms in this study to build a prediction model using the same multi-class dataset are stated in the table 5.5.

Table 5.5: Classifiers model building time performance result

Algorithm	Building time in second
J48	0.94
RF	10.89
MLP	34.37

As we have stated in the table 5.5 the time taken to a model building with J48, RF, and ANN (MLP) are 0.94, 10.89, and 37.34 seconds respectively. so that the J48 algorithm is more time-efficient compared to RF and MLP while MLP is time inefficient compared to J48 and RF. The model building time efficiency of RF is faster when compared to MLP however it is slower when relative to J48.

The effectiveness of algorithms can be measured based on terms of overall correctly predicted instances rate out of all instances or accuracy, error rate, precision, recall, and f-measure. Accuracy, weighted average precision, weighted average recall, weighted average f-measure of selected prediction algorithms are described as follows in figure 5.4.

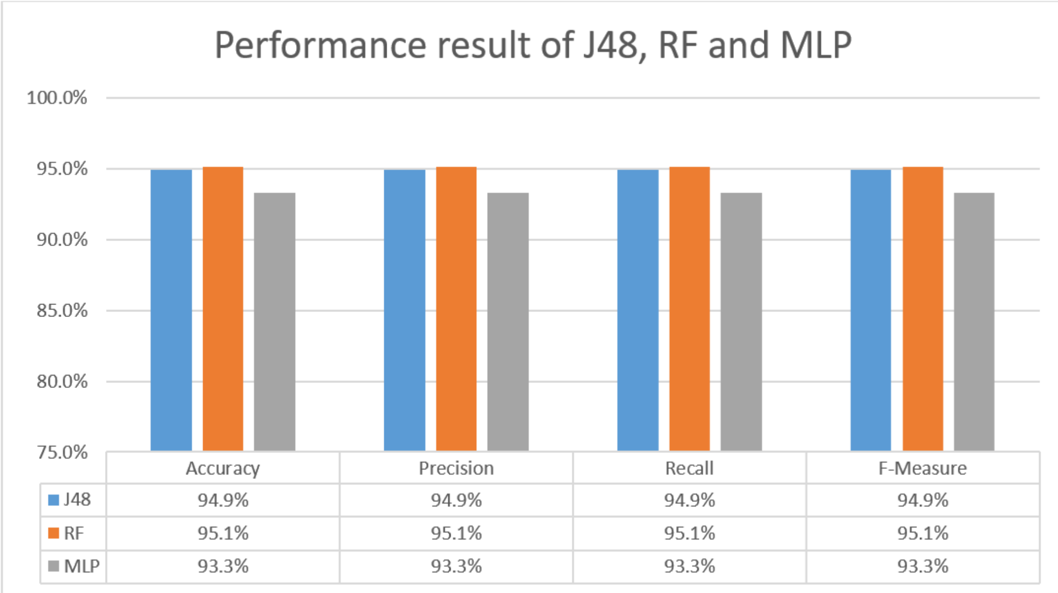

Figure 5.4: Performance result of J48, RF and MLP

As we can see on the above Fig 5.4 accuracy of J48, RF and MLP are 94.9%, 95.1%, and 93.3% respectively. The performance of precision recorded on J48, RF and MLP are 94.9%, 95.1%, and 93.3% respectively. Performance of recall scored on J48, RF and MLP are 94.9%, 95.1%, and 93.3% respectively. The f1-score measured on J48, RF, and MLP are 94.9%,95.1%, and 93.3% respectively.

From the figure 5.4 and table 5.5, we can observe that results of all performance metrics recorded by J48 and RF are better and almost the same except model building time whereas minimum performance result generated with MLP.

6

CONCLUSION AND RECOMMENDATION

6.1 CONCLUSION

Service termination is an issue for telecom operators, including monopolistic markets such as ethio telecom in Ethiopia, and have a greater impact on revenue loss as well as the operator's reputation and retention rate. From the customers' standpoint of view, service terminations might be voluntary or involuntary.

Voluntary service termination and involuntary service termination may have different causes like service quality for voluntary while fraud and unpaid bill for involuntary. Voluntary service terminations request initiated from customers whereas involuntary from telecom operators and conduct involuntary service termination without knowledge about terminated customers reaction may consider as advertise service termination as one business strategy.

A mechanism, which is used to control service termination is necessary to reduce revenue loss happened in case of service termination and its subsequent. The main purpose of service termination or customer churn prediction model is to forecast customers will be terminated their service or churn to another service provider nearly and helps to enable retain them by conducting appropriate retention strategy plans and retention activities proactively in result reducing revenue losses. The main goal of this research analyzed three selected algorithms which are J48, Random Forest, and ANN(MLP) using a prepared multi-class balanced dataset from usage CDR and customer profile data of postpaid mobile service subscribers. Cross-validation techniques (k=10) with WEKA tool have been used to evaluate the performance of the classifiers. We observed that from our result the accuracy of RF, J48, and ANN(MLP) are 95.1%, 94.9%, and 93.3% respectively.

Random Forest is the most accurate algorithm even if without hyperparameter adjustment it gives maximum accurate values. The main drawback of the RF algorithm taking a significant time to build its model compared to J48 but it is faster relative to MLP. J48 is moderate and almost the same as the RF classification algorithm and the disadvantage of J48 is overfitting maybe happen when used to build a model but can build a more interpretable model quickly. ANN(MLP) is the least accurate classification algorithm as compared to J48 and Random Forest. In multilayer perceptron computations are difficult and time-consuming when building a classification model, and the functions of the model depend on the training data quality while can applied to complex non-linear problems.

In this research, best features have been selected using correlation and gain ratio feature selection evaluation techniques. To identify the best features when using two types of feature selection evaluation techniques we used two approaches to remove or select the same attributes from the two-feature selection evaluation technique after deciding the ranking threshold. The first approach removes attributes not selected on the top ten (10) in both techniques (Correlation and Gain ratio) and the second approach selects attributes that have been selected in common on the top ten (10) in both techniques. Evaluate the approach's performance and observe that using the first approach, which is removed attributes not selected on the top ten (10) in both techniques (Correlation and Gain ratio) is better.

6.2 RECOMMENDATION

Predicting postpaid mobile service termination is not a one-time task rather a continuum. Thus, future works proposed to proceed to start from this research include:

- Clustering identified terminated services: This is important to know the root cause of service termination and prepare an appropriate action plan to retain.
- Employed by using big data and additional features may give better classification models.
- Comparing classification models with customer experience: This helps to identify the real problems of termination behind the figures in prediction models especially for voluntary service termination.

BIBLIOGRAPHY

[1] Yi-Bing Lin, Ming-Feng Chang, and Herman Chung-Hwa Rao. "Mobile prepaid phone services." In: *IEEE Personal Communications* 7.3 (2000), pp. 6–14.

[2] Adnan Ghayas. *Prepaid Vs Postpaid -What's The Difference?* Last accessed on 2021-10-15. 2020. URL: https://commsbrief.com/?s=Prepaid+Vs+Postpaid+-What%E2%80%99s+The+Difference%3F.

[3] Yongjun Liao, Wei Du, Márton Karsai, Carlos Sarraute, Martin Minnoni, and Eric Fleury. "Prepaid or Postpaid? That Is the Question: Novel Methods of Subscription Type Prediction in Mobile Phone Services." In: *Social Network Based Big Data Analysis and Applications*. Springer, 2018, pp. 165–181.

[4] Kiflemariam Mehari. "Predicting Termination of Mobile Subscribers Using Classification Algorithms." MSC Thesis. Addis Ababa University Addis Ababa, Ethiopia, 2019.

[5] Amin Nazifi, Dahlia El-Manstrly, Angela Tregear, and Kristina Auxtova. "The impact of termination severity on customers' emotional, attitudinal and behavioral reactions." In: *Journal of Service Theory and Practice* (2020).

[6] ethiotelecom. *Mobile Service Subscribers*. Last accessed on 2021-11-1. 2021. URL: https://www.ethiotelecom.et/.

[7] Irfan Ullah, Basit Raza, Ahmad Kamran Malik, Muhammad Imran, Saif Ul Islam, and Sung Won Kim. "A churn prediction model using random forest: analysis of machine learning techniques for churn pre-

diction and factor identification in telecom sector." In: *IEEE Access* 7 (2019), pp. 60134–60149.

[8] Andreea Dumitrache, Monica Mihaela Matei Maer, Stelian Stancu, and Oana Popescu. "The Factors that Influence Postpaid Subscribers to Make Churn in a Romanian Telecommunication Company." In: *Proceedings of the International Conference on Applied Statistics*. Vol. 1. 1, pp. 240–247.

[9] Vishal Mahajan, Richa Misra, and Renuka Mahajan. "Review of data mining techniques for churn prediction in telecom." In: *Journal of Information and Organizational Sciences* 39.2 (2015), pp. 183–197.

[10] Ahsan Rehman and Abbas Raza Ali. "Customer churn prediction, segmentation and fraud detection in telecommunication industry." In: *Proceedings of the 4th ASE International Conference on Big Data, Harvard University*. 2014.

[11] C Arun Kumar, MP Sooraj, and S Ramakrishnan. "A comparative performance evaluation of supervised feature selection algorithms on microarray datasets." In: *Procedia computer science* 115 (2017), pp. 209–217.

[12] Utsav Mishra. *Binary and Multiclass Classification in Machine Learning*. Last accessed on 2021-10-28. 2021. URL: https://www.analyticssteps.com/blogs/binary-and-multiclass-classification-machine-learning.

[13] Dejan Radosavljevik and Peter van der Putten. "Preventing churn in telecommunications: The forgotten network." In: *International Symposium on Intelligent Data Analysis*. Springer. 2013, pp. 357–368.

[14] M Jaeyalakshmi, S Gnanavel, K S Guhapriya, S Harshini Phriyaa, K Kavya Sree. *Prediction of Customer Churn on e-Retailing*. Last accessed on 2021-10-20. 2020. URL: https://www.ijrte.org/wp-content/uploads/papers/v8i6/F9550038620.pdf.

[15] Emilía Huong Xuan Nguyen. "Customer churn prediction for the Icelandic mobile telephony market." MSC Thesis. 2011.

[16] Clement Kirui, Li Hong, Wilson Cheruiyot, and Hillary Kirui. "Predicting customer churn in mobile telephony industry using probabilistic classifiers in data mining." In: *International Journal of Computer Science Issues (IJCSI)* 10.2 Part 1 (2013), p. 165.

[17] Yi-Shun Wang, Hsin-Hui Lin, and Pin Luarn. "Predicting consumer intention to use mobile service." In: *Information systems journal* 16.2 (2006), pp. 157–179.

[18] Kristina Heinonen and Minna Pura. "Classifying mobile services." In: *Helsinki Mobility Roundtable* (2006), pp. 1–2.

[19] Omae Malack Oteri, Langat Philip Kibet, and EN Ndung'u. "Mobile subscription, penetration and coverage trends in Kenya's Telecommunication Sector." In: *International Journal of Advanced Research in Artificial Intelligence* 4.1 (2015), pp. 1–7.

[20] Keng Siau and Zixing Shen. "Mobile communications and mobile services." In: *International Journal of Mobile Communications* 1.1-2 (2003), pp. 3–14.

[21] S Sriram, Pradeep K Chintagunta, and Puneet Manchanda. "Service quality variability and termination behavior." In: *Management Science* 61.11 (2015), pp. 2739–2759.

[22] Jainam D Shah, Fenil D Shah, and Mrugendra Rahevar. "Customer Churn Prediction Analysis." In: *International Journal of Computer Applications* 975 (), p. 8887.

[23] V Umayaparvathi and K Iyakutti. "A survey on customer churn prediction in telecom industry: Datasets, methods and metrics." In: *International Research Journal of Engineering and Technology (IRJET)* 3.04 (2016).

[24] Sahar F Sabbeh. "Machine-learning techniques for customer retention: A comparative study." In: *International Journal of Advanced Computer Science and Applications* 9.2 (2018).

[25] Mohssen Mohammed, Muhammad Badruddin Khan, and Eihab Bashier Mohammed Bashier. *Machine learning: algorithms and applications*. Crc Press, 2016.

[26] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. "Machine learning basics." In: *Deep learning* 1.7 (2016), pp. 98–164.

[27] MiguelA Carreira-Perpinán. "CSE176 Introduction to Machine Learning—Lecture notes." In: (2016).

[28] Wei-Lun Chao. "Machine learning tutorial." In: *Digital Image and Signal Processing* (2011).

[29] Jason Brownlee. *Types of Classification Tasks in Machine Learning*. Last accessed on 2021-11-5. 2020. URL: https://https://machinelearningmastery.com/types-of-classification-in-machine-learning/.

[30] Rachel Wolff. *Multi-Label Classification: Overview & How to Build A Model*. Last accessed on 2021-12-29. 2020. URL: https://monkeylearn.com/blog/multi-label-classification/.

[31] N Saravana and Dr V Gayathri. "Performance and classification evaluation of J48 algorithm and Kendall's based J48 algorithm (KNJ48)." In: *Int. J. Comput. Trends Technol.(IJCTT)–Volume* 59 (2018).

[32] J.GUPTA, J.PATRICK, S.PHOON. *CLINICAL SAFETY INCIDENT TAXONOMY PERFORMANCE ON J48 AND RANDOM FOREST*. Last accessed on 2021-10-29. 2019. URL: https://www.hisa.org.au/slides/hic19/tue/JaiGupta.pdf.

[33] Sultan Aljahdali, Syed Naimatullah Hussain, et al. "Comparative prediction performance with support vector machine and random forest classification techniques." In: *International journal of computer applications* 69.11 (2013).

[34] Niklas Donges. *A Complete Guide to the Random Forest Algorithm*. Last accessed on 2021-10-25. 2021. URL: https://builtin.com/data-science/random-forest-algorithm#procon.

[35] Berke Akkaya and Nurdan Çolakoğlu. "Comparison of Multi-class Classification Algorithms on Early Diagnosis of Heart Diseases." In: ().

[36] FY Osisanwo, JET Akinsola, O Awodele, JO Hinmikaiye, O Olakanmi, and J Akinjobi. "Supervised machine learning algorithms: classification and comparison." In: *International Journal of Computer Trends and Technology (IJCTT)* 48.3 (2017), pp. 128–138.

[37] Sotiris B Kotsiantis, Dimitris Kanellopoulos, and Panagiotis E Pintelas. "Data preprocessing for supervised leaning." In: *International journal of computer science* 1.2 (2006), pp. 111–117.

[38] Vaishali Chandrakant Wangikar and Ratnadeep R Deshmukh. "Data Cleaning: Current Approaches and Issues." In: ().

[39] Maurizio Lenzerini. "Data integration: A theoretical perspective." In: *Proceedings of the twenty-first ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*. 2002, pp. 233–246.

[40] Jürgen Bernard, Marco Hutter, Matthias Zeppelzauer, Dieter Fellner, and Michael Sedlmair. "Comparing visual-interactive labeling with active learning: An experimental study." In: *IEEE transactions on visualization and computer graphics* 24.1 (2017), pp. 298–308.

[41] Teodor Fredriksson, David Issa Mattos, Jan Bosch, and Helena Holmström Olsson. "Data labeling: an empirical investigation into industrial challenges and mitigation strategies." In: *International Conference on Product-Focused Software Process Improvement*. Springer. 2020, pp. 202–216.

[42] Harshada C Mandhare and SR Idate. "A comparative study of cluster based outlier detection, distance based outlier detection and density based outlier detection techniques." In: *2017 International Conference on Intelligent Computing and Control Systems (ICICCS)*. IEEE. 2017, pp. 931–935.

[43] SS Sreevidya et al. "A survey on outlier detection methods." In: *IJCSIT) International Journal of Computer Science and Information Technologies* 5.6 (2014).

[44] Abebe Kelemework. "Comparison of Supervised Machine Learning Algorithms on Detection of Signalling DoS attack to the 3G (UMTS) mobile network-in the case of ethio telecom." MSC Thesis. Addis Ababa University, Addis Ababa, 2020.

[45] Amal S Ghanem, Svetha Venkatesh, and Geoff West. "Multi-class pattern classification in imbalanced data." In: *2010 20th International Conference on Pattern Recognition*. IEEE. 2010, pp. 2881–2884.

[46] Mahendra Sahare and Hitesh Gupta. "A review of multi-class classification for imbalanced data." In: *International Journal of Advanced Computer Research* 2.3 (2012), p. 160.

[47] M Vijayakumar and E Prabhakar. "A hybrid combined under-over sampling method for class imbalanced datasets." In: *International Journal of Research and Advanced Development (IJRAD)* 2.05 (2018), pp. 27–33.

[48] Astha Agrawal, Herna L Viktor, and Eric Paquet. "SCUT: Multi-class imbalanced data classification using SMOTE and cluster-based undersampling." In: *2015 7Th international joint conference on knowledge discovery, knowledge engineering and knowledge management (IC3k)*. Vol. 1. IEEE. 2015, pp. 226–234.

[49] Tina R Patil. "MSSS Performance analysis of naive bayes and J48 classification algorithm for data classification. Intl." In: *Journal of Computer Science and Applications* 6.2 (2013).

[50] Márcio Guia, Rodrigo Rocha Silva, and Jorge Bernardino. "Comparison of Naïve Bayes, Support Vector Machine, Decision Trees and Random Forest on Sentiment Analysis." In: *KDIR*. 2019, pp. 525–531.

A

APPENDIX

A.1 Confusion matrix result of proposed algorithms

Table A.1: Confusion matrix result of proposed algorithms

Algorithm	Confusion Matrix			classified as
J48	a	b	c	a = Involuntary terminated
	6918	181	227	b = Active service
	174	7006	146	c = Voluntary terminated
	215	180	6931	
RF	a	b	c	a = Involuntary terminated
	6937	181	208	b = Active service
	159	7017	150	c = Voluntary terminated
	195	187	6944	
MLP	a	b	c	a = Involuntary terminated
	6745	276	305	b = Active service
	183	6955	188	c = Voluntary terminated
	263	248	6815	

A.2 DETAIL PERFORMANCE RESULT OF PROPOSED ALGORITHMS

Table A.2: Detail performance result of J48 algorithm

	Detailed accuracy by class label								
Algorithm	**TP**	**FP**	**Precision**	**Recall**	**F-Measure**	**MCC**	**ROC Area**	**PRC Area**	**Classa Label**
J48	0.944	0.027	0.947	0.944	0.946	0.918	0.969	0.938	Involuntary terminated
	0.956	0.025	0.951	0.956	0.954	0.93	0.975	0.936	Active service
	0.946	0.025	0.949	0.946	0.948	0.921	0.972	0.941	Voluntary terminated
weighted average	0.949	0.026	0.949	0.949	0.949	0.923	0.972	0.938	

Algorithm	**Performance Metrics**	**Performance result**
J48	Kappa statistic	0.9234
	Mean absolute error	0.0412
	Root mean squared error	0.175
	Relative absolute error	9.27%
	Root relative squared error	37.13%

Algorithm	**Number of Leaves**	**Size of the tree**	**Time taken to build model**
J48	383	740	0.94 seconds

Table A.3: Detail performance result of RF algorithm

	Detailed accuracy by class label								
Algorithm	**TP**	**FP**	**Precision**	**Recall**	**F-Measure**	**MCC**	**ROC Area**	**PRC Area**	**Classa Label**
RF	0.947	0.024	0.951	0.947	0.949	0.924	0.994	0.989	Involuntary terminated
	0.958	0.025	0.95	0.958	0.954	0.931	0.996	0.992	Active service
	0.948	0.024	0.951	0.948	0.949	0.924	0.995	0.99	Voluntary terminated
weighted average	0.951	0.025	0.951	0.951	0.951	0.926	0.995	0.99	

Algorithm	**Performance Metrics**	**Performance result**
RF	Kappa statistic	0.9263
	Mean absolute error	0.0593
	Root mean squared error	0.1573
	Relative absolute error	13.34%
	Root relative squared error	33.37%
	Time taken to build model	10.89 seconds

Table A.4: Detail performance result of MLP algorithm

	Detailed accuracy by class label								
Algorithm	**TP**	**FP**	**Precision**	**Recall**	**F-Measure**	**MCC**	**ROC Area**	**PRC Area**	**Classa Label**
MLP	0.921	0.03	0.938	0.921	0.929	0.894	0.973	0.957	Involuntary terminated
	0.949	0.036	0.93	0.949	0.94	0.909	0.987	0.974	Active service
	0.93	0.034	0.933	0.93	0.931	0.897	0.982	0.966	Voluntary terminated
weighted average	0.933	0.033	0.933	0.933	0.933	0.9	0.981	0.966	

Algorithm	Performance Metrics	Performance result
MLP	Kappa statistic	0.9002
	Mean absolute error	0.0602
	Root mean squared error	0.1934
	Relative absolute error	13.53%
	Root relative squared error	41.03%
	Time taken to build model	34.37 seconds

Table A.5: Detail comparison result of proposed algorithms with including and excluding bill cycle ID attribute

	Before excluding bill cycle ID attribute								After excluding bill cycle ID attribute								
Algorithm	**TP**	**FP**	**Precision**	**Recall**	**F-Measure**	**MCC**	**ROC Area**	**PRC Area**	**TP**	**FP**	**Precision**	**Recall**	**F-Measure**	**MCC**	**ROC Area**	**PRC Area**	**Classa Label**
J48	0.944	0.027	0.947	0.944	0.946	0.918	0.969	0.938	0.871	0.042	0.912	0.871	0.891	0.839	0.944	0.9	Involuntary terminated
	0.956	0.025	0.951	0.956	0.954	0.93	0.975	0.936	0.949	0.034	0.932	0.949	0.941	0.911	0.976	0.937	Active service
	0.946	0.025	0.949	0.946	0.948	0.921	0.972	0.941	0.904	0.061	0.881	0.904	0.893	0.838	0.948	0.876	Voluntary terminated
weighted average	0.949	0.026	0.949	0.949	0.949	0.923	0.972	0.938	0.908	0.046	0.908	0.908	0.908	0.863	0.956	0.904	
RF	0.947	0.024	0.951	0.947	0.949	0.924	0.994	0.989	0.893	0.033	0.931	0.893	0.912	0.869	0.98	0.97	Involuntary terminated
	0.958	0.025	0.95	0.958	0.954	0.931	0.996	0.992	0.955	0.029	0.942	0.955	0.949	0.923	0.994	0.987	Active service
	0.948	0.024	0.951	0.948	0.949	0.924	0.995	0.99	0.927	0.05	0.902	0.927	0.914	0.871	0.985	0.966	Voluntary terminated
weighted average	0.951	0.025	0.951	0.951	0.951	0.926	0.995	0.99	0.925	0.038	0.925	0.925	0.925	0.888	0.986	0.974	
MLP	0.921	0.03	0.938	0.921	0.929	0.894	0.973	0.957	0.828	0.037	0.918	0.828	0.871	0.813	0.943	0.918	Involuntary terminated
	0.949	0.036	0.93	0.949	0.94	0.909	0.987	0.974	0.944	0.04	0.921	0.944	0.932	0.898	0.986	0.971	Active service
	0.93	0.034	0.933	0.93	0.931	0.897	0.982	0.966	0.912	0.08	0.85	0.912	0.88	0.818	0.957	0.891	Voluntary terminated
weighted average	0.933	0.033	0.933	0.933	0.933	0.9	0.981	0.966	0.895	0.053	0.896	0.895	0.894	0.843	0.962	0.926	