Telecom Customer Churn Prediction Model for Ethiopian Telecommunication

Yerosan Birhanu Zewde

A Thesis submitted to Department of Computer Science and Engineering

School of Electrical Engineering and Computing

Presented in Partial Fulfillment of the Requirement for the Degree of Master's in Computer Science and Engineering

Office of Graduate Studies

Adama Science and Technology University

June 2023,

Adama, Ethiopia

Telecom Customer Churn Prediction Model for Ethiopian Telecommunication

Yerosan Birhanu Zewde

Advisor: Teklu Urgessa (PhD)

A Thesis submitted to Department of Computer Science and Engineering

School of Electrical Engineering and Computing

Presented in Partial Fulfillment of the Requirement for the Degree of Master's in Computer Science and Engineering

Office of Graduate Studies

Adama Science and Technology University

June 2023,

Adama, Ethiopia

DECLARATION

I hereby declare that this Master Thesis entitled ‒**Telecom Customer Churn Prediction Model for Ethiopian Telecommunication**‖ is my original work. That is, it has not been submitted for the award of any academic degree, diploma, or certificate in any other university. All sources of materials that are used for this thesis have been duly acknowledged through citations.

Yerosan Birhanu

Name of the student	Signature	Date

RECOMMENDATION

I, the advisor of this thesis, hereby certify that I have read the revised version of the thesis entitled "**Telecom Customer Churn Prediction Model for Ethiopian Telecommunication''** prepared under my guidance by Yerosan Birhanu submitted in partial fulfillment of the requirements for the degree of Masters of Science in Computer Science and Engineering. Therefore, I recommend the submission of a revised version of the thesis to the department following the applicable procedures.

_____ _____ _____

Major Advisor Signature Date

APPROVAL SHEET

I the advisor of the thesis entitled "**Telecom Customer Churn Prediction Model for Ethiopian Telecommunication''** and developed by "**Yerosan Birhanu**" hereby certify that the recommendation and suggestion made by the board of examiners are appropriately incorporated into the final version of the thesis.

Dr. Teklu Urgessa	_____	_____
Major Advisor/Supervisor	Signature	Date

We, the undersigned, members of the Board of Examiners of the thesis by Yerosan Birhanu have read and evaluated the thesis entitled "**Telecom Customer Churn Prediction Model for Ethiopian Telecommunication''** and examined the candidate during the open defense. This is, therefore, to certify that the thesis is accepted for partial fulfillment of the requirement of the degree of Masters of Science in Computer Science and Engineering.

_____	_____	_____
Chairperson	Signature	Date
_____	_____	_____
Reviewer 1	Signature	Date
_____	_____	_____
Reviewer 2	Signature	Date

Final approval and acceptance of the thesis are contingent upon submission of its final copy to the Office of Postgraduate Studies (OPGS) through the Department Graduate Council (DGC) and School Graduate Committee (SGC).

_____	_____	_____
Department Head	Signature	Date
_____	_____	_____
School Dean	Signature	Date
_____	_____	_____
Office of Postgraduate Studies, Dean	Signature	Date

ACKNOWLEDGEMENT

I would like to thank **God** for being with me throughout my life. I am grateful for His love, grace, and mercy that have sustained me in every situation.

I am particularly thankful to my advisor **Dr. Teklu Urgessa** for his invaluable guidance, and support and for helping me complete the research on time. He has encouraged me with his patience and kindness.

I would like to express my sincere appreciation to my **friends** who supported me throughout my research project. They have provided me with constructive feedback, useful suggestions, and honest opinions. Finally, I would like to thank my **family** for their unwavering love, support, and encouragement.

Table of Contents

Contents Page

LIST OF FIGURES

LIST OF TABLES

LIST OF EQUATIONS

LISTS OF ABBREVIATIONS AND ACRONYMS

AI	Artificial Intelligence
ANN	Artificial Neural Network
AUC	Area Under Curve
CI	Computational Intelligence
CRM	Customer Relationship Management
DM	Data Mining
DT	Decision Tree
ECA	Ethiopian Communication Authority
ECA	Ethiopian Communication Authority
ETB	Ethiopian Birr
FN	False Negative
FP	False Positive
GBM	Gradient Boosting Machine
GSM	Global System for Mobile Communications
GTP	Growth and Transformation Plan
IDE	Integrated Development Environment
KNN	K-Nearest Neighbors
LR	Logistic Regression
MB	Mega Byte
ML	Machine Learning
ML	Machine Learning
MLE	Maximum Likelihood Estimation
PDF	Portable Document Format
RF	Random Forest
RoI	Return on Investment
RPS	Risk Profiling Score
SME	Small & Medium Enterprises
SMS	Short Message Service
SNA	Social Network Analysis

SOHO	Small Office-Home Office
SOM	Self-Organizing Map
SRM	Structural Risk Minimization principle
SVM	Support Vector Machine
TN	True Negative
TP	True Positive
XGBOOST	Extreme Gradient Boosting

ABSTRACT

Machine learning is a valuable tool for industries that deal with large amounts of data and want to enhance their operations. Ethio Telecom is one of these industries, which collects, processes, and stores huge volumes of records regularly. A major challenge for this industry is customer churn, which occurs when customers switch to another service provider and cause a loss of revenue. In industries that are based on subscriptions and face high competition and dynamism, customer churn is a serious problem. The telecommunication industry uses customer relationship management (CRM) databases to store a lot of data that contains hidden knowledge and patterns that can help them understand their customers better and make smarter decisions. However, because of the large amount of data that accumulates over time, CRM needs to apply business intelligence that can examine the causes and behaviors of churn customers from the existing data. The goal of this research is to create a machine-learning model that can classify customer data as churner or non-churner based on the customer's past data. The dataset for the experiment is obtained from Ethiopian telecommunication. The proposed model used random forest, support vector machine, logistic regression, and K-nearest neighbors' algorithms toclassify customer data. A confusion matrix was used to measure the accuracy, precision, recall, and f1 score and assess the performance of the model. The research results showed that the RF classifier achieved an accuracy of 96.6%, SVM achieved an accuracy of 94.55%, the KNN classifier achieved an accuracy of 92.88% and LR achieved an accuracy of 87.25%. Based performance comparison of these classification algorithms, the best classifier for the customer churn classification problem is random forest.

***Keywords**: Customer churn, machine learning, Telecom Customer Churn, Ethiopian Telecommunication*

CHAPTER ONE

1 INTRODUCTION

1.1 Background of the Study

One of the most valuable resources in the dynamic, cutthroat, and subscription-based company and service industries is the customer (Nur Amir Sjarif et al., 2019). Businesses are putting a lot of effort into thriving in cutthroat markets by employing a range of strategies. The major strategies that have been put up in the literature as ways to increase revenue are increasing the number of new customers you have, increasing the number of products or services you sell to current customers, and increasing the length of time that current customers stay with you (Ullah I. R., 2019). The third option, which shows that maintaining an existing customer costs much less than acquiring a new one, is the most profitable when comparing different strategies taking into account each one's Return on Investment (Ahmad, 2019).

A key factor in telecommunication companies' ability to sustain stable revenue is their customer base (Idris, 2012). Thus, keeping clients is given great consideration. Users of voice and data services in the telecom industries have the flexibility to choose a service provider from a different business and to freely transfer their rights to a different service provider they believe to be superior (Jain H. K., 2021). Because switching between providers is easy and convenient at all times, people are continually looking for better services at lower prices (Ullah I. R., 2019). Therefore, businesses should look into several methods that can forecast churn rates and aid in creating efficient client retention strategies.

‒Churn represents the problem of losing a customer when a customer switches from one service provider to another which leads to serious profit loss‖ (Amin A. A.-O., 2020). In the field of mobile communications, churn is the loss of customers who change service providers within a specific time frame (Gaur, 2018). This occurs when a significant number of customers are dissatisfied with any telecom provider's services (Ullah I. R., 2019).

In businesses with high levels of turnover, such as the telecommunications sector, customer relationship management (CRM) databases are frequently maintained. These databases hold a lot of undiscovered data as well as certain patterns that can be exploited to swiftly obtain consumer

information for business decision-making (Amin A. A., 2016). However, it can be difficult to find information in such comprehensive CRM databases because they frequently hold thousands or millions of records about clients. As a result, many industries must unavoidably rely on client attrition prediction models to compete in the market (Tamuka, 2020). As a result, the customer relationship management department of a telecom business looks for a reliable churn prediction model to identify potential churners (Idris, 2012).

The main contribution of this paper is developing an efficient and effective churn prediction model for Ethio Telecom using machine learning methods to determine whether a customer churns or not depending on historical data of the customer that has provided by Ethio Telecom.

The Churn Prediction model assists in analyzing past data that is accessible within the business to identify a list of customers who are likely to leave (Shirazi, 2019). This will enable telecom providers to prioritize investing in consumers who have a high likelihood of leaving rather than making an offer to all customers. Building a churn prediction model will make it easier to retain customers, which will help telecommunications firms thrive in the current market that is becoming more and more competitive (García, 2017). Therefore, a company can give plans that are more effective for the customer to stop them from churning.

Using a variety of approaches, several churn prediction models have been developed over several years (Sana, 2022). For churn prediction, the use of machine learning has grown in popularity. With the use of strong algorithms, machine learning can learn from previous observations to produce precise predictions. At the same time, data mining has advanced significantly (Jain, Khunteta, & Srivastava, 2020). The collection and processing of data now include a wide variety of innovative approaches and techniques. It is possible to describe data mining as the process of obtaining important information from data.

With a given training data set, classification is one of the supervised machine learning models that may be used to predict future events (Osisanwo, 2017). Decision trees, Bayesian classification, rule-based methods, memory-based learning, and support vector machines are a few examples of learning techniques that are used in classification algorithms to predict future events using historical data sets. These algorithms have proven to be accurate and correct by being regularly utilized in a range of applications (Shaik & Srinivasan, 2019).

1.2. Motivation

Ethio Telecom is one of the highly profitable companies in Ethiopia. However, customer churn is a tremendous issue that could influence the organization in any event, losing its standing on the off chance that it isn't identified at the beginning phase. Many businesses frequently lose customers, when they fail to anticipate when customers will leave (Jain, Khunteta, & Srivastava, 2020). Organizations that confronted the postponed forecast of clients attempt to obtain new clients which become extravagant with regards to time and money. In this manner, the organization should foresee the clients to stir early so it helps in holding them by upgrading administration or giving a proposal to the right clients. The main motive of this study is to provide Ethio-telecom with an easy and effective way to predict customers who are going to churn at a very early stage. By using this approach, the telecom company can determine how many customers have left or are likely to leave their service.

1.3 Statement of the Problem

Machine learning can be used to find models, assist decisions, and learn from data in a variety of business fields, including banking, insurance, e-commerce, and telecommunications (Tyagi,2019). When it comes to the telecom sector, the telecommunication industry generates and stores a tremendous amount of data that needs the application of data mining to uncover usefulinformation from these data sets and create a model that can learn from the data for bringing enhanced service using machine learning algorithms (Ahmad A. K., Customer churn prediction in telecom using machine learning in big data platform., 2019).

As any telecom company's primary objective is to maximize profit, however, churn is one of the major problems that businesses must address to thrive in the cutthroat market (Ullah I. R., 2019). To handle this problem company needs to know the percentage of customers who leave for another company in a given period, come up with a detailed analysis of the causes of the churn, and understand the behavior of customers that unsubscribe and move to another business competitor.

Despite being the only telecom service provider in the nation, Ethio Telecom has focused on relationship marketing and customer retention to keep its clients. However, where Ethio Telecom had direct client contact, customer confidence in the company's dependability, honesty, and loyalty was waning due to complaints about the service and problems with commitment,

communication, and compliance handling (Tadesse, 2013). Because of this customers always pursue comparing service quality and product cost to benefits with another compatible service provider. Therefore, churn prediction is vital for Ethio-telecom to retain its loyal and valuable customers and to enhance its Customer Relationship Management (CRM) administration.

Safaricom Ethiopia is a new rival firm that has lately begun to offer services and has significantly increased the rivalry for Ethio-telecom as a result. Customers now have an alternative option of mobile operators to choose from. So, this is where the competition beginsfor Ethio- telecom, and therefore the telecom has to shift from its acquisition strategy to a retention strategy to save its valuable customer from churning.

1.4 Research Questions

In the end, this study attempts to answer the following **research questions**:

RQ1. Which attributes are more significant to predict customers churning at Ethio-telecom?

RQ2. To what extent do the proposed models perform in customer churn prediction?

RQ3. Which classification algorithm is more suitable for constructing the telecom customer churn prediction model?

1.5 Objective of the study

1.5.1 General Objective

The general objective of this research is to construct a prediction model for determining Ethio-telecom customer churn using Machine Learning Algorithms.

1.5.2 Specific Objectives

For the realization of the general objective stated above, the following specific objectives are formulated.

- To collect an appropriate dataset
- To prepare a dataset with relevant attributes that can help to predict telecom customer churns
- To prepare data for training and testing.
- To select suitable machine learning algorithms for customer churn prediction.

- To construct a customer churn prediction model using the selected machine learning algorithms.
- To evaluate the model using performance measures.
- To select a best-performing model.

1.6 Significance of the Study

Customers quit different firms for a variety of reasons. Product owners frequently lack knowledge about numerous possible pain spots including poor quality construction and missing features, unattractive design, and bad customer service. The main goal of this research is to employ a vast amount of customer-related telecom data to discover valuable churn consumers to improve customer churn prediction models and practices.

Customers would profit from the work as a result of the improved service, with Ethio-telecom as the primary beneficiary. This study aims to create a prediction model that can categorize consumers into churner and non-churner divisions based on customer history, hence minimizing and limiting telecom customer churns. This study will improve the way the company manages its present client connections and foster consumer loyalty. Customer attrition can prevent Ethio-telecom from losing its companies and seeing a reduction in profit.

Additionally, the study adds to the body of knowledge, particularly for other researchers with an interest in the same field. It will be crucial for later scholars to widen their scope and conduct a comprehensive analysis of the issue at hand. The study reveals which machine learning (ML) algorithms are best for predicting customer attrition in the telecom business domain.

1.7 Scope and limitation of the Study

The main task here is to forecast customer that is likely to churn from Ethio Telecom. In particular, machine learning algorithms are trained to learn the behavior patterns of clients who have already ceased their services. Correlations between the behaviors of churning and non- churning clients are then performed. As a result, the system can identify the consumers who are most likely to leave. Depending on the customer's intention to terminate the service, churning can be classified as either voluntary or non-voluntary (Bhatnagar, A., & Srivastava, S., 2020). When a client decides to switch services or businesses voluntarily due to discontent, betteroffers, or other factors, this is known as churn. Customer satisfaction, loyalty, switching costs,

service quality, and competitive pressure all have an impact on voluntary churn, whereas non-voluntary churn occurs when a company terminates a contract on its own or a customer terminates a contract without intending to switch to a competitor (Amity, 2018). Our work is limited for a customer that will churn voluntarily.

The customer Relationship Management (CRM) framework's churn management approach entails two significant analytical modeling initiatives (Shirazi F. &., 2019): While the second assesses the best way a service provider might respond to keep clients, the first one foresees customers who are ready to go. This study concentrated on the earlier study, which only dealt with foreseeing which consumers would leave the company.

The study's service category is restricted to mobile, data, and internet consumers alone. The research was performed based on the secondary data that is already recorded for customer management.

1.8. Application Result

These days, basically everybody has a telecom membership, and how much telecom information is quickly developing as a result of simpler correspondence and some specialist organizations (Liyanage, 2022). This makes it hard for CRM frameworks to assemble client information from many resources, or channels, between the client and the business to decide the situation with clients and administration utilization information. After successful completion of the work, it willbe possible to tell whether a customer churns or not depending on the historical data of the customer. This will help the customer management system to enhance customer retention strategy by identifying the behavior of churners in identifying factors behind churners. In addition to this telecom subscribers can benefit from the works as a result of service enhancement.

1.9 Organization of this thesis

This research study is organized into five chapters. Chapter 1 serves as the introduction, offering an overview of the background of the study, statement of problems, general and specific objectives, scope and limitation of the research, and significance of the research. Chapter 2 presents a comprehensive literature review, summarizing existing knowledge and identifying gaps in the field of the Telecom Customer Churn Prediction Model. In Chapter 3, the research

methodology presents an overview of data collection followed by data preparation and proposed model architecture. Chapter 4 focuses on the experimentation and result analysis. The evaluation and discussion of the model's results are also presented in this chapter. Finally, we concluded the study by summarizing key findings and providing recommendations for further improvements.

CHAPTER TWO

2 LITERATURES

Overview

The telecommunications industry is made up of businesses that enable global communication, whether it is via the phone, the internet, radio waves, or cables. These businesses build systems that enable the transmission of data, including text, voice, audio, and video, throughout the globe.

A telecom service is a good offered by a telecom provider or a specific set of user-information transfer capabilities made available to a user group by a telecom system. The user is accountable for the message's informational content when using telecommunications services. The responsibility for message acceptance, transmission, and delivery is on the telecommunications service provider.

Additionally, broadcast communications organizations are putting more and more emphasis on productivity and rapid development as a result of rapidly progressing innovation and intense industry competition. In essence, both people and society are affected by these developments. The management of data streams, capital, and intercompany cooperation as well as interpersonal, financial, and natural risk factors that affect supportability are also related to the goals of the inventory network (Chen et al., 2021).

2.1 Telecom in Ethiopia

It is a well-known fact that Ethiopian Telecommunications is one of the earliest public telecommunications services in Africa, having created the first telephone line between Harar and Addis Ababa (Adame, 2021). The inter-urban network began to reach from Addis Ababa to various locations. The first telegraph link between Djibouti and the Ethiopian city of Dire Dewa was set up in the year 1906.[1]

Three different Earth satellite stations for international communication were developed in Sululta between 1979 and 1987. After establishing an electronic exchange in 1988, the provider

[3] https://www.ethiotelecom.et

developed an electronic radio link in 1989. Moreover, four high-capacity 60-channel national satellite stations were set up in 1994 at Addis Ababa, Gode, Mekele, and Humera.[2]

Since November 29, 2010, Ethio Telecom has been a part of Ethiopia's 2005/06 - 2009/10 GTP, according to the federal government's determination to devote resources to upgrading telecom services, which it sees as crucial to the future growth of the country. As a result, the nation's telecom products and facilities have been changed to the best levels feasible, boosting the nation's development and resulting in a significant paradigm shift in the sector (Adame, 2021).

Ethiopia's government has been moving towards liberalizing and privatizing key businesses, notably the telecom sector, since 2018. It created the Ethiopian Communication Authority (ECA) in 2019 to license commercial telecom operators and regulate the whole sector (Adame, 2021). When ECA granted Safaricom Ethiopia the nation's first private sector operator license in 2021, the country's telecom monopoly came to an end. The following year, Safaricom Ethiopia's operations got underway. Smaller cities like Diredawa and Bahirdar were the first to receive service from Safaricom Ethiopia, which is currently expanding its network to include Addis Ababa. Another potential in Ethiopia's telecom business is emerging in addition to operating licenses, such as mobile money, fintech, and network towers (Analytica, 2023; Vodafone, 2020). The mobile money service TeleBirr was launched by the state-owned mobile provider Ethio Telecom in 2021. Only Ethio Telecom is currently authorized to offer mobile financial services.[3]

2.2 Customer Relationship Management

Customer relationship management (CRM) has taken center stage as today's competitive markets have become more saturated and intense (Rafiki, 2019). More than ever, achieving a company’s objectives depends on having the capacity to comprehend and control a strong relationship with the consumer (Ahn et al., 2020). The current market paradigm is shifting from a product-centered stage to a customer-centered to obtain information about clients, their usage, and even projections based on past behavior, with the help of modern database technology.

Customer relationship management (CRM) is the process of locating customers, getting to know them, forming connections with them, and influencing how they perceive a company's goods or

[2] https://www.ethiotelecom.et

[3] https://www.ethiotelecom.et

services (Eskinder, 2021). According to (Vafeiadis et al., 2015) It is a broad approach toproducing, keeping, and strengthening loyal and lasting client connections. It is widely accepted and widely utilized in a variety of industries, including banking, insurance, telecommunications, and the retail market. CRM has four fundamental parts that cooperate to close how fruitful it is. Technology, people, processes, information, and insight all contribute to the company's ability tomake better decisions and take better actions regarding customer relationship management, witha focus on connections with customers and the ultimate goal of ensuring that customers are satisfied (Eskinder, 2021).

Since the client's steady loss rate fundamentally influences the organization's monetary market worth, organizations that effectively coordinate CRM into their work processes generally partake in an improvement in client maintenance power or the probability that a client will not leave (Rafiki, 2019). Because of this, most organizations assess their client esteem on a month-to- month or quarterly premise. Subsequently, most organizations survey their client esteem on a month-to-month or quarterly premise (Gebreegziabher & To, 2022). Developing a coordinated relationship with customers helps businesses retain their most valuable resource, their customers.Organizations are focusing on the adaptation and coordination of new developments, new cycles,people, and relationship marketing to aid them and to help their clients and retain them for a lifetime.

2.3 Dimensions of Customer Relationship Management

Business analysts and CRM analysts must comprehend the reasons why customers leave as well as the trends in behavior revealed by previous data on customer churn (Ullah et al., 2019). CRM may increase productivity, offer appropriate promotions to the group of customers that are most likely to churn based on similar behavior patterns, and significantly enhance the company's marketing campaigns by identifying the main reasons why consumers leave (Ullah et al., 2019).

There are numerous data mining and machine learning technologies available that can be used to analyze such data as a result of recent developments in the field of big data. These methods examine the data to determine the causes of client churn. CRM must use these strategies to increase its profits (Ullah et al., 2019). As a result, an effective churn prediction model is needed for telecommunications firms to predict possible churners, retain their important clients, and improve their customer relationship management (Idris, 2012).

Dimensions utilized in CRM to improve customer management strategy include customer identification, customer attraction, customer retention, and customer development. Customer identification and retention are CRM's top priorities (Rodrigues Chagas et al., 2019). To realize this objective ideas like artificial intelligence (AI), computational intelligence (CI), machine learning (ML), and data mining (DM) were created to manage and use customer data. These technologies are used by CRM to make it possible to collect, analyze, and share information on current and potential customers to more precisely pinpoint their needs and forge intelligent connections (Rodrigues Chagas et al., 2019).

2.4 Background of customer churn

The term churn was defined in multiple ways in various works. According to (Ullah I. R., 2019) churn is a problem that arises when a customer leaves one service provider and changes to another, resulting in a significant loss of revenue. This occurs When a substantial percentage of clients are unhappy with the services offered by any telecom operator. The percentage ofcustomers who sever ties with a business or service in a specific time frame is known as thechurn rate (Jaber et al., n.d.).

Many firms are now forced to provide identical goods with the same degree of customer service and product quality due to increasing market rivalry. On the other hand, client demands rise and customer loyalty falls as competition grows. Customers today need responsive service in the shortest amount of time, efficient service delivery techniques, value-added products that cater to their particular needs, comparably low transaction costs, simple or easy procedures, and appealing services that are tailored just for them (Gebreegziabher & To, 2022).

Customer churn is a big issue in service businesses when services are intensely competitive. Customer-based businesses worry about customers who decide to leave because it costs substantially more to recruit new customers than it does to keep an existing one (Jaber et al., n.d.). Differently, anticipating which customers are most likely to leave the company early on could offer significant additional revenue sources (Ahmad et al., 2019).

Low levels of buyer happiness, tough rivalry strategies, novel goods, restrictions, etc. could all contribute to churning in today's changing market climate. Churn models seek to recognize early churn signals and consumers who are more likely to depart freely (Vafeiadis et al., 2015).

Churning also has a negative impact on a business's identity and standing in general. A loyal customer generates substantial income for the company and gets rarely impacted by competitors. These customers help the business succeed by referring it to others in their networks (Ullah et al., 2019).

Losing a customer to a competitor is a serious issue in the telecom industry as well. The telecom industry is expanding quickly, making it simple for clients to shift to another operator. For telecom businesses, knowing customers who are inclined to leave is essential. Churn prediction is, therefore, an invaluable company statistic and one of the most crucial machine-learning applications in the telecom sector (Jaber et al., n.d.).

2.5 Machine learning

Overview

Because of the availability and affordability of modern computing technology, organizations and firms now record everyday events and data created by a wide range of sources including employees, machinery, or monitoring devices, and these data are preserved in a variety of forms (Gebreegziabher & To, 2022). It is no longer a secret that big data has impacted the success of hundreds of important professional enterprises. However, as more companies use it to store, examine, and extract value from their enormous amounts of data, it gets challenging for them to make the greatest possible utilization of the data they collect.

A machine learning algorithm is a type of computing approach that, by modeling how humans attain knowledge, uses input data to perform an intended action without being manually programmed. Numerous fields, including discovering patterns, imagery processing, rocket construction, banking, recreation, and biological computing have effectively used algorithmic learning techniques (Bell, 2022).

2.5.1 Random Forest

A Random Forest is a supervised machine-learning method that is used to resolve regression and classification issues. It labels data by building different classifiers to increase accuracy in prediction. As the trees are built and the resulting individuals are integrated to predict the class label, the data set is tested using Random Forest approaches. Because of this, it is frequently

employed in situations where an extensive amount of information is required to be grouped using a Decision Tree Classification Algorithm (Shaik & Srinivasan, 2019).

It is now commonplace to employ random forests in many different applications. The primary aspect influencing the random forest's attractiveness is its capacity to handle non-linear classification jobs well. Especially for large datasets, random forest is widely known for addressing data disparities in many classifications (Paul et al., 2018).

Numerous studies show the advantages of random forests, including their capacity of generating error, deliver accurate classification results, handle enormous volumes of training data instantly, and offer reliable methods to approximate missing data. Since it requires a lot much data and builds numerous tree models on a single processor to create random trees, long computational times are common problems when applying random forest (Azizah et al., 2019).

2.5.2 K-Nearest Neighbor

K-Nearest Neighbor is a non-parametric supervised machine learning approach. It classifies objects with regard to their categories of their immediate neighbors and is distance-based. Whilst it can be applied to regression problems, KNN has been commonly utilized for issues with classification. In order to classify objects, K-Nearest Neighbor (KNN) compares prior andcurrent data to find the data that is closest to the object. KNN employs proximity formulas to the learning process to determine the distance to the nearest neighbor (Lubis et al., 2020).

The KNN algorithm's accuracy can be affected by the value of the hyperparameter k, which the user must select. Usually, a higher k value will result in a more accurate classification, but it can also result in overfitting. Based on the particular data set, k's ideal value will vary. Using KNN, items are grouped based on learning data that is nearer to the object. This approach serves to classify new objects using features and training data. After getting a query point, it will find a group of K objects or training points that are the closest. The anticipated outcome of the query will be determined based on the neighbor classification. Before performing computations using the K-Nearest Neighbour methods, it is important to first identify the training and test data (Lubis et al., 2020).

2.5.3 Logistic regression

Logistic regression is a supervised learning algorithm that is used to predict the probability of a categorical outcome. The outcome can be binary or multiclass. Because it is a linear model, it assumes that the connection between the independent variables and the dependent variable is linear. The logistic function is used to compute the probability that serves as the model for the dependent variable (Jain et al., 2020).

The logistic function transfers real numbers to the range [0, 1] using a sigmoid function. This indicates that a probability, which is a number between 0 and 1, may be modeled using the logistic function. With the help of the maximum likelihood estimation (MLE) algorithm, the logistic regression model is trained. The MLE approach identifies the model's parameters that maximize the likelihood that the observed data will occur (Ray, 2019).

Once trained, the model may be used to estimate the likelihood that a new observation would belong to a certain class, allowing the user to choose the class with the highest likelihood. A versatile tool, logistic regression can be used to address several issues. It works well in situations where the outcome is categorical and there is a linear relationship between the independent and dependent variables (Vafeiadis et al., 2015).

2.5.4 Support vector machine

Support Vector Machines are supervised learning models that can be used for classification, prediction, and clustering problems. An SVM builds a model that can categorize incoming data into one of two classes using a collection of input observations and related binary outputs. Support vector machine is mostly used because it produces great accuracy while using minimal computing power (Gebreegziabher & To, 2022).

SVMs are based on the idea of finding a hyperplane that separates the data points into two classes. The hyperplane is chosen such that it maximizes the margin between the two classes. For classification tasks, the SVM algorithm finds a hyperplane that separates the data points into two classes. For regression tasks, the SVM algorithm finds a hyperplane that minimizes the error between the predicted values and the actual values. They are particularly well-suited for tasks

where the data is not linearly separable. SVMs are also relatively robust to noise and outliers. However, SVMs can be computationally expensive to train, especially for large data sets which are also can be difficult to interpret (Cervantes et al., 2020).

2.6 Review of Related Works

Numerous types of research have examined from various angles and viewed how machine learning techniques might be used in the global environment to develop customer churn prediction models in the telecom industry. Churn prediction has been done in the literature using a range of techniques, including machine learning, data mining, and hybrid methods. Datamining was the most popular and commonly used technique for predicting telecom churn. These methods aid in decision-making and assist businesses to recognize, anticipate, and keep churningcustomers. Most of the researchers used a few ways to present the comparison study (Ullah et al.,2019).

2.6.1 Foreign-Related Works

Predicting customer churn prediction in the telecom sector using various machine learning techniques (Gaur, 2018): In this study, classification models for predicting client turnover are developed utilizing machine learning methods such as Logistic Regression, SVM, RandomForest, and Gradient Boosted Tree. Gradient boosting, with an AUC value of 84.57%, produces the most accurate models, according to the calculated AUC values. Logistic Regression, SVM, Random Forest, and Gradient Boosted Tree were the four machine learning models that were trained for this investigation, according to the report. Gradient Boosting performs better than the other three models, Random Forest, and SVM, than Logistic Regression.

A Churn Prediction Model using Random Forest: Analysis of Machine Learning Techniques for Churn Prediction and Factor Identification in Telecom Sector (Ullah I. R., A churn prediction model using random forest: analysis of machine learning techniques for churn prediction and factor identification in telecom sector, 2019): This study not only identifies churning customers through classification and clustering methods, but it also offers a churn prediction model that identifies the causes of client departure from the telecom industry. In order to deliver group- based retention offers, the proposed approach first classifies the data and then divides the data of the churning consumers into groups. Two datasets are used in this investigation. A South Asian

GSM mobile service provider made the first dataset available for use in assessing customer churn prediction issues. The second set of data is a churn-big dataset that is openly accessible. The suggested churn prediction model improved churn classification and customer profile, according to the results, by using the RF algorithm and k-means clustering. This research shows that the Random Forest (RF) approach performed effectively, correctly identifying 88.63% of samples.

BaYcP: A novel Bayesian customer Churn prediction scheme for the Telecom sector (Bhattacharya, Ladha, Kumar, & Verma, 2020): The study presented a novel, two-phased BaYcP strategy to deal with the problem of customer churn. In the first stage, using client data sets and decision trees, a risk profiling score (RPS) is created. It is then compared to a threshold value. Based on scores that are higher than a threshold, an ideal prediction model is subsequently constructed using a Bayesian classifier with carefully picked features. After being trained and tested, the model surpasses other cutting-edge techniques with an accuracy of 97.89%.

A Review and Analysis of Churn Prediction Methods for Customer Retention in Telecom Industries (Ahmed, 2017): This paper summarizes the churn prediction techniques to have a deeper understanding of customer churn and it shows that the most accurate churn prediction is given by the hybrid models rather than single algorithms so that telecom industries becomeaware of the needs of high-risk customers and enhance their services to overturn the churn decision. According to the paper hybrid methods either hybrid machine learning, hybrid meta- heuristics, or sometimes both, provide high accuracy of churn prediction. Hybrid models using SVM, ANN, SOM, etc. provide high accuracy with reduced computation complexity.

Churn Prediction using Ensemble Learning (Wang, 2020): This study contrasts the most widely used methods for categorizing the problem of customer churn in the telecom sector. The main goal of this study is to assess and compare the performance of a few widely used classification algorithms using a publicly accessible dataset. To evaluate the accuracy of each candidate model, a publicly available dataset from the Telecommunication Company was used. Based on model performance, the top four models—Light GBM, XGBoost, Random Forest, and Decision Tree—have been selected.

Customer churn prediction in telecom using machine learning in big data platforms (Ahmad et al., 2019): The study's key contribution is the creation of a churn prediction model that helps telecom companies identify consumers who are most likely to experience churn. The model

utilized in this work establishes a novel method for feature engineering and selection using machine learning techniques on a massive data platform. The model's effectiveness was assessed using the Area Under Curve (AUC) standard measure, and a value of 93.3% was found. Utilizing Social Network Analysis (SNA) features to add custom social networks into the prediction model is another significant contribution. The model's performance improved from 84 to 93.3% when compared to the AUC benchmark because of the usage of SNA. The model was built and tested in the Spark environment using a large dataset that was produced by converting vast amounts of raw data supplied by the telecom company SyriaTel. The dataset that SyriaTel used to train, test, and evaluate the system contains all client data from the previous nine months. The algorithms Decision Tree, Random Forest, Gradient Boosted Machine Tree ("GBM"), and Extreme Gradient Boosting ("XGBOOST") were tested using the model. However, employing the XGBOOST approach led to the best results.

Table 1: Related works

S/N	Title	Authors	Objective	Gap	Algorithm	performance
1	Churn Prediction using Ensemble Learning	Wang, X., Nguyen, K., & Nguyen, B. P	Presents a comparative study of the most widely used classification methods on the problem of customer churning in the telecommunication sector. benchmark the performance of some widely used classifications Using public churn dataset from the Kaggle competition	Could not find real-time dataset and realtime user attributes.	Light GBM, XGBoost,Rondam forest, Decision tree	0.6861, 0.6793, 0.6664, 0.645 with AUC metrics
2	A Churn Prediction Model Using Random Forest: Analysis of Machine Learnin	Ullah, I., Raza, B., Malik, A. K., Imran, M., Islam, S. U., & Kim, S. W	proposes a churn prediction model that uses classification, as well as clustering techniques to identify churn customers using classification algorithm cluster churn customer in a group for a retention offer, provides the factors behind the churning of customers in the telecom sector	Performance is not excellent	Random Forest J48 k-means	88.63% Correctly classified data and 88.58% accuracy

	g Techniques for Churn Prediction and Factor Identification in Telecom Sector					
3	predicting customer churn prediction in telecom sector using various machine learning techniques	Gaur, A., & Dubey, R.	Focuses on various machine learning techniques such as Logistic Regression, SVM, Random Forest, and Gradient boosted tree and also compares the performance of these models.	Performance only relatively good	Logistic Regression, SVM, Random Forest and Gradient	Gradient boosting 84.57%, most accurate LR and RF is an average and SVM underperformed

	Customer churn prediction in telecom using machine learning in big data platform	Abdelrahim Kasem Ahmad, Assef Jafar and Kadan Aljoumaa	This study provides the best understanding of the telecom dataset, and features used in different research Another main contribution is to use of customer social networks in the prediction model by extracting Social Network Analysis (SNA) features	Unable to finalize the reasons for customer churn	DT, RF, GBM and XGBOOST‖	93.3 90.89 78.47 72.2 for XGBoost, GSM, Rondam forest, decision tree respectively

2.6.2 Gap Analysis

Although there are many studies on customer churn prediction using machine learning techniques in different industries and contexts, there is a lack of research that focuses on the specific case of Ethio Telecom. Ethio Telecom has its characteristics and challenges that may affect customer churn behavior and the performance of machine learning algorithms. Therefore, there is a need for research that addresses the customer churn problem for Ethio Telecom using their own data and evaluates the effectiveness of different machine learning techniques for predicting churn. This research aims to fill this gap by conducting a comprehensive analysis of customer churn prediction for Ethio Telecom using machine learning techniques.

In Related Works above, previous research works have shown that the performance of these algorithms is often limited in terms of the accuracy of the models in classification and prediction problems. Therefore, enhancing the performance is essential for developing a robust and reliable model that can achieve high accuracy and efficiency.

CHAPTER THREE

3 METHODOLOGIES

Overview

In the modern world, telecom businesses produce enormous amounts of data at a very fast rate (Ullah I. R., 2019). Data mining techniques are methods to discover useful patterns and insights from these large data sets. There are many kinds of data mining techniques, such as pattern mining, text mining, and web mining that can be applied to different types of data and problems. This work aims to investigate the existing techniques in machine learning and data mining to propose a model for telecom customer churn predictions. Data mining techniques can permit these companies the ability to mine historical data to predict when a customer is likely to leave. The proposed customer churn prediction model involves four main processes, namely data acquisition, preprocessing, model construction, and validation.

- **Data collection -> Preprocessing -> model-> validation**

3.1 Data Collection

The size and quality of the dataset affect how well a classifier performs. Attrition analysis and churn rejection, therefore, depend on the correctness of the data obtained, and the efficacy of the prediction model is dependent on the knowledge gained from the information that is accessible. In this study, Ethio-telecom is used to collect experimental data on customer account informationand status, demographic data, and service consumption.

3.1.1 Description of Original Data

We made an effort to incorporate significant knowledge about customer-related data that is available and recommended by the literature on customer retention to successfully finish the ML project. Both qualitative and quantitative client data were collected for this study using a tabular data-gathering method. The following client information was gathered in data of type Excel format.

- Demographic data
- Customer service usage data
- Customer account and status

Customer Demographics: Customer demographics are the location and population data for a specific customer or details on a group residing in a specific area. Demographic predictors relating to customers are collected, including age, gender, area, administrative region, and zone.

A firm can use consumer demographic data for a variety of objectives, including marketing, manufacturing, and customer service. Businesses can segment clients into distinct groups based on their similarities and distinctions by knowing their demographic information and then modifying offers and communications to each group's requirements and preferences.

Service usage: Telecom industries provide multiple services like data retail, Internet retail, voice retail, and wholesale. Data about how and when clients use telecom services, such as voice calls, text messages, data traffic, etc., are referred to as service utilization data in the telecom industry. Call data records (CDRs), which are files that contain information about each call made or received by a customer, such as caller and callee numbers, the date and time of the call, the duration of the call, and the type of service used, have historically been used to collect service usage data in Ethiopian telecommunication. Customer service usage is customer information mainly related to the service the customer use and demonstrates the preference of specific customer which helps the industry to predict the overall customer interest in both pre-paid and post-paid telecom service.

Customer account and status: Customer account information, which is a subset of customer account and status data in telecom, is the data that contains the fundamental details of a customer who subscribes to a telecom service, such as name, address, phone number, email, etc. Information from customer accounts helps locate and contact clients, as well as confirm their identification and suitability for particular services or promotions.

3.1.2 Ethiopian Telecommunication Customer Classification

Ethio-telecom classifies its customers into five dimensions to ensure close follow-ups, proper documentation, and customer due diligence. This includes customer identification, segmentation, profiling, risk assessment, and monitoring. By categorizing customers based on these dimensions, Ethio-telecom can better understand and manage potential risks and issues.

Residential: These customers include both sizable customer groups as identified by Ethio Telecom and those who buy Ethio goods and services for their own usage. They are individual

customers who are able to enroll in any service. Despite making up the majority of the division, they do not bring in the majority of the revenue. This group also receives additional attention as part of the company's social and development objectives, regardless of how it affects revenue (Tadesse, 2013).

Key Account Customers: These are major organizations that rely heavily on telecom services and have ongoing follow-up requirements. Not only is that, but the majority of the revenue generated by these organizations is derived from this particular consumer base. These groups are in charge of producing significant revenue for the country. Their Ethio Telecom membership includes more than one service type, and a dedicated task team is in charge of managing their accounts. Examples include NGOs, embassies, financial institutions, universities and colleges, and the Ministry of Defense (Tadesse, 2013).

SME (Small & Medium Enterprises): These groups are more modest in size than significant account clients in terms of their revenue-generating and business-sensitive contributions. They play a key role in the financial growth of the country in addition to their contribution to income generation. PLCs and small-cap corporations are included in this category.

SOHO (Small Office-Home Office): They are the smallest in the Enterprise category despite being slightly bigger than Residential. In addition to using telecom services, they run a very small office. Both internet cafes and business centers fall under this category.

Ethio Telecom Employees: One of the consumer segments in Ethiopia is the employees of Ethio Telecom. They can take advantage of several savings and benefits offered by Ethio Telecom, including free voice calls, mobile data, and internet access. Additionally, they are very devoted to and trusting of the business.

3.2 Data preprocessing

The classifier is able to reach the necessary training level with the aid of the properly preprocessed dataset, which finally results in the desired performance. The large size, high complexity, and imbalanced nature of telecommunication datasets, according to the research, are the key obstacles to achieving the desired performance for churn prediction (Özmen, 2020). The telecom industry generates an immense amount of data, but that data has missing values, which

makes prediction models perform poorly. Similar to this, there are considerably fewer churners in the telecommunications sector than there are non-churners, which could lead to subpar learning by a classifier. The choice of features in current models is another significant issue. For the classifier to train well, the preparation phase fundamentally calls for data cleaning, feature extraction, and reduction procedures.

3.2.1 Data cleaning

Data cleaning, the initial step in the preprocessing of data included locating and eliminating or correcting any errors, discrepancies, outliers, or missing values that can impair the performance or accuracy of the models. We handled the inaccurate or incomplete data using a variety of strategies, including deletion, which involved removing any duplicate records from the dataset that would have introduced noise or bias into the data analysis. We looked over the data, corrected any typos or spelling errors that would cause confusion or misinterpretation of the data values incorrectly, and substituted estimated values for any missing numbers based on other available information. As a simple technique that can maintain the data's distribution, we employed mean imputation. The mean of the non-null values in the same column is used in this approach to replace each null value.

3.2.1.1 Noise Removal

Since noisy data could result in bad outcomes, making the data meaningful is essential. The telecom dataset contains a large number of empty rows, and incorrect values like null and unbalanced characteristics. Our dataset has a total of 15 attributes. We looked at the dataset for filtering to reduce the number of features and make sure the dataset only contains useful features.

The 15 features of the dataset that Ethio Telecom provided, together with a description of each feature, are listed in the table below.

Table 2: Dataset features

No	Features	Description
1	SERVICE NUMBER	Phone number of the customer
2	CUSTOMER CATEGORY	The type of customer that the customer belongs to is based on the Ethio Telecom customer definition.
3	CATEGORY	Prepaid and postpaid which shows payment plans for telecom services
4	ADMINISTRATIVE REGION	The geographical area that shows the admirative division of the customers
5	ZONE SUB-CITY	Subdivision of a city the customer belongs to
6	GENDER	Gender of the customers
7	AGE	Age of the customer
8	TOTAL TRAFFIC MINUTE	Total traffic usage of the customer per minute
9	TOTAL FEE ETB	Total traffic usage of the customer in ETB
10	SMS_USAGE_MIN	SMS usage of the customer per min
11	SMS_FEE_ETB	SMS usage of the customer per ETB
12	DATA USAGE_MB	Data usage of the customer per MB
13	DATA REVENUE_ETB	Data usage of the customer per ETB
14	VOICE USAGE_MIN	Voice usage of the customer per min
15	CHURN	A feature that shows whether the customer is churn or not

3.2.1.2 Feature selection

Feature selection, a crucial preprocessing step in machine learning, involves choosing the most pertinent and instructive features from the data that might enhance the functionality or accuracy of the models. In order to assess and rank the features according to their result score, we used a variety of algorithms, such as information gain and correlation coefficient.

It chooses the most crucial characteristics by removing extraneous or redundant features from the initial feature set. Most feature selection algorithms place a high priority on reducing duplicate information and maximizing relevant information (Zhou et al., 2022). Finding the best set of

features that makes it possible to create optimized models is the aim of feature selection approaches in machine learning. The data variables you use to train your models have a significant impact on the performance you can achieve. Choosing variables involves manually or automatically selecting the characteristics that are essential to your prediction variable or desired result. Feature selection is a method for enhancing the generalizability and accuracy of a prediction model by removing noisy and redundant information.

According to the relationship with classifiers, feature selection methods can be divided into three categories which are filter methods, wrapper methods, and embedded methods (De, 2021). In this paper filter feature selection method has been employed because the choice of the filter feature does not depend on the classifier. Filter feature selection method is a type of feature selection method that applies a statistical measure to assign a score to each feature (Saheed, 2021). The features are ranked according to the score and either chosen to be preserved in the dataset or rejected. The approaches are frequently univariate and take the feature into account either separately or in relation to the dependent variable. The process first selects the features for the data set, and then it trains the classifier. The choice of features has no bearing on the subsequent classifier. It is comparable to filtering the initial features using the feature selection approach before training the model using the filtered features. The expense of this procedure is rather inexpensive (Zhou et al., 2022). The following feature selection strategies are used in this study.

3.2.1.2.1 Information Gain

Information gain calculates the reduction in entropy from the transformation of a dataset. It can be used to estimate how much a specific transformation reduces the entropy of a dataset. This measure is frequently used in feature selection, where the knowledge acquired from each variable is assessed in relation to the desired outcome (Saheed, 2021). In essence, information gain measures the amount of new and pertinent information a particular variable adds to thetarget variable, and in this way, it can be used to identify the most significant features in a datasetto build more precise and successful models for prediction and classification tasks.

3.2.1.2.2 Correlation Coefficient

Correlation is a method for determining the linear relationship between two or more variables. Through correlation, we can predict one variable based on another. The argument for using

correlation in feature selection is that good variables have a significant connection with the aim. Additionally, variables should not be associated with one another but rather with the goal. If two variables are associated, we can predict one variable from another. The model just needs one connection because the second feature doesn't add any additional information.

3.2.2 Data Representation

In this step, we perform data transformation on the features that have text-based values. We transform the text data into numerical representations that the machine learning algorithms can use as input. We can standardize the scale and format of the text data through this approach, which also makes it compatible and comparable to machine learning techniques.

The process of encoding and storing information in a way that a computer can comprehend and manipulate is known as data representation. It entails changing the data's format, such as going from text to binary. Choosing the right format and structure for the data, such as using bits, bytes, characters, pixels, or samples, is another aspect of data representation. By lowering the variability in attribute values in particular areas, data aggregations and representation are crucial for producing results that are more expressive and understandable. The dataset contains six separate values for the client category, encoded numerically as follows: domestic (1), small office home office (2), small and medium enterprise (3), corporate business (4), Ethio employee (5), and large enterprise (6). Additionally, category variables that have been converted to binary values are included in the dataset. The category variable, for instance, is dichotomized into prepaid (1) and postpaid (0). The age variable is recorded as male (1) and female (2) as well. Target attributes are also changed to churn (1) and non-churn (2), and the values of zone sub-cityattribute values are changed to numbers between 1 and 83.

3.2.3 Data normalization

Normalization or Standardization or Feature Scaling involves re-scaling of values to be able to make processing easier. Even if the data are compared on several scales, standardization helps the comparison of values appear simpler. After data has been collected from various sources and before it is posted into the target, standardization involves the transformation of datasets. To get the best results, feature scaling is crucial (Raju et al., 2020).

The optimum pre-processing method for data before the application's training phase is normalization. To achieve strong prediction outcomes in a shorter amount of time, normalization employing metrics with criteria is considered to be of utmost importance (Raju et al., 2020).

3.2.4 Dataset splitting

The process of splitting the available data into distinct subsets for training, testing, and assessing the model is known as dataset splitting in machine learning. To avoid overfitting or underfitting and to make sure the model is trained and evaluated correctly, it is crucial to take this step.

We employed the train-test split approach, which separates the data into training and testing sets. The testing set is used to evaluate the model's performance after it has been trained using the training set. In line with the traditional split of 70–80% for training and 20–30% for testing, this study splits training and testing at 80% and 20%, respectively.

3.3 Modeling

The rapidly expanding discipline of data science includes machine learning as a key element. Algorithms are trained to create classifications or predictions and to find significant insights in data mining projects through the use of statistical approaches. These conclusions then influence decision-making within applications and enterprises, hopefully having an impact on important growth KPIs.

The main goal of this study is to forecast customer attrition by using computational methods to analyze current Ethio Telecom customer behavior. The research also tries to pinpoint the machine learning approach that predicts customer attrition the most accurately. Several machine learning algorithms were studied in this work, and effective ones were used to find a model in the data. In earlier related work, the effectiveness of each machine-learning model was investigated and analyzed.

Each machine learning algorithm was chosen based on its benefits and previous success in other studies. K-nearest neighbors (KNN), support vector machines (SVM), logistic regression, and random forest were the machine learning methods chosen for this study. Using those approaches, we created a variety of machine learning models, which we subsequently trained and tested using test data. Finally, we compared them based on how well they performed in terms of developing an efficient churn prediction model.

3.3.1 Random Forest

The random forest creates decision trees from several samples, using the majority vote for classification and the average in the case of regression. A random forest is composed of a specified number of binary decision trees. Each tree in the forest is grown using a bootstrap sample from the training dataset (Paul et al., 2018).

Two key ideas, bagging and random feature selection are the foundation of random forests. Bagging is the technical term for bootstrap aggregation, which means that each tree is trained on a random sample of the data with replacement, meaning that some data points may appear more than once or not at all in each sample. As a result, individual trees' variation is decreased, and they are less likely to overfit (Shumaly et al., 2020). At each node of the tree, just a random subset of the characteristics is taken into account for splitting rather than all of the features. This is known as random feature selection. As a result, the trees are less correlated with one another and become more resilient and diversified. The majority vote (for classification) or mean (for regression) of all the forest's trees' predictions is used to determine the random forest's final forecast. The primary benefit of employing RF is that there is no requirement for dimensionality reduction or feature selection because the data is already extremely highly dimensional. Additionally, parallel models have a higher training rate and are simpler to utilize (Rahman & Kumar, 2020).

3.3.2 K-Nearest Neighbor

KNN is instance-based learning, often known as lazy learning. When all of the training data is accessible, KNN performs at its best. The ideal value for k must be selected based on the available data. Larger values of k often reduce the effect of noise on classification, but they also make the distinctions between classes more hazy (Wang, 2019). A churn prediction model can be

developed using the K-Nearest Neighbor (KNN) approach. One advantage of this approach is that it can do classification jobs without knowing the distribution of data beforehand (Nur Amir Sjarif et al., 2019). For the majority of similar chemicals, the KNN technique helps forecast a customer's features based on trial data.

The K-Nearest neighbor (KNN) algorithm uses similarity metrics to categorize a set of data points. The method compares the weights of many existing properties to determine how similar two instances are (Lubis et al., 2020). In our case, the closeness of each customer case to each previous churn case is assessed to determine which customers will be churned.

3.3.3 Logistic regression

Logistic regression can be binary, multinomial or ordinal (Jain et al., 2020). A binary logistic regression is used in this instance. The logistic regression uses real-valued inputs to predict whether an input class belongs to the non-churners or churners groups.

Using logistic regression, we have developed a function that takes customer data as input and produces a chance of churn. We can label a person as a churner if the probability is more than 0.5; otherwise, we can label them as a non-churner. We must identify the coefficients that best suit the data if we are to learn the logistic function from the data. Several techniques can be used to achieve this, such as gradient descent, which iteratively updates the coefficients by moving in the direction of the steepest descent on a cost function that gauges how well the logistic function matches the data (Jain et al., 2020).

3.3.4 Support Vector Machine

The main goal of pattern classification is to construct a model that performs as well as possible given the training data. The models are created to accurately identify each input-output combination as belonging to the class to which it belongs using traditional training techniques. If the classifier is too well-suited to the training set, the model starts memorizing the training set rather than learning to generalize, which diminishes the classifier's potential for generalization (Cervantes et al., 2020).

The main goal of pattern classification is to construct a model that performs as well as possible given the training data. The models are created to accurately identify each input-output combination as belonging to the class to which it belongs using traditional training techniques. If

the classifier is too well-suited to the training set, the model starts memorizing the training set rather than learning to generalize, which diminishes the classifier's potential for generalization (Cervantes et al., 2020).

The primary goal of SVM is to determine the best hyperplane, which is the one that is farthest from both classes. This is accomplished by identifying various hyperplanes that best categorize the labels, after which it will select the hyperplane that is the furthest away from the data points or the one with the largest margin. The margin is the separation between the closest data point for either class and the hyperplane. Support vectors are the data points that are on the edge of or nearest to the hyperplane because they help define or support the hyperplane (Cervantes et al., 2020). SVM uses constraints to determine the optimum hyperplane by minimizing the norm of the hyperplane's normal vector while still ensuring that every data point is correctly classified or has a minimal error. Numerous techniques, including quadratic programming and gradient descent, can be used to address this optimization challenge.

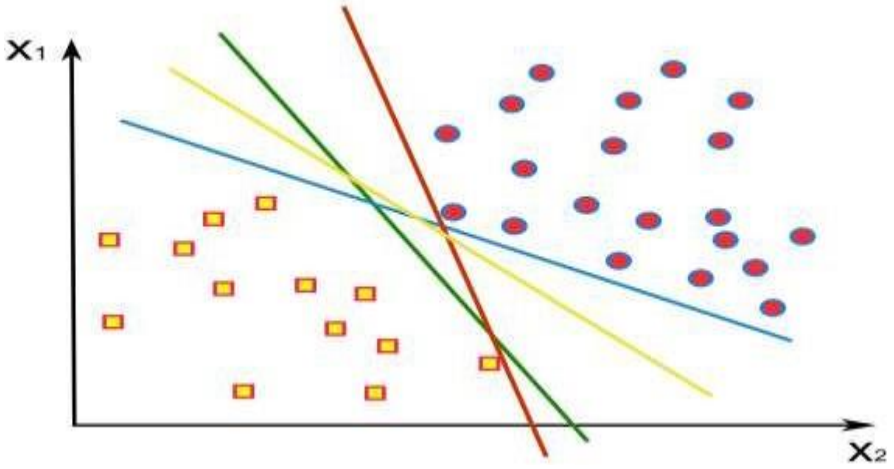

Figure 1: Separation of the hyperplane (Cervantes et al., 2020)

3.4 Proposed Architecture

The architecture shows a high-level procedure that is followed in the prediction of customer churn for Ethio-telecom. The gathering of customer information, which serves as the churn prediction model's input, is the initial step in this process. After preprocessing, the data is checked to make sure it can be used in the model. Preprocessing may entail activities like feature

engineering, feature selection, and data cleansing. This phase seeks to guarantee that the information is correct, comprehensive, and pertinent to the work at hand.

The data is divided into a training dataset and a testing dataset once it has undergone preprocessing. The machine learning algorithms that will be utilized in the model-buildingprocess are trained using the training dataset. The model's performance is assessed, and its abilityto forecast customer turnover is determined using the testing dataset.

The churn prediction model is built using a number of supervised machine learning techniques, such as Random Forest, Support Vector Machines (SVM), K-Nearest Neighbors (KNN), and Logistic Regression. These algorithms are selected for their propensity to manage the complex and high-dimensional nature of consumer data as well as their capacity to recognize patterns and linkages in the data.

The process of customer churn prediction comes from the selection of the best model out of those that have been developed and tested. The model of choice should be the most accurate and reliable at predicting which consumers are most likely to do business elsewhere. The customer churn forecast may then be applied to both new and existing customers using this model,assisting the company in keeping them as clients.

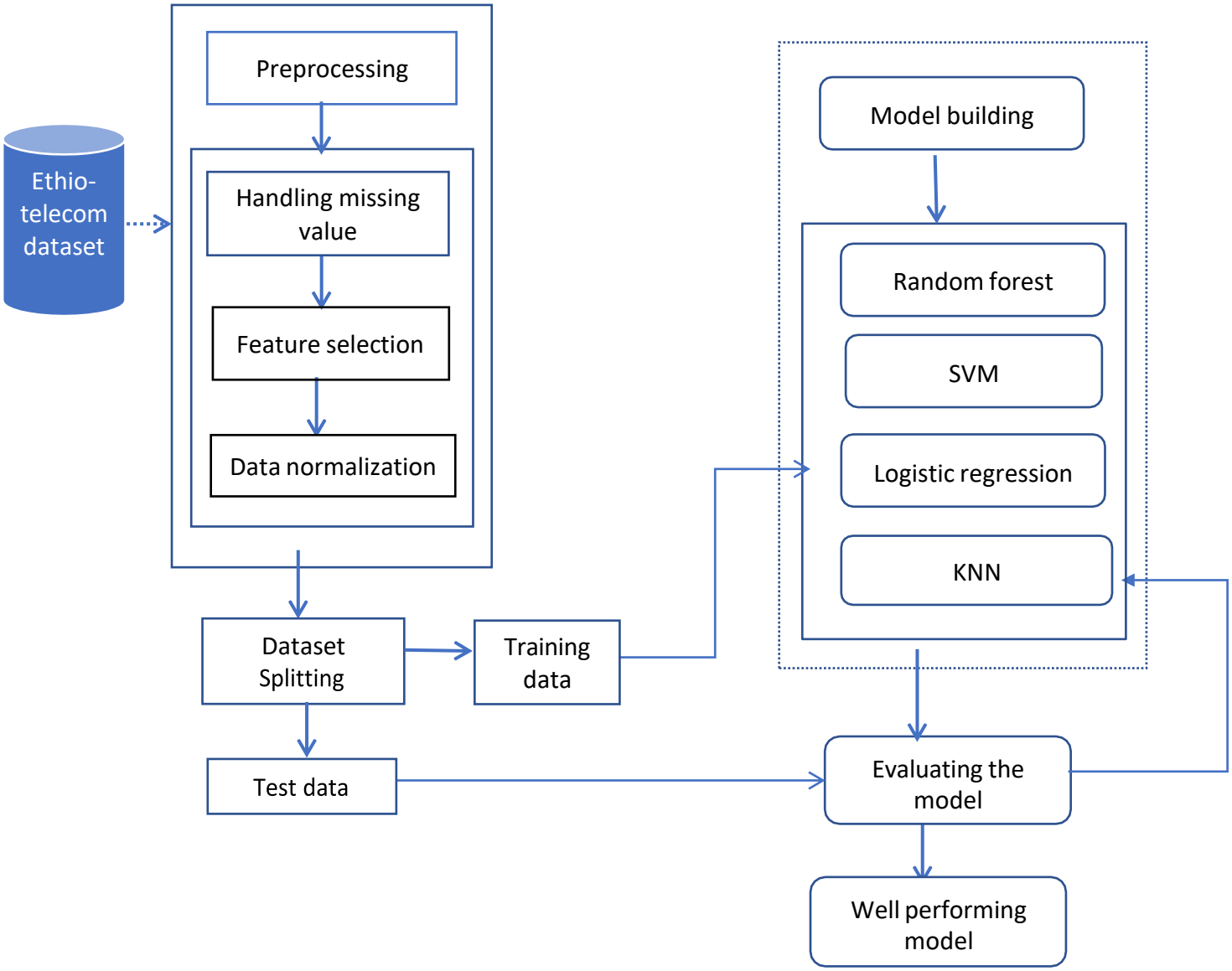

Figure 2: Proposed model architecture

3.5 Model Evaluation

This section outlines the evaluation procedure utilized for determining how well the customer churn prediction models work. Confusion matrices are frequently used in the domains of statistics, AI, data mining, machine learning, and similar ones to evaluate performance. To evaluate how well a task with two classes is being performed, confusion matrices are typically employed. Because the Confusion Matrix contains evidence regarding the actual and expected

classification carried out by the suggested prediction and classification model, it was used in the research as an evaluation tool.

3.5.1 Confusion Matrix

The confusion matrix is used to measure the performance of two class problems for the given dataset. Instances are accurately classified by the right diagonal elements TP (true positive) and TN (true negative), while Instances are mistakenly classified by FP (false positive) and FN (false negative). TN stands for the quantity of false negative classifications for an instance. FP is the quantity of false positives a certain instance has received. FN stands for the number of correctly classified instances as negative. TP is the proportion of correctly categorised instances that are Positive (Rahmad, 2020).

3.5.2 Kappa Statistic

The Cohen's Kappa coefficient (k), which compares the data to the fundamental truth and determines the accuracy of the random classifier in terms of measured and expected accuracy, is a measurement of the percentage of examples in a machine learning model that are properly classified. When (1)/(k) is the range of accuracy for the randomness. The accuracy is 50% in the case of binary classification when k = 2 (Vujović, 2021).

$$K = \frac{(p0 - pe)}{(1 - pe)} \quad \text{Equation 1}$$

p0 - total accuracy of the module,

pe - random accuracy (random accuracy of the model).

3.5.3 Accuracy

Accuracy is the percentage of instances that were correctly classified by the model. It determines the right prediction value for the classification issue as the total number of the model's correct predictions divided by the total number of data examples utilised in the model (Dj Novakovi et al., 2017). Accuracy is calculated using the equation:

$$Accuracy = \frac{TN+TP}{\text{total number of instance data}} \quad \text{Equation 2}$$

3.5.4 Precision

The percentage of positive cases that were projected to occur is known as precision. Precision quantifies the anticipated value as being true, and it reveals how often the model makes a correct prediction. Precision answer what proportion of identifications was correct. Precision calculated using the equation:

$$Precision = \frac{TP}{TP+FP} \quad \text{Equation 3}$$

3.5.5 Recall

Recall measures the proportion of positive events that the model properly classified (Yacouby, 2020). What percentage of real positives was accurately identified? Calculate using the following formula.

$$Recall = \frac{TP}{TP+FN} \quad \text{Equation 4}$$

3.5.6 F1 score

The precision and recall components of the F1 score are weighted harmonic means. The weighted harmonic mean of the test's precision and recall is known as the F measure, sometimes known as the F1-score or F-score, and it serves as a measure of test accuracy. When the dataset has an unequal class distribution, F1 is a more relevant metric than accuracy (Dj Novakovi et al., 2017). This score is calculated using this equation:

$$F1\ score = 2 * \frac{(Recall*Precision)}{(Recall+Precsion)} \quad \text{Equation 5}$$

CHAPTER FOUR

4 RESULTS AND DISCUSSION

Overview

This chapter outlines the primary experiments carried out to develop the best model that achieves the goal stated in chapter one, as well as the experimental approach and model-building process based on the suggested framework architecture. Additionally, we provide an analysis of the findings from each step.

The experimental evaluation shows that the suggested framework architecture is workable. It describes how support vector machines (SVM), logistic regression, random forest, and K-nearest neighbors (KNN) were used to develop machine learning models.

4.1 Implementation Environment

This study used multiple programming tools and packages to provide a solution for predicting customer turnover in Ethio Telecom. The data preprocessing, model construction and evaluation phases of the solution all used Python. Python was chosen because it is well-liked and appropriate for programmers, academics, and data scientists who deal with machine learning models. The following sections will discuss the hardware and software utilities used in training and testing the models.

4.1.1 Hardware Utilities

Processor: Intel(R) Core (TM) i7-1065G7 CPU @ 1.30GHz
RAM: 12 GB
Hard disk: 500 GB

4.1.2 Software Utilities

Operating System: 64-bit Windows 10
Python: 3.7.11
Jupiter Notebook: 6.4.6
The software tools and python packages that were used in this study, along with their versions are described below.

Table 3: Software tools

Tools	Version	Description
Application Software		
Google Chrome	114.0.5735.110	Web browser used to access Google colab.
Microsoft Excel	2010	It is utilized to process data by cleaning, filtering, and sorting the acquired information.
Mendeley Desktop	2.85.0	Mendeley Desktop allows users to easily import and store documents from various sources such as databases, websites, and PDF files and automatically generates citations and bibliographies in different citation styles.
Editor		
Google colab	2.7.0	It is a Python web IDE that allows users to write and run any Python code through a web browser. It is particularly useful for machine learning, data analysis, and education.
Programming language		
Python	3.7.11	It is a high-level, general-purpose programming language. Due to its ease of use, adaptability, and extensive community, it is frequently utilized for machine learning applications.
Python Packages		

sklearn	1.2.2	It is a free, open-source machine-learning library for Python. It provides a wide range of supervised and unsupervised learning algorithms.
google.colab	0.0.1a2	It is a collection of functions and classes that allow you to interact with Google Colaboratory. It provides access to the Collaboratory environment, such as the runtime, file system, and session.
Pandas	1.5.3	It is a Python library providing high-performance, easy-to-use data structures, and data analysis tools. It is widely used by data scientists and analysts for data cleaning, analysis, and visualization.
numpy	1.22.4	It is a scientific computing Python library. It offers arrays, which are quick and effective data structures for storing and handling numerical data, a high-level interface.
matplotlib	3.7.1	It is a Python library for creating static, animated, and interactive visualizations.
seaborn	0.12.2	It is a Python visualization library based on matplotlib.

4.2 Dataset for Experiment

For this study, data was collected from Ethio Telecom on February 27, 2023. The data collected contains a total of 7205 customer records with 15 distinct attributes originally. The dataset covers the client's account information and status, demographic information, and service usage while one of the attributes is the dependent field which indicates whether an event is a churn or not.

According to Chapter 3, We went through multiple rounds of data preprocessing to confirm the accuracy and compatibility of the data before applying the suggested models to the dataset. Data cleaning was the initial step, which required finding and eliminating any errors, discrepancies, outliers, or missing numbers that would have an impact on the models' performance or correctness. To deal with inaccurate or incomplete data, we employed several strategies, including deletion and imputation. The dataset size is reduced to 6239 records after data cleaningactivities were carried out on it.

4.3 Feature Selection

This section describes the process of feature selection, including the steps taken, the outcomes of those steps, and the decisions made based on the outcomes. Feature selection is a process of selecting the most relevant features from a dataset to improve the performance of a machine learning model (Zhou et al., 2022). In this study, the aim variable was churn, which refers to when a customer cancels their subscription, and two feature selection methods were used to assess the link between each feature and the target variable.

4.3.1 Information Gain

Information gain is a measure of how much information a feature provides about the target variable (Saheed, 2021). To perform feature selection, necessary libraries are imported and loaded into the dataset. We imported train_test_split, mutual_info_classif from sklearn.model_selection, sklearn.feature_selection library of python respectively to split a datasetinto random train and test subsets and to measure the dependency between two random variables.We have also imported files from google.colab which help to browse and upload our dataset. Thefollowing code illustrates how to do that:

```
import numpy as np
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.feature_selection import mutual_info_classif
from google.colab import files

uploaded = files.upload()
df = pd.read_csv('telecom_churn_dataset.csv')
```

- Then, we need to split the dataset into training and testing data and apply mutual information classification to them.

```
X_train,X_test,y_train,y_test=train_test_split(df.drop(labels=['churn'], axis=1),
   df['churn'],
   test_size=0.2,
   random_state=0)
mutual_info = mutual_info_classif(X_train, y_train)
```

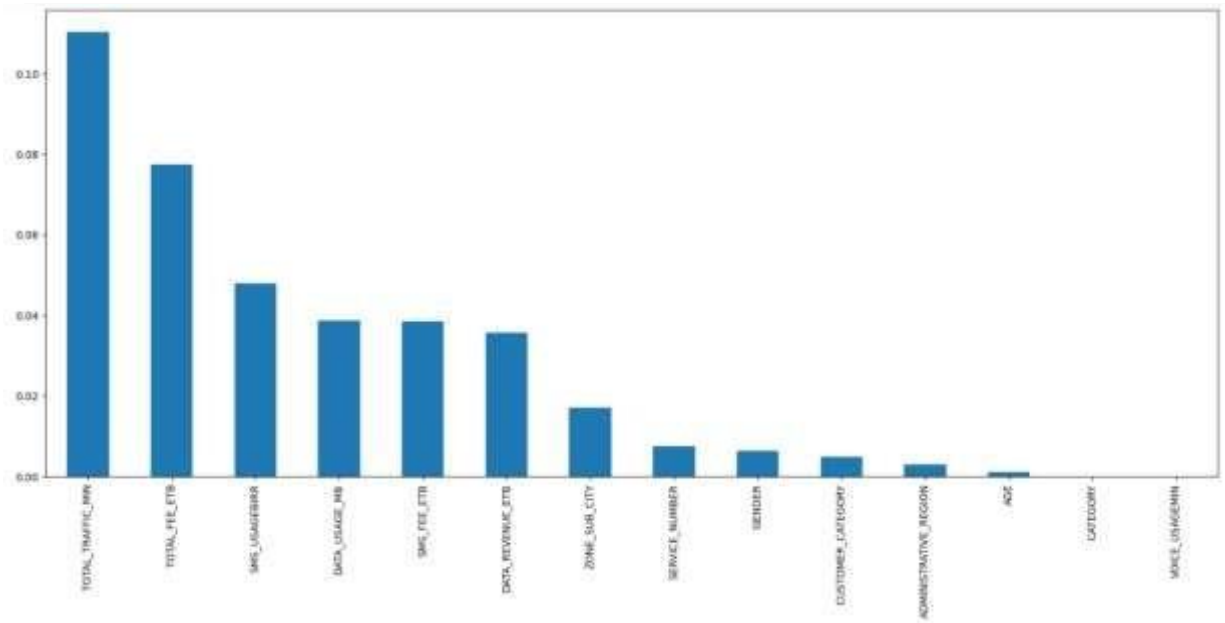

Figure 3: Information gain

The output of the information gain for each feature is shown in the above bar chart. The feature with the highest information gain is TOTAL_TRAFIC_MIN, while the features with the lowest information gain are CATEGORY and VOICE_USAGE_MIN. This suggests that CATEGORY and VOICE_USAGEMIN it is not very useful for predicting the target variable because they are not informative. Therefore, they are dropped from the dataset. This will help to improve the performance of the model by reducing the amount of noise in the data.

4.3.2 Correlation Coefficient

A statistical measure that measures the linear relationship between two variables is the correlation coefficient. Here, we employed the Pearson correlation coefficient to accomplish feature selection by locating traits that are significantly associated with the target variable and with one another. The intensity and direction of the link between two variables is represented bya number between -1 and 1 (Mu et al., 2018).

Two features were eliminated from the current feature selection after considering the outcomes of the previous feature selection, leaving the remaining 13 features. There are two approaches to calculate the correlation coefficient: Between-feature feature correlation and target feature correlation (Al-Mashraie et al., 2020). The association between several characteristics in adataset is referred to as feature correlation between features. This can be used to spot featuresthat overlap or are potentially redundant (Sharma, 2020). Accordingly, if two features have a high degree of correlation, only one of them may be required for the model to learn.

Feature correlation to the target refers to the relationship between a feature and the target variable. This can be used to identify features that are important for predicting the target variable(Zhou et al., 2022). Therefore, if a feature is highly correlated with the target variable, then it is likely to be a useful feature for the model to learn.

- **Feature correlation between features**

To perform a correlation test between features we have imported pandas, trian_test_split, and files from Python libraries to partition the dataset into training and testing sets and perform a correlation analysis on the features. The output is visualized as the following using the Seaborn heatmap diagram.

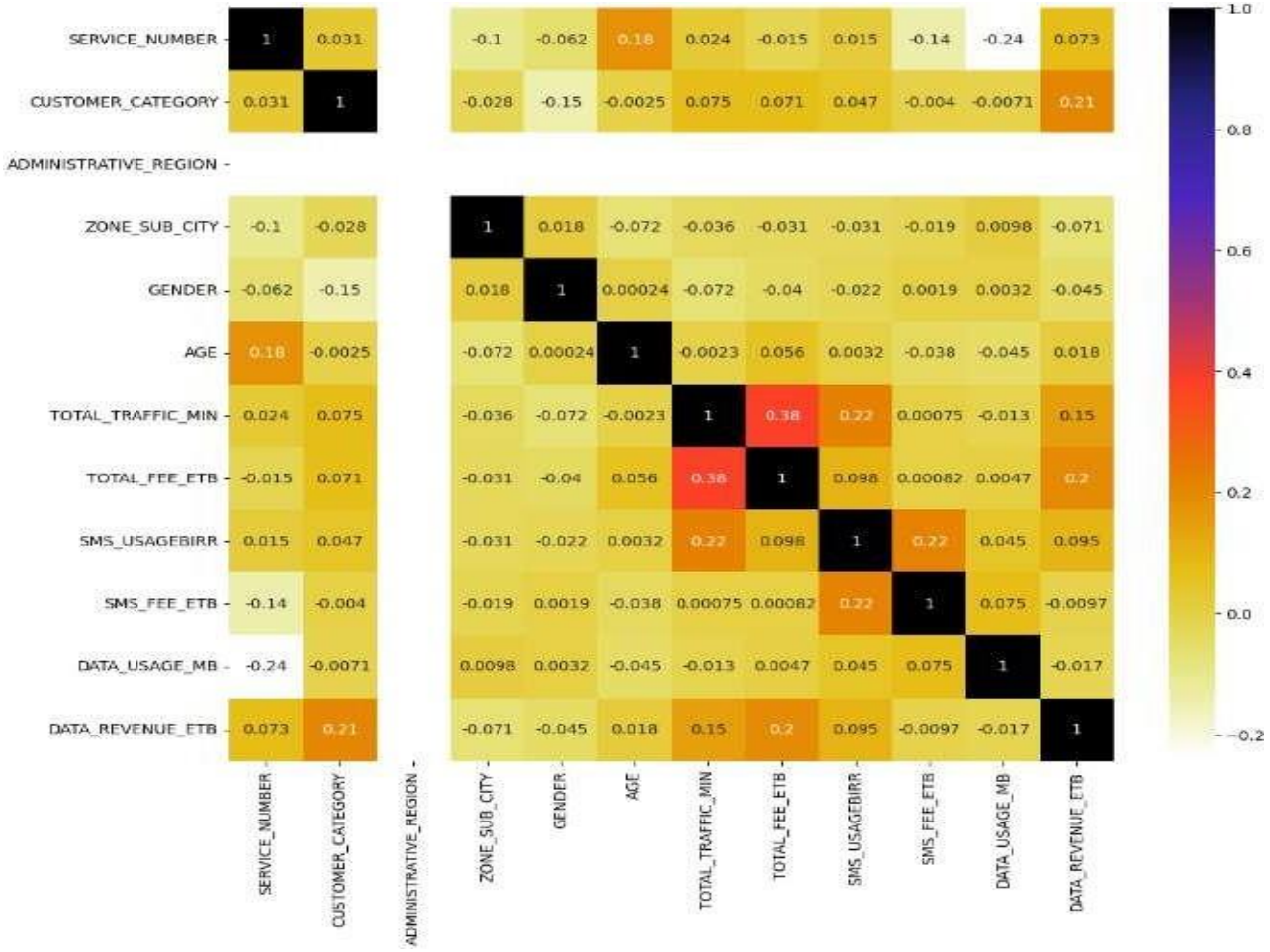

Figure 4: Feature correlation between features

From the diagram, we can analyze that: There are positive correlations between total traffic minutes and total fee (ETB), SMS usage (birr) and SMS fee (ETB), and data usage (MB) and data revenue (ETB), which means that these variables increase together. On the other hand, there is a negative correlation between age and total traffic minutes, which implies that these variables move in opposite directions.

- **Feature correlation to the target**

The below output shows that there is a negative correlation between churn and total traffic minutes, total fee (ETB), SMS usage (birr), SMS fee (ETB), data usage (MB), and data revenue

(ETB). This means that as the values of these variables increase, the likelihood of churn decreases.

CUSTOMER_CATEGORY -0.011708

ADMINISTRATIVE_REGION NaN

ZONE_SUB_CITY -0.005057

GENDER 0.011291

AGE -0.014111

TOTAL_TRAFFIC_MIN -0.172786

TOTAL_FEE_ETB -0.118422

SMS_USAGEBIRR -0.065078

SMS_FEE_ETB -0.013219

DATA_USAGE_MB -0.009479

DATA_REVENUE_ETB -0.054006

churn 1.000000

The correlation between churn and service number, customer category, administrative region, zone sub-city, gender, and age is very weak or non-existent. This means that these variables are not very good predictors of churn.

Overall, the correlation data provided suggests that the variables total traffic minutes, total fee (ETB), SMS usage (birr), SMS fee (ETB), data usage (MB), and data revenue (ETB) are good predictors of churn.

Based on the two analyses of correlation, we can drop two features (service number and administrative region). After dropping these two features, we will be left with a dataset of 10 features that are more likely to be helpful for building efficient models.

4.4 Data normalization

The process of normalization involves rescaling the values of numerical columns in a dataset to a standard scale without distorting variations in the value ranges (Raju et al., 2020a). This is done to make the data easier to analyze and to improve the performance of machine learning algorithms.

We applied minmax normalization, which includes dividing by the range of values (maximum value minus minimum value) after removing the minimum value from each value in the column. By doing this, the column's values will all lie between 0 and 1 (Raju et al., 2020b)

We used min-max normalization because it has some advantages over other normalization techniques such as z-score, Log scaling and clipping, such as simplicity, interpretability, and robustness to outliers and it is simpler to implement because it only requires the minimum and maximum values of the feature. This can be done by setting the normalized value of all continuous features to a specific value. In contrast to normalization approaches that rely on the data's mean and standard deviation, might fluctuate over time while minmax method is less sensitive to outliers or changes in the data over time, as long as the minimum and maximum values remain fixed (Aldalan & Almaleh, n.d.).

Once the data has been normalized, we need to analyze it. This includes calculating the mean and standard deviation of the data. The mean is the average value of the data, and the standard deviation is a measure of how spread out the data is. The distribution of the data shows how the values are clustered around the mean.

OUTPUT:

Mean of normalized data: 0 0.007052

Standard deviation of normalized data: 0 0.083689

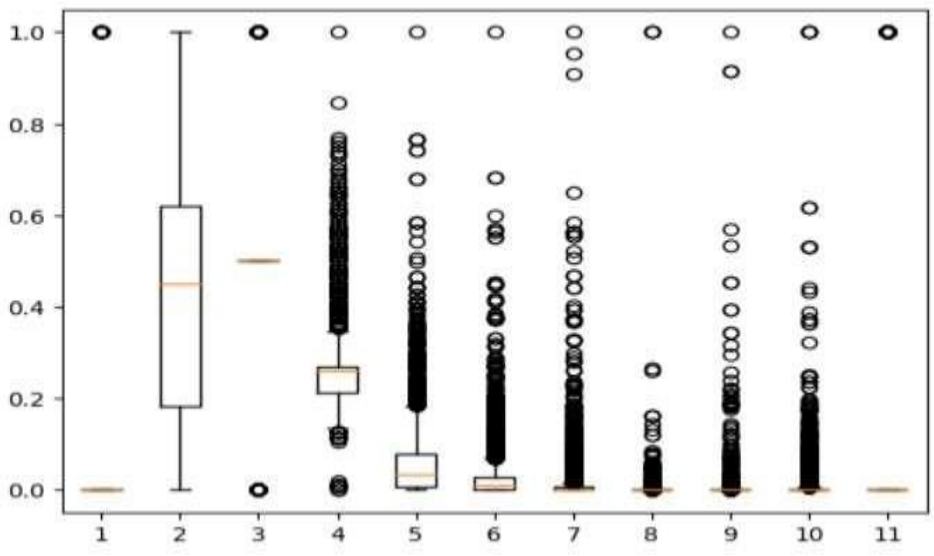

Figure 5: Data normalization

Based on the mean and standard deviation of the normalized data, we can conclude that: The normalized data has some characteristics that indicate its distribution and variability. The mean of the normalized data is 0.007052, which means that the data is slightly skewed to the right but mostly centered on this value. The standard deviation of the normalized data is 0.083689, which

means that the data has a low dispersion around the mean, and most of the values are close to it. There are no outliers in the normalized data, as all of the values fall within 3 standard deviations of the mean, which is a common criterion for identifying outliers.

Overall, the normalized data appears to be evenly distributed and there are no outliers. This suggests that the normalization was successful and that the data is now ready for further analysis.

4.5 Creating Training and Test Data

In this study, a supervised model was trained and evaluated using an 80:20 split ratio between training and testing datasets. train_test_split function from sklearn.model_selection library of pythons was used. Essentially, 80% of the dataset was utilized for training the model, while the remaining 20% was reserved for assessing the model's performance.

To begin, we import the necessary libraries and split the data into training and test data. The two main packages are imported.

```
from sklearn.model_selection import train_test_split
from sklearn.ensemble import RandomForestClassifier
```

After importing the necessary packages, we split the data into training and testing datasets.

```
x_train, x_test, y_train, y_test = train_test_split (
df.drop(labels=[10], axis=1),
   df[10],
  test_size=0.2,
  random_state=0)
```

4.6 Creating Model

To create machine learning models using the proposed algorithms, we used the Python scikit- learn library package. We followed the conventional method of importing the necessary library package for creating the models. Then, we created the models from the respective algorithm class for each model.

4.6.1 Experiment One –Random Forest

In this experiment, we have imported the RandomForestClassifier function fromsklearn.ensemble library of Python and created the random forest classifier model with about 11 hyperparameters. These parameters can be used to fine-tune the random forest model to achievethe desired accuracy and performance.

clf.fit (x_train, y_train) method is used to train a random forest model. It takes two arguments:

- x_train: APandas Data Frame containing the training data features.
- y_train: A Pandas Series containing the training data labels.

The model is made up of a number of decision trees, each of which is trained on a randomly selected subset of the training data. The model then uses the decision trees to make predictions on new test data.

4.6.1.1 Evaluating Model Performance

We have evaluated the performance of the model using the evaluation metrics described in the previous chapter. Evaluation is important to assess the accuracy and generalization ability of the model and to compare the models and choose the best one for the problem. The following output shows the performance found using each evaluation metric.

OUTPUT:

Confusion matrix:

[[1416 14]

[39 91]]

Kappa statistic: 0.7563218390804598

accuracy: 0.966025641025641

precision: 0.8666666666666667

recall: 0.7

f1 score: 0.774468085106383

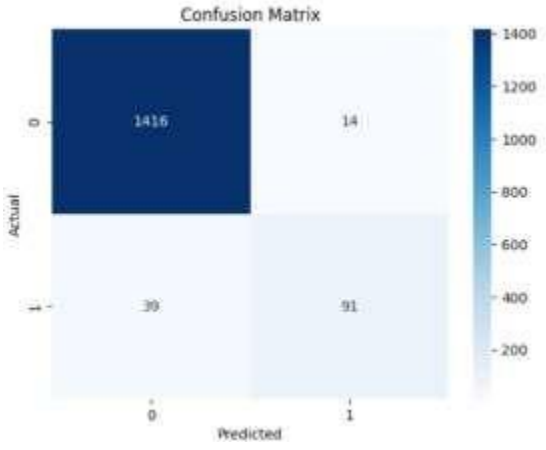

Figure 7: Random Forest Confusion Matrix

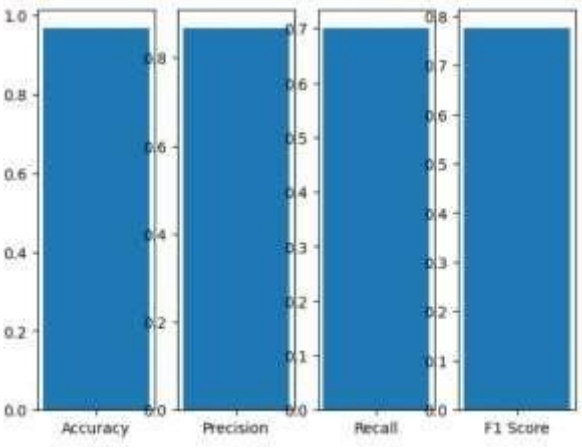

Figure 6: Random Forest Evaluation Bar chart

4.6.1.2 Cross-validation

Cross-validation is a technique that can be used to evaluate the performance of a model on unseen data. It works by dividing the data into a training set and a test set. After the model has been trained on the training set, its effectiveness is assessed on the test set. Each time, different data points are used for the training and test sets in this procedure, which is repeated numerous times. The model's overall performance is then evaluated using the model's average performance throughout all cross-validation iterations (Dutta & Kumar, 2020).

To perform cross-validation, we imported the cross_val_score () function from the sklearn.model_selection module and then used it to evaluate the performance of the model clf on the training data x_train and the target values y_train using 30 folds. The cross_val_score () function returns a list of scores, each of which is the evaluation of the model on one of the folds of the data.

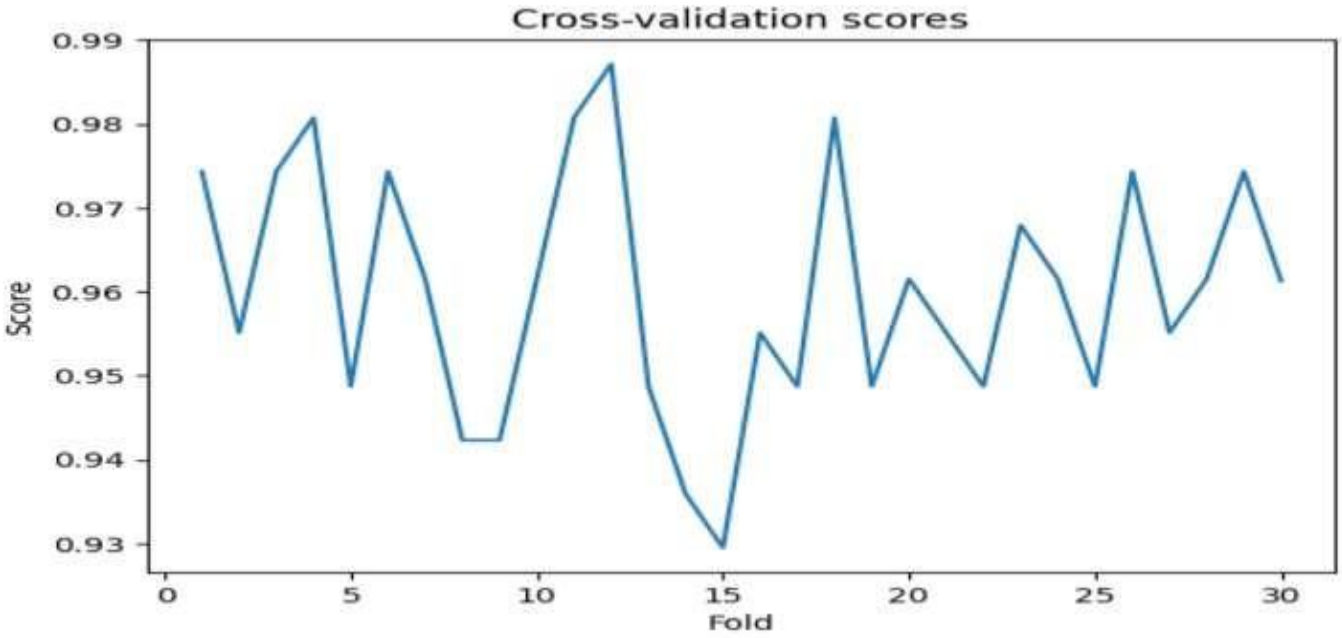

Figure 8: Random Forest Cross validation

By analyzing the provided graph, we can conclude that the model is performing well, with an average accuracy of 96.129032%. There is some variation in the model's performance; however, the overall performance is excellent.

4.6.1.3 Improve the Performance of the Model

The performance of the model can be improved by changing the hyperparameters. Hyperparameters are the settings that control the behavior of the model. By changing the hyperparameters, we can improve the performance of the model. For example, we can increase the number of trees in the forest, or we can use a different sampling method.

Hyperparameter tuning is the process of finding the optimal values for the model's hyperparameters. Some common hyperparameters include the number of trees in a random forest model, the maximum depth of a decision tree, etc. The optimal values for these hyperparameters will vary depending on the dataset and the type of model being used (Paul et al., 2018). To improve the performance of the model, we monitored accuracy, precision, recall, and F1 scores for a range of numbers of trees.

We also find the number of trees that achieves the highest accuracy and this will be repeated for four of the evaluation metrics as follows.

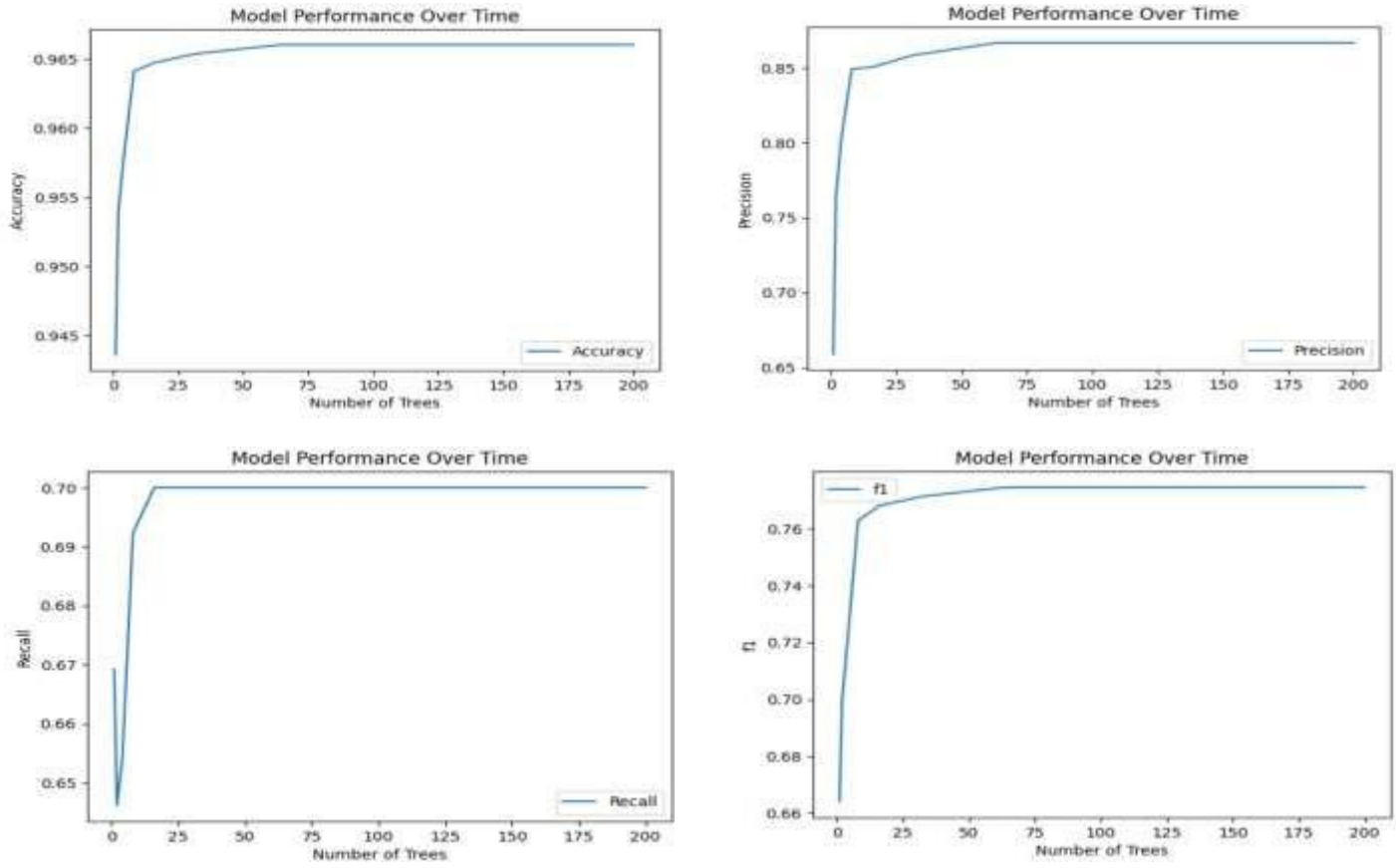

Figure 9: Random Forest model performance over different number of trees

As can be seen from the figures, all four metrics (accuracy, precision, recall, and F1 score) reach their highest values when the number of trees is 64. This suggests that the model is performing best when it is trained on 64 trees.

4.6.2 Experiment Two - K-Nearest Neighbors (KNN)

K-nearest neighbors' classification is an algorithm that classifies a new data point by finding the K most similar data points in the training set and then predicting the label of the new data point based on the labels of the k nearest neighbors (Zhang et, 2018). This subsection discusses the development, optimization, and interpretation of the k nearest neighbors' prediction model. In this experiment, we have imported the KNeighborsClassifier class from sklearn. neighbors and created the k neighbor classifier.

4.6.2.1 Evaluating Model Performance

The output of the KNN model evaluation is also measured using the metrics mentioned before.

OUTPUT:

Confusion matrix:

[[1417 13]

[97 33]]

Kappa statistic: 0.3465346534653466

Accuracy: 0.9294871794871795

Precision: 0.717391304347826

Recall: 0.25384615384615383

F1 score: 0.375

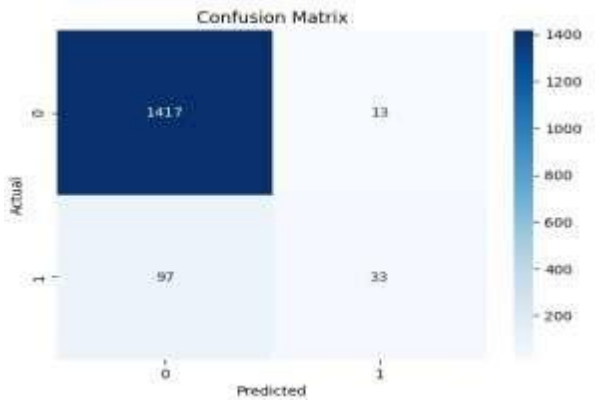

Figure 10: KNN confusion matrix

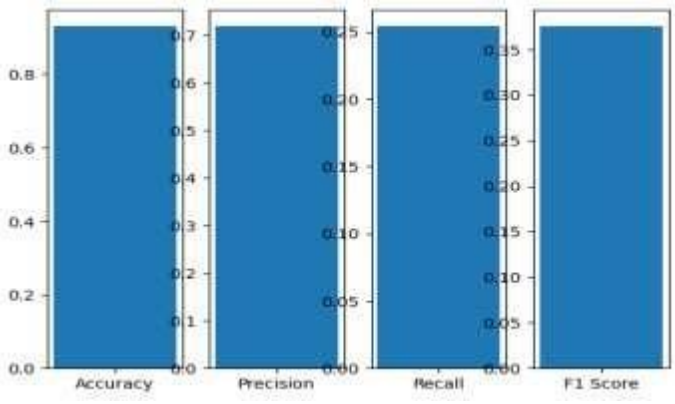

Figure 11: KNN evaluation the four metrics

From the output of the KNN evaluation, we can conclude that the model has a high accuracy of 92.95%. This means that the model correctly classified 92.95% of the test data.

4.6.2.2 Cross-validation

After performing cross validation for the KNN model we could able to conclude about the model's performance.

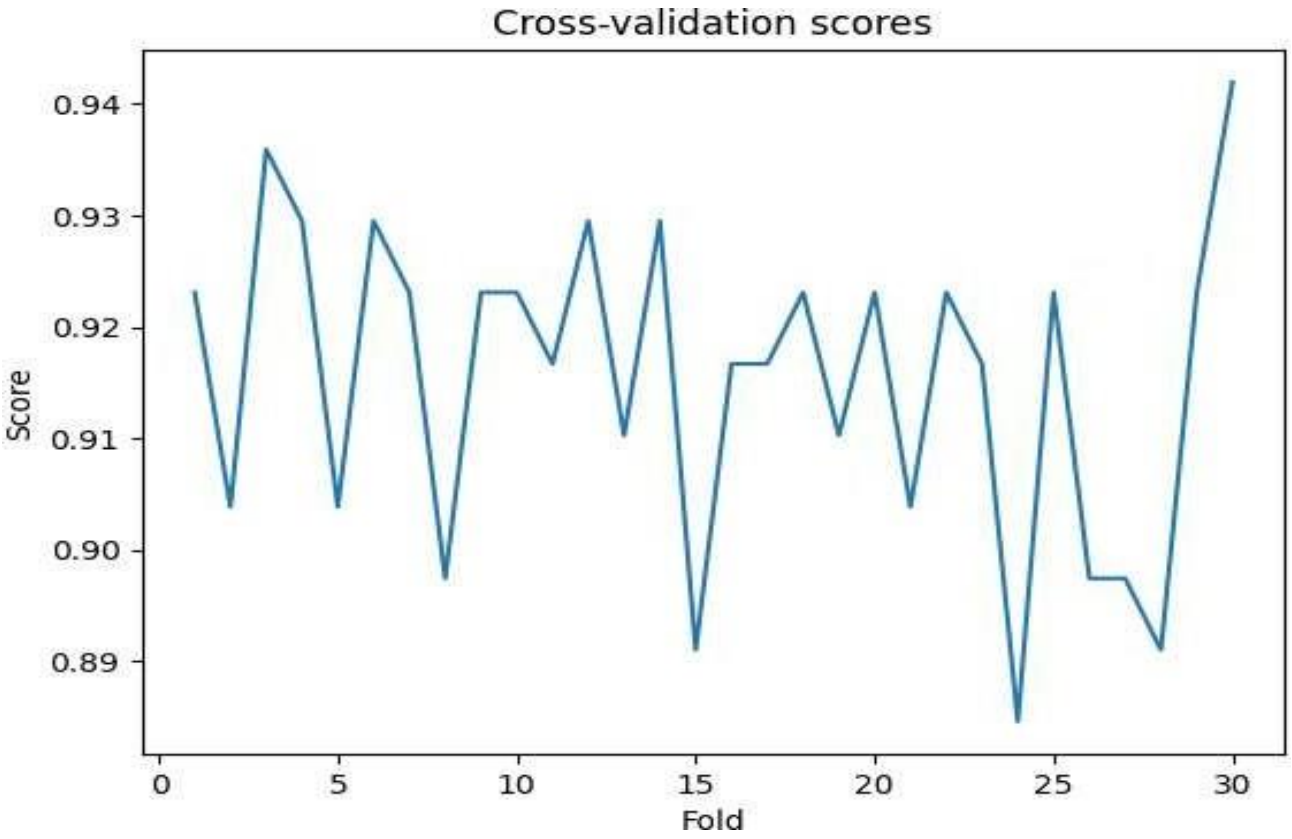

Figure 10: KNN cross validation

The output of cross-validation can be used to draw several conclusions about the model's performance. First, the average accuracy of the model is 0.92094017, which means that the model correctly classified 92.09% of the data in the cross-validation set. Second, the standard deviation of the accuracy is 0.01077027, which indicates that the model's accuracy is consistent, with only a small amount of variation. Third, the distribution of the accuracy is evenly spread around the mean, with no obvious outliers.

In general, a high average accuracy and a low standard deviation are desirable. A distribution of accuracy that is evenly spread around the mean is also desirable. Based on these conclusions, it can be said that the KNN model has a high level of accuracy and consistency. This suggests that the model is likely to perform well on unseen data.

4.6.2.3 Improve the Performance of the Model

To improve the performance of the model, we monitored accuracy, precision, recall, and F1 scores for a range of numbers of k.

We have created a K-nearest neighbors (KNN) model with a different number of neighbors for each value in the neighbors. The model is then trained on the training data and evaluated on the test data. The accuracy of the model has stored in the accuracies list. This will be repeated for four of the metrics and the outputs are displayed in the next diagrams.

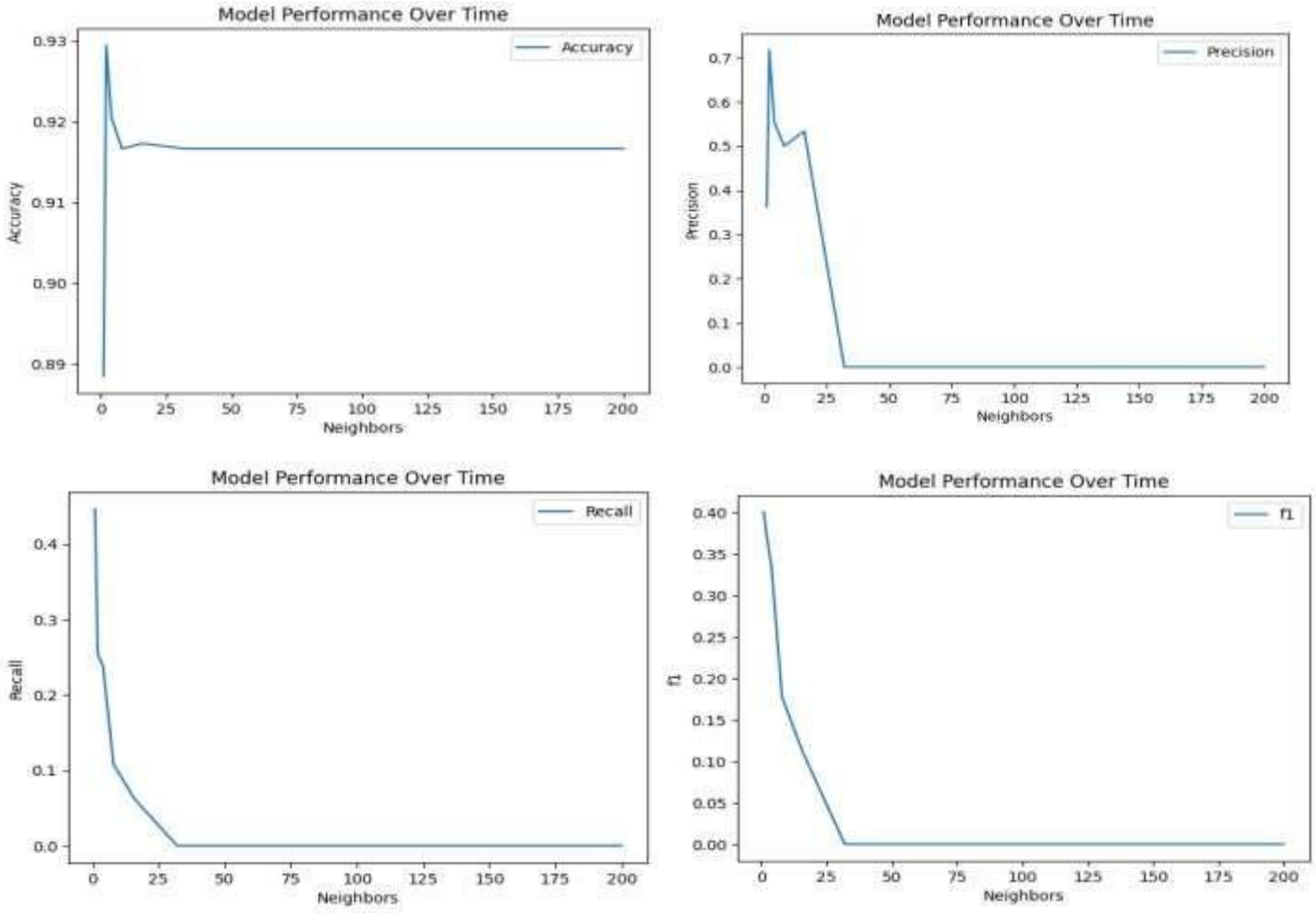

Figure 11: KNN performance over different K value

As shown in the graph, accuracy and precision are maximized when k = 2. Recall and F1 scores, on the other hand, are maximized when k = 1. However, it is recommended that k be set to 2 because accuracy and precision are more important for this predictive model than recall and F1 scores.

To improve the performance of the model, we monitored accuracy, precision, recall, and F1 scores for a range of leaf sizes. We created a K-nearest neighbors (KNN) model with a different number of leaf_size for each value in the n_leaf_size list. The model will then be trained on the training data and evaluated on the test data. The performance of the model is shown in the figures.

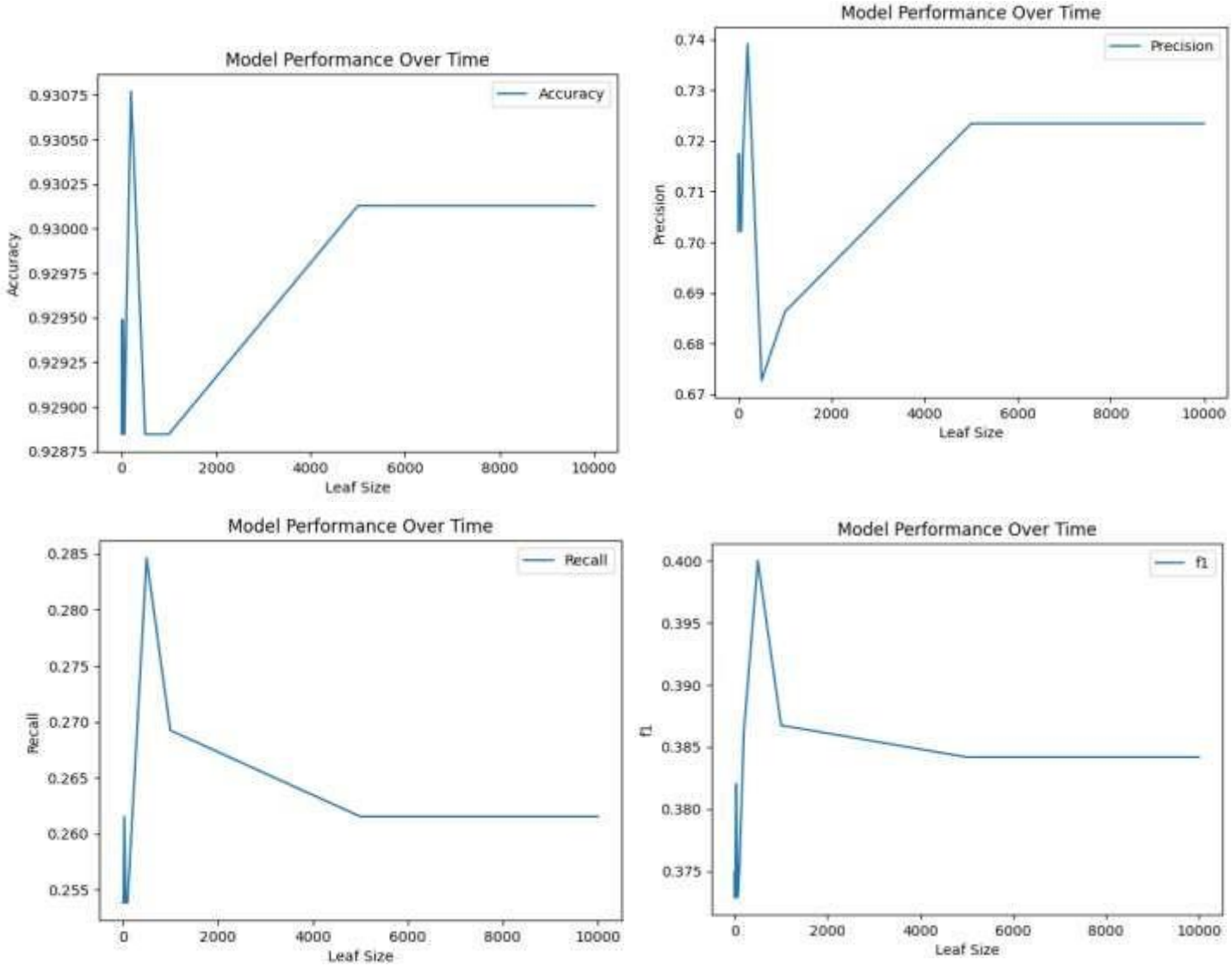

Figure 12: KNN performance over different leaf size

Notably, accuracy and precision are both maximized when the leaf size is 200. Recall and F1 scores, on the other hand, are maximized when the leaf size is 500. Based on this, it is recommended that the leaf size be set to 200 to achieve maximum accuracy and precision.

4.6.3 Experiment Three – Support Vector Machine (SVM)

In this experiment, we have imported the SVC class from sklearn. model selection and created an SVM model. classifier.fit (x_train, y_train), trains the classifier on the training data x_train and y_train. Once the classifier is trained, it is used to make predictions on new data.

4.6.3.1 Evaluating Model Performance

To evaluate the model, we import the necessary libraries and predict the label for the test case. Now, we visualize the output of the model evaluation.

Confusion matrix:

[[1110 27]

[52 59]]

Kappa statistic: 0.5652219929089273

Accuracy: 0.936698717948718

Precision: 0.955249569707401

Recall: 0.9762532981530343

F1 score: 0.9656372335798173

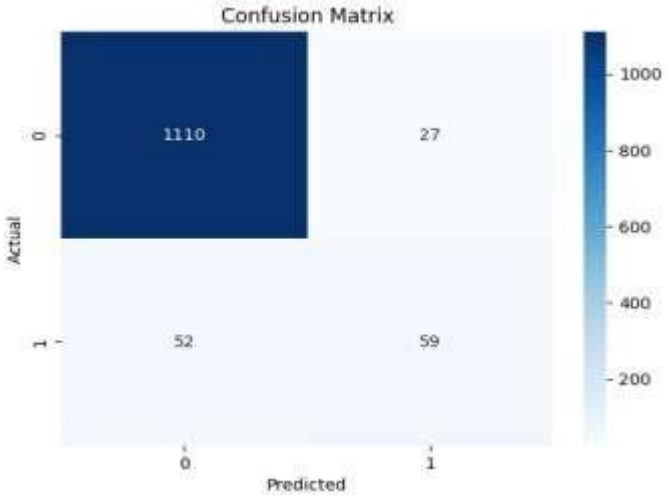

Figure 14: SVM confusion matrix

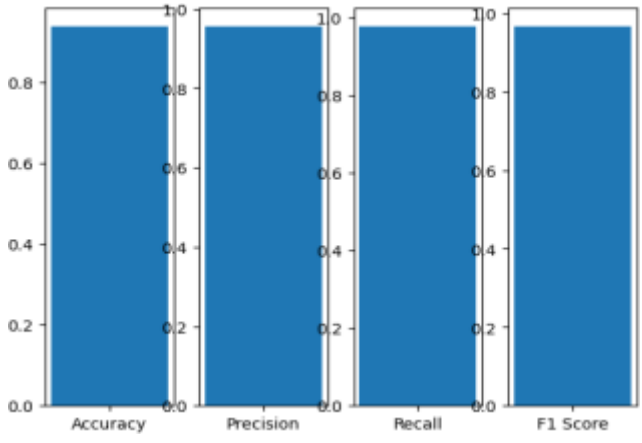

Figure 13: SVM evaluation using the four metrics

The confusion matrix shows that your SVM predictive model correctly classified 1110 instances of the positive class and 59 instances of the negative class. It incorrectly classified 27 instances of the positive class as negative and 52 instances of the negative class as positive. A kappa of 0.565 indicates that there is a moderate agreement between the two raters.

An accuracy of 0.937 and a precision of 0.955 indicates that your SVM predictive model is very accurate and also precise. In another way, a recall of 0.976 indicates that your SVM predictive model is very good at identifying positive instances. An F1 score of 0.966 indicates that your SVM predictive model is very good at both precision and recall. Overall, the output of yourSVM predictive model is very good. It is very accurate, precise, and has a high recall.

4.6.3.2 Cross-validation

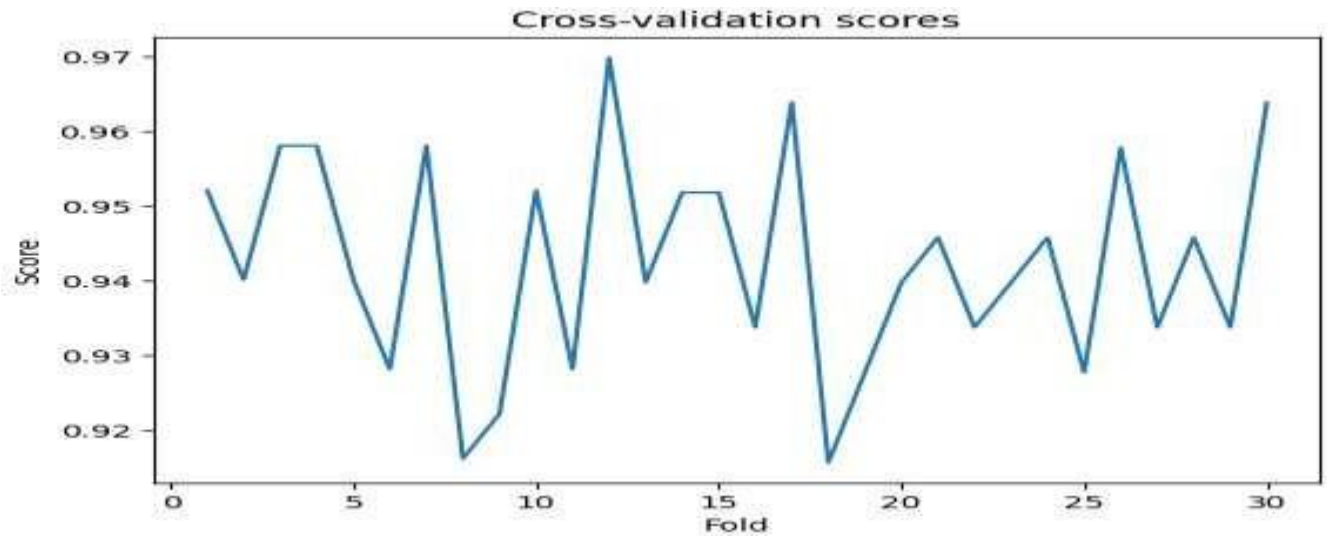

Figure 15: SVM cross validation

The cross-validation accuracy scores are all above 90%, which indicates that the model is performing well. However, there is some variation in the scores, which suggests that the model may not be as robust as it could be. For example, the lowest score is 0.915, which is below 90%. This suggests that the model may not be able to accurately predict all of the data points in the dataset. Overall, the cross-validation accuracy scores suggest that the model is performing well.

4.6.3.3 Improve the Performance of the Model

It loops over a list of cost values, and for each cost value, it creates an SVC model, trains it on the training data, makes predictions on the test data, calculates the F1 score of the predictions, and appends the F1 score to a list of accuracies. After looping over all of the cost values, the function plots a graph of the accuracies versus the cost values. The graph allows you to see how the accuracy of the model changes as the cost value changes. The cost value that produces the highest accuracy is also identified. This will be repeated for four of the metrics and the outputs are displayed in the next diagrams.

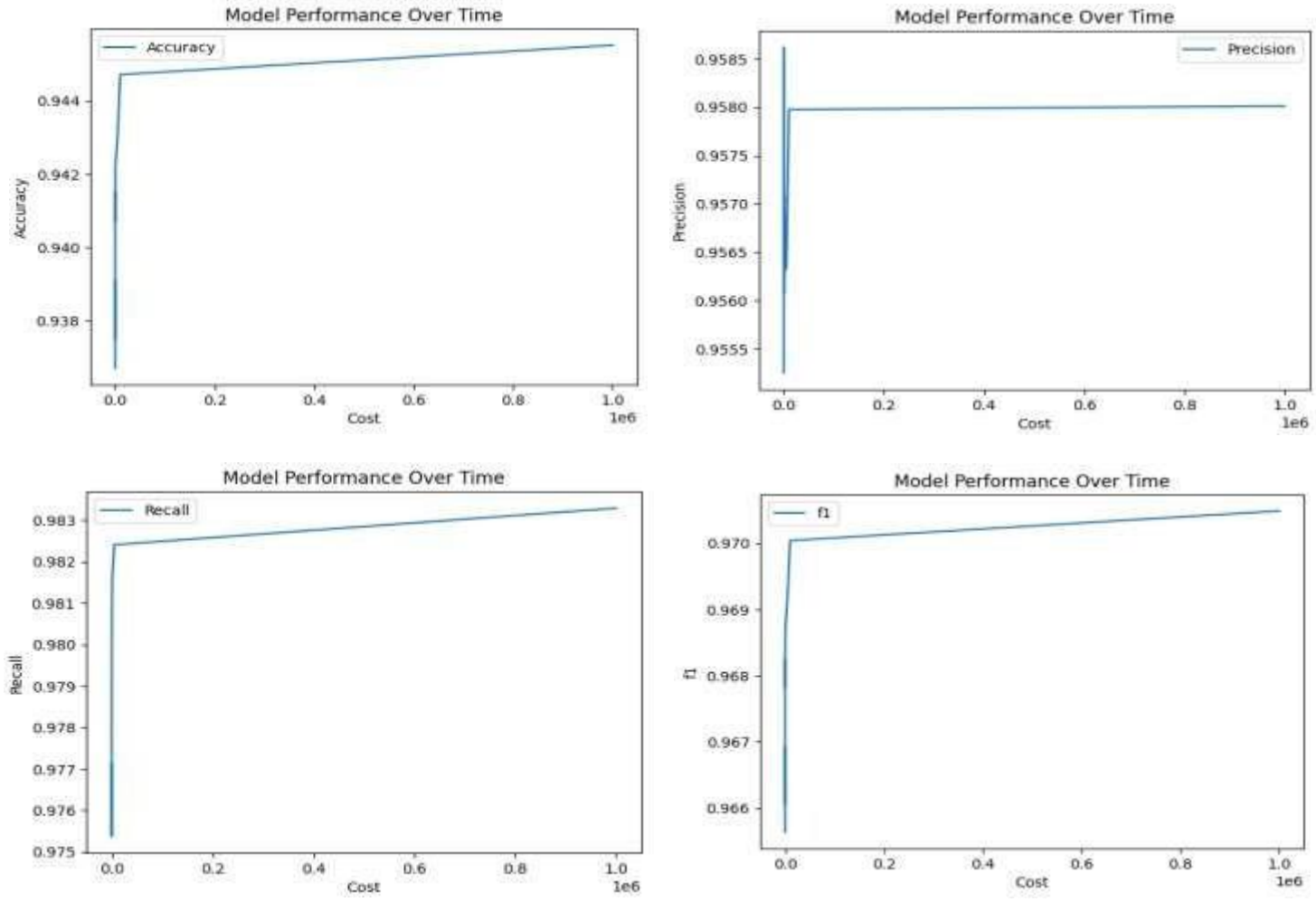

Figure 16: SVM performance improved

As shown in the graph, accuracy, recall, and f1 scores are maximized when the cost (C) is at its maximum. But precision, on the other hand, is maximized to 95.862 when C = 32 and stays at 95.80. Therefore, the overall performance of the model is maximized when C is at its maximum.

4.6.4 Experiment four – Logistic Regression

In this experiment, we have imported the LogisticRegression class from sklearn. linear_model and created a Logistic Regression model with 4 hyperparameters. It is created with L2 regularization, which penalizes the size of the coefficients to prevent overfitting, and an inverse regularization strength of 1.0, which balances the complexity and simplicity of the model. The model also uses the LBFGS solver, which is a gradient-based solver that is faster than other solvers for LogisticRegression and adds a constant term to the model, which is helpful for classification problems. The model is then fit to the training data using the fit () method, which takes the training data and the target variable as arguments.

4.6.4.1 Evaluating Model Performance

To evaluate the model, we imported the necessary libraries and predict the label for the test case. The output of the model evaluation visualized in the figures below.

Confusion matrix:

[[1079 79]

[83 7]]

Kappa statistic: 0.009756670976842519

Accuracy: 0.8701923076923077

Precision: 0.9285714285714286

Recall: 0.9317789291882557

F1 score: 0.9301724137931036

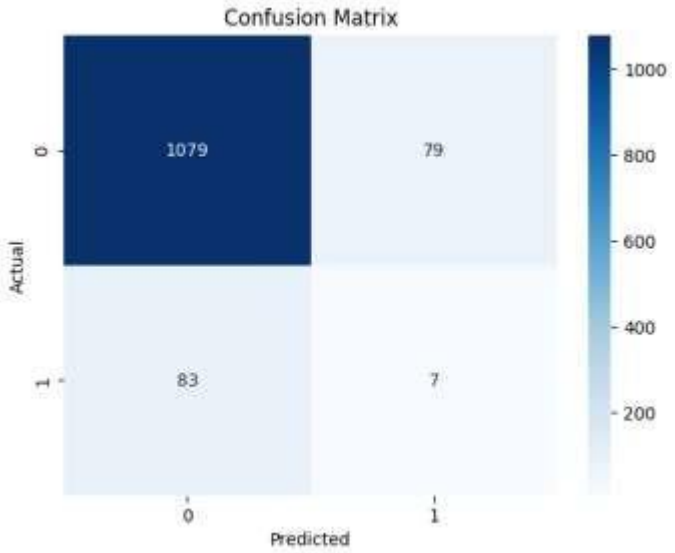

Figure 19: Logistic regression confusion matrix

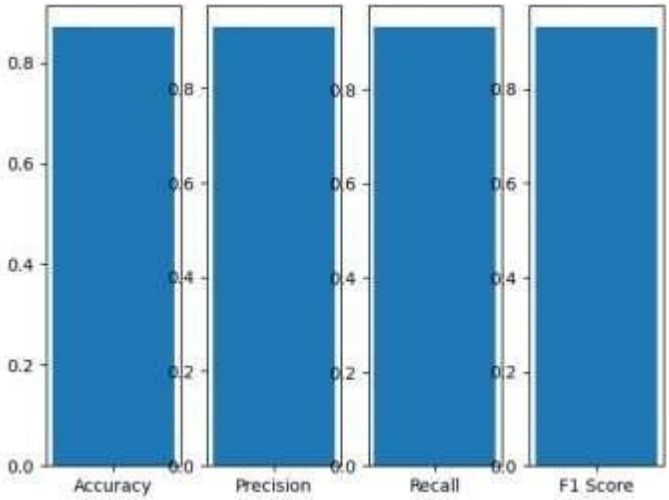

Figure 20: Logistic regression performance using the four metrics

From the confusion matrix, we can see that the model correctly classified 1079 instances of the positive class and 79 instances of the negative class. The model incorrectly classified 83 instances of the positive class as negative and 7 instances of the negative class as positive. The kappa statistic is a measure of how well the model performs compared to random chance. A kappa statistic of 0 indicates that the model performs no better than random chance, while akappa statistic of 1 indicates that the model performs perfectly. The kappa statistic for this model is 0.009756670976842519, which indicates that the model performs slightly better than random chance.

The accuracy for this model is 0.8701923076923077, which indicates that the model correctly classifies 87.02% of instances. The precision for this model is 0.9285714285714286, which indicates that the model correctly identifies 92.86% of instances that it predicts to be positive.

The recall for this model is 0.9317789291882557, which indicates that the model correctly identifies 93.18% of positive instances. The f1 score is a measure of the model's accuracy and precision. The f1 score for this model is 0.9301724137931036, which indicates that the model has a good balance of accuracy and precision. Overall, the model performs well in terms of accuracy and precision. However, the kappa score suggests that the model does not perform well

overall. This is likely because the model is not able to distinguish between the positive and negative classes very well.

4.6.4.2 Cross-validation

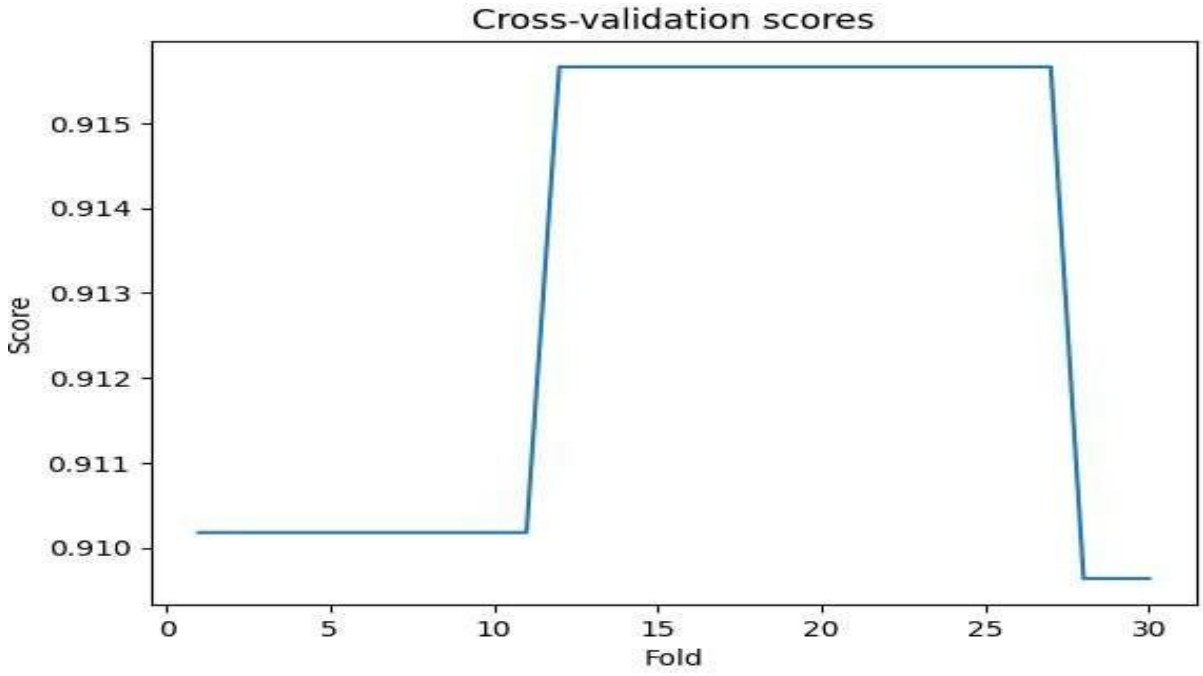

Figure 17: Logistic Regression Cross validation

The cross-validation evaluation of the model shows that the model performs consistently well, with accuracy scores ranging from 90.96% to 91.56%. This suggests that the model is not overfitting the training data and can generalize to unseen data. The average accuracy score of 91.26% is also within an acceptable range.

4.6.4.3 Improve the Performance of the Model

To improve the performance of the model, we monitored the accuracy, precision, recall, and F1 scores for a range of values of C. We first imported the following metrics libraries from sklearn. metrics: confusion_matrix, accuracy_score, precision_score, recall_score, and f1_score. We also imported the LogisticRegression class from sklearn. linear_model.

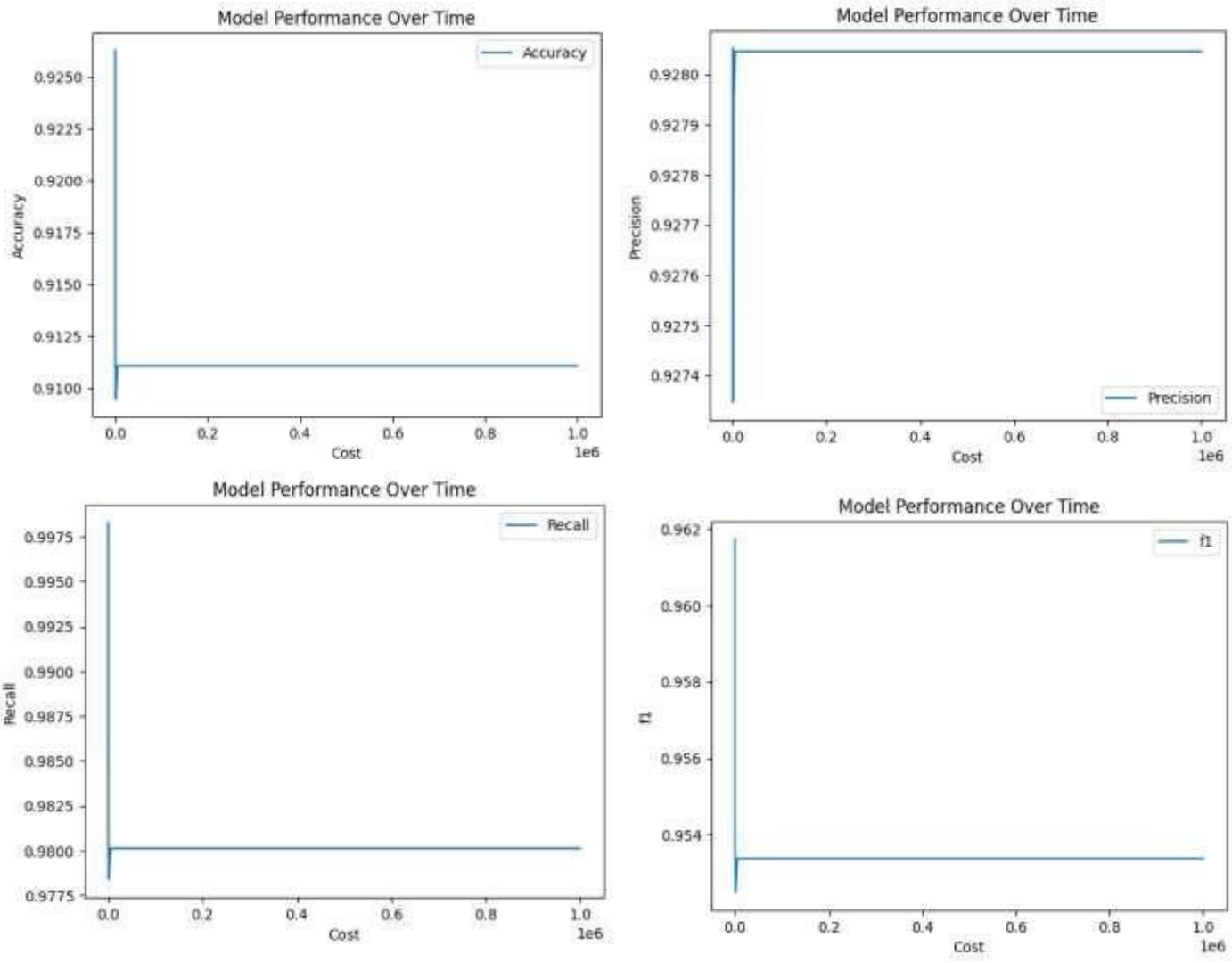

Figure 22: Logistic regression with different cost value

As shown in the graph, accuracy, recall and f1 scores are maximized when C = 1. But precision, on the other hand, are maximized when C = 1. Therefore, the overall performance of the model is maximized when C = 1.

4.7 Discussion

There are 7205 customer records total in the data set, each with a unique set of 15 attributes. Demographic details (gender, age, administrative region, zone, and sub-city), as well as usage details (total traffic minutes, total charge in Ethiopian Birr, SMS usage birr, SMS fee in Ethiopian Birr, data usage MB, data income in Ethiopian Birr, and voice usage minutes), are included in the attributes. Churn is a dependent variable that shows whether a client has abandoned Ethio Telecom. Following data preparation, the dataset was reduced to 6,239 records and 11 attributes. The dataset declined to 6,239 records and 11 attributes after data preprocessing.

We used min-max normalization to rescale the values of numeric columns in the dataset to a common scale, without distorting differences in the ranges of values. This was done to simplify data analysis and boost the effectiveness of machine learning algorithms. Min-max normalization entails dividing by the range of values (maximum value minus minimum value) after removing the minimum value from each value in the column. By doing this, the column's values will all lie between 0 and 1. There were no outliers and the normalized data seemed to be uniformly distributed, indicating that the normalization was successful and the data was ready for further analysis.

We conducted a customer churn prediction model for Ethiopian Telecommunication company using various features related to customer behavior and profile. Our goal was to identify the most important factors that influence customer retention and loyalty. We used information gain and correlation as the criteria to select the relevant features for our model. Information gain measures how much a feature reduces the entropy or uncertainty of the target variable, while correlation measures how much a feature is linearly related to the target variable.

We found that the features TOTAL_TRAFIC_MIN, TOTAL_FEE_ETB, SMS_USAGE_BIRR, SMS_FEE_ETB, DATA_USAGE_MB, and DATA_REVENUE_ETB had the highest information gain and correlation with the target variable, which indicates customer churn. These features represent the total amount of traffic, total fee, usage, and revenue of voice, SMS, and data services for each customer. These features are important for customer churn prediction because they reflect the customer's satisfaction, value, and engagement with the telecom

services. Customers who have low or declining traffic, fee, usage, or revenue are more likely to churn than customers who have high or increasing traffic, fee, usage, or revenue.

On the other hand, we found that the features CATEGORY, VOICE_USAGE_MIN, ADMINISTRATIVE_REGION, and SERVICE_NUMBER had the lowest information gain and correlation with the target variable. These features represent the customer's category (prepaid or postpaid), voice usage in minutes, administrative region (city or rural), and service number (phone number). These features are not important for customer churn prediction because they do not capture the customer's satisfaction, value, or engagement with the telecom services. Customers who have different categories, voice usage, regions, or service numbers are not significantly different in their likelihood to churn.

Based on these findings, we decided to remove the features CATEGORY, VOICE_USAGE_MIN, ADMINISTRATIVE_REGION, and SERVICE_NUMBER from our
dataset. By removing these irrelevant features, we improved the performance of our model by reducing noise and dimensionality. We also improved the interpretability and explainability of our model by focusing on the most relevant features for customer churn prediction.

When comparing the models, it was discovered that the Random Forest algorithm had the best overall results. Support Vector Machine (SVM), logistic regression, and K-nearest neighbors (KNN) models all fared well in terms of accuracy and precision at the prediction level, though. The following measures, including accuracy, precision, recall, F1 score, confusion matrix, and kappa statistics, were used to assess and compare these machine learning algorithms. The results from all four models are summarized in the table below.

Table 4: Summary of result comparison

Evaluation Criteria	Random Forest	KNN	SVM	Logistic Regression
Accuracy	96.60%	92.88%	94.55%	87.25%
Precision	86.67%	70.21%	95.80%	92.80%
Recall	70%	25.38%	98.32%	93.52%
F1_score	77.44%	37.28%	97.04	93.16%

Kappa Stat	75.63%	34.38%	61.70%	20.02%
Positive-Positive	1416	1417	1110	1079
Positive-Negative	14	13	27	79
Negative-Positive	39	97	52	83
Negative-Negative	91	33	59	7
Average accuracy	96%	91.55%	94.25%	91.3

The random forest model achieved an accuracy of 96.60%, a precision of 86.67%, a recall of 70%, an F1-score of 77.44%, a kappa stat of 75.63%, and an average accuracy of 96%. The model correctly identified 1416 customers that will not churn and 91 customers who will churn. The model incorrectly identified 14 customers who would not churn as churners and 39 customers who would churn as non-churners.

The KNN model achieved an accuracy of 92.88%, a precision of 70.21%, a recall of 25.38%, an F1-score of 37.28%, a kappa stat of 34.38%, and a cross-validation average accuracy of 91.55%. The model correctly identified 1417 customers who will not churn and 33 customers who will churn. The model incorrectly identified 13 customers who would not churn as churners and 97 customers who would churn as non-churners.

The SVM model achieved an accuracy of 94.55%, a precision of 95.80%, a recall of 98.32%, an F1-score of 97.04, a kappa stat of 61.70%, and a cross-validation average accuracy of 94.25%. The model correctly identified 1110 customers who will not churn and 59 customers who will churn. The model incorrectly identified 52 customers who would not churn as churners and 59 customers who would churn as non-churners.

The logistic regression model achieved an accuracy of 87.25%, a precision of 92.80%, a recall of 93.52%, an F1-score of 93.16%, a kappa stat of 20%, and a cross-validation average accuracy of 91.3%. The model correctly identified 1079 customers who will not churn and 7 customers who will churn. The model incorrectly identified 79 customers who would not churn as churners and 83 customers who would churn as non-churners.

The random forest model performed the best, followed by the SVM model, the KNN model, and the logistic regression model. The random forest model had the highest accuracy, precision, recall, F1-score, and kappa stat. The SVM model had the second-highest accuracy, precision, recall, F1-score, and kappa stat. The KNN model had the third highest accuracy, precision, recall,

F1-score, and kappa stat. The logistic regression model had the lowest accuracy, precision, recall, F1-score, and kappa stat.

The random forest model was able to correctly identify more customers who would churn than the other models. The SVM model was able to correctly identify more customers who would not churn than the other models. The KNN model was able to correctly identify more customers whowould churn and more customers who would not churn than the logistic regression model.

The random forest model was the most accurate, but it was also the most computationally expensive model. The SVM model was less computationally expensive than the random forest model, but it was not as accurate. The KNN model was the least computationally expensive model, but it was also the least accurate.

The logistic regression model was the simplest model to train and interpret, but it was also the least accurate model. Overall, the random forest model was the best-performing model, followed by the SVM model, the KNN model, and the logistic regression model.

Hyperparameters play a crucial role in improving the accuracy of machine learning models such as random forests, KNN, SVM, and logistic regression. For random forests, tuning the hyperparameter of the number of trees helps prevent overfitting and improve accuracy. For KNN, tuning the value of k helps control the bias-variance tradeoff and improves performance. Similarly, in SVM and logistic regression, tuning the cost value helps control the tradeoffbetween bias and variance and improves performance. In conclusion, using hyperparameters allows us to optimize and fine-tune models to improve their accuracy, making them moreeffective for predicting customer churn.

CONCLUSION AND RECOMMENDATION

Overview

This chapter synthesizes the key findings of the study and presents them in two main sections: conclusion and recommendation. The conclusion section draws on the main implications and contributions of the research, while the recommendation section suggests directions for future work based on the gaps and limitations identified in the study.

Conclusion

In this study, the experimental data were obtained from Ethio-telecom. The data was structured in an Excel sheet and the total dataset dimension was 7205. The dataset comprises extraneous attributes and numerous null entries. The dataset undergoes preprocessing and 6239 instances are rendered for the experimental analysis.

The construction of classification and prediction models was conducted using the RandomForest, KNN, SVM, and Logistic Regression algorithms. Jupiter Notebook, google Colab, and Python programming were utilized as tools to simulate every aspect of the experiment. Various performance metrics were employed to evaluate and contrast the efficiency of the models. When it comes to accurately classifying instances, the three models (Random Forest, SVM, and KNN) produced impressive results, with Random Forest having an accuracy value of 96.6%, SVM having an accuracy value of 94.55%, and KNN having accuracy values of 92.88% and Logistic regression having an accuracy of 87.25%.

All of the research questions are answered as follows:

- The first research question is ‒Which attributes are more significant to predict customers churning at Ethio-telecom?‖ The following variables are determined to be effective and relevant predictors for predicting customer churn in Ethiopian Telecom. They are: TOTAL_TRAFFIC_MIN, TOTAL_FEE_ETB, SMS_USAGE_BIRR, SMS_FEE_ETB, DATA_USAGE_MB, and DATA_REVENUE_ETB.
- According to the research question two: ‒To what extent do the proposed models perform in customer churn prediction?‖ Random forest performed an accuracy of 96.6%, SVM 94.55%, KNN 92.88 %, and LR 87.25% in making decisions about whether a customer churns or not.

- As for the third study question –Which classification algorithm is more suitable for constructing Ethio-telecom customer churn prediction model?‖ With an accuracy of 96.6% and precision of 86.67%, random forest was found as the most effective customer churn prediction model.

The results of the studies indicate that telecoms can utilize machine learning algorithmic techniques to create customer churn predictive models that can accurately forecast the likelihood of customers leaving their service. These models can help the telecoms to identify the customers who are at risk of churning and take appropriate actions to retain them. The models can also provide insights into the factors that influence customer churn and satisfaction.

Therefore, these machine learning algorithms can be used in the telecom industry to calculate the likelihood of client attrition. It is anticipated that by putting this plan into practice, Ethio-telecom will be able to maximize its return on investment and perform well in terms of identifying and forecasting client attrition. This would help them to reduce the loss of revenue and customers, and increase their customer loyalty and satisfaction.

Recommendations

Based on the current state of research and the potential benefits of machine learning algorithms for predicting customer attrition in the telecom industry, it is recommended that future work should focus on expanding the dataset size and incorporating multiple features. Increasing the dataset size can help to improve the accuracy and robustness of the machine learning models, as well as enhance the generalizability of the results. Additionally, incorporating multiple features can provide a more comprehensive understanding of the factors that influence customer attrition, enabling telecom companies to make more informed decisions and develop more effective strategies for retaining customers.

Possible future work for this research is to improve the data collection method by conducting interviews with experts in the telecom industry, especially because the domain is new and there is limited literature and data available. This would help to identify the most effective features that determine the churn possibility of customers, as well as to fill the gaps and address the challenges in the current data collection and analysis. Interviews with experts could also provide insights into the trends and innovations in the telecom industry, as well as the expectations and preferences of customers. By incorporating expert knowledge into the data collection and feature

engineering process, the research could achieve more comprehensive and relevant results, as well as more practical and feasible recommendations for telecom companies.

Furthermore, it is suggested that future researchers explore the use of advanced machine learning techniques, such as deep learning and ensemble learning, to further improve the predictive capabilities of the models. This can be particularly beneficial for handling large and complex datasets, and for capturing and analyzing non-linear relationships between variables. Most of the existing studies use customer data features that are related to demographic, product usage, and revenue aspects, which may not capture the full spectrum of customer behavior and preferences. Therefore, there is a need for more studies that use data features that are related to customer feedback, social network, sentiment analysis, or other aspects that can provide more insights into customer churn and retention.

Overall, by incorporating these recommendations into future research, it is expected that the telecom industry can make significant strides in predicting and addressing customer attrition, ultimately leading to improved customer retention and business performance.

***

Special Acknowledgment

This research was sponsored by Adama Science and Technology University with a grant number

ASTU/SM-R/873/23

Adama, Ethiopia

***

References

Adame, B. O. (2021). The Ethiopian telecom industry: gaps and recommendations towards meaningful connectivity and a thriving digital ecosystem. *Heliyon*, *7*(10). https://doi.org/10.1016/j.heliyon.2021.e08146

Ahmad, A. K., Jafar, A., & Aljoumaa, K. (2019). Customer churn prediction in telecom using machine learning in big data platform. *Journal of Big Data*, *6*(1). https://doi.org/10.1186/s40537-019-0191-6

Ahn, J., Hwang, J., Kim, D., Choi, H., & Kang, S. (2020). A Survey on Churn Analysis in Various Business Domains. *IEEE Access*, *8*, 220816–220839. https://doi.org/10.1109/ACCESS.2020.3042657

Aldalan, A. M., & Almaleh, A. (n.d.). *Customer Churn Prediction Using Four Machine Learning Algorithms Integrating Feature Selection and Normalization in the Telecom Sector*.

Al-Mashraie, M., Chung, S. H., & Jeon, H. W. (2020). Customer switching behavior analysis in the telecommunication industry via push-pull-mooring framework: A machine learning approach. *Computers and Industrial Engineering*, *144*. https://doi.org/10.1016/j.cie.2020.106476

Amity University. Dubai Campus, Institute of Electrical and Electronics Engineers. UAE Section, & Institute of Electrical and Electronics Engineers. (n.d.). *Abstract proceedings of International Conference on Computation, Automation and Knowledge Management (ICCAKM-2020) : 9th-10th January 2020*.

Azizah, N., Riza, L. S., & Wihardi, Y. (2019). Implementation of random forest algorithm with parallel computing in R. *Journal of Physics: Conference Series*, *1280*(2). https://doi.org/10.1088/1742-6596/1280/2/022028

Bell, J. (2022). What is machine learning? In *Machine Learning and the City: Applications in Architecture and Urban Design* (pp. 209–216). Wiley Blackwell. https://doi.org/10.1007/978-3-319-18305-3_1

Cervantes, J., Garcia-Lamont, F., Rodríguez-Mazahua, L., & Lopez, A. (2020). A comprehensive survey on support vector machine classification: Applications, challenges and trends. *Neurocomputing*, *408*, 189–215. https://doi.org/10.1016/j.neucom.2019.10.118

Chen, W. K., Nalluri, V., Ma, S., Lin, M. M., & Lin, C. T. (2021). An exploration of the critical risk factors in sustainable telecom services: An analysis of indian telecom industries. *Sustainability (Switzerland)*, *13*(2), 1–22. https://doi.org/10.3390/su13020445

Dj Novakovi, J., Veljovi, A., Ili, S. S., Zeljko Papi, ˇ, & Tomovi, M. (2017). Evaluation of Classification Models in Machine Learning. In *Theory and Applications of Mathematics & Computer Science* (Vol. 7, Issue 1).

Dutta, S., & Kumar Bandyopadhyay, S. (2020). Employee attrition prediction using neural network cross validation method. *International Journal of Commerce and Management Research*. www.managejournal.com

Eskinder, M. (2021). *THE EFFECT OF CUSTOMER RELATIONSHIP MANAGEMENT ON CUSTOMER SATISFACTION IN THE CASE OF ETHIO TELECOM, ADDIS ABABA ADDIS ABABA*.

Gebreegziabher, B., & To, S. (2022). *BANK CUSTOMER CHURN PREDICTION MODEL: THE CASE OF COMMERCIAL BANK OF ETHIOPIA*.

Jaber, K. M., Institute of Electrical and Electronics Engineers. Jordan Section, Institute of Electrical and Electronics Engineers. Region 8, & Institute of Electrical and Electronics Engineers. (n.d.). *2019 IEEE Jordan International Joint Conference on Electrical Engineering and Information Technology (JEEIT) : proceedings : April 9-11, 2019, Le Royal Amman Hotel, Jordan*.

Jain, H., Khunteta, A., & Srivastava, S. (2020). Churn Prediction in Telecommunication using Logistic Regression and Logit Boost. *Procedia Computer Science*, *167*, 101–112. https://doi.org/10.1016/j.procs.2020.03.187

Lubis, A. R., Lubis, M., & Al-Khowarizmi. (2020). Optimization of distance formula in k-nearest neighbor method. *Bulletin of Electrical Engineering and Informatics*, *9*(1), 326–338. https://doi.org/10.11591/eei.v9i1.1464

Mu, Y., Liu, X., & Wang, L. (2018). A Pearson's correlation coefficient-based decision tree and its parallel implementation. *Information Sciences*, *435*, 40–58. https://doi.org/10.1016/j.ins.2017.12.059

Nur Amir Sjarif, N., Rusydi Mohd Yusof, M., Hooi-Ten Wong, D., Ya, S., Ibrahim, R., Zamri Osman, M., & Teknologi Malaysia Kuala Lumpur Jalan Sultan Yahya Petra, U. (2019a). A Customer Churn Prediction using Pearson Correlation Function and K Nearest Neighbor Algorithm for Telecommunication Industry. In *Int. J. Advance Soft Compu. Appl* (Vol. 11, Issue 2).

Nur Amir Sjarif, N., Rusydi Mohd Yusof, M., Hooi-Ten Wong, D., Ya, S., Ibrahim, R., Zamri Osman, M., & Teknologi Malaysia Kuala Lumpur Jalan Sultan Yahya Petra, U. (2019b). A Customer Churn Prediction using Pearson Correlation Function and K Nearest Neighbor Algorithm for Telecommunication Industry. In *Int. J. Advance Soft Compu. Appl* (Vol. 11, Issue 2).

Paul, A., Mukherjee, D. P., Das, P., Gangopadhyay, A., Chintha, A. R., & Kundu, S. (2018). Improved Random Forest for Classification. *IEEE Transactions on Image Processing*, *27*(8), 4012–4024. https://doi.org/10.1109/TIP.2018.2834830

Rahman, M., & Kumar, V. (2020). Machine Learning Based Customer Churn Prediction in Banking. *Proceedings of the 4th International Conference on Electronics, Communication and Aerospace Technology, ICECA 2020*, 1196–1201. https://doi.org/10.1109/ICECA49313.2020.9297529

Raju, V. N. G., Lakshmi, K. P., Jain, V. M., Kalidindi, A., & Padma, V. (2020a). Study the Influence of Normalization/Transformation process on the Accuracy of Supervised Classification. *Proceedings of the 3rd International Conference on Smart Systems and Inventive Technology, ICSSIT 2020*, 729–735. https://doi.org/10.1109/ICSSIT48917.2020.9214160

Raju, V. N. G., Lakshmi, K. P., Jain, V. M., Kalidindi, A., & Padma, V. (2020b). Study the Influence of Normalization/Transformation process on the Accuracy of Supervised Classification. *Proceedings of the 3rd International Conference on Smart Systems and Inventive Technology, ICSSIT 2020*, 729–735. https://doi.org/10.1109/ICSSIT48917.2020.9214160

Rodrigues Chagas, B. N., Nogueira Viana, J. A., Reinhold, O., Lobato, F., Jacob, A. F. L., & Alt, R. (2019). Current Applications of Machine Learning Techniques in CRM: A Literature Review and Practical Implications. *Proceedings - 2018 IEEE/WIC/ACM International Conference on Web Intelligence, WI 2018*, 452–458. https://doi.org/10.1109/WI.2018.00-53

Shaik, A. B., & Srinivasan, S. (2019). A brief survey on random forest ensembles in classification model. In *Lecture Notes in Networks and Systems* (Vol. 56, pp. 253–260). Springer. https://doi.org/10.1007/978-981-13-2354-6_27

Shumaly, S., Neysaryan, P., & Guo, Y. (2020). Handling Class Imbalance in Customer Churn Prediction in Telecom Sector Using Sampling Techniques, Bagging and Boosting Trees. *2020 10h International Conference on Computer and Knowledge Engineering, ICCKE 2020*, 82–87. https://doi.org/10.1109/ICCKE50421.2020.9303698

Tadesse, M. (2013). *FACULTY OF BUSINESS DEPARTMENT OF MARKETING MANAGEMENT THE PRACTICE OF RELATIONSHIP MARKETING AND CUSTOMER LOYALTY WITH REFERENCE TO ethio telecom*.

Ullah, I., Raza, B., Malik, A. K., Imran, M., Islam, S. U., & Kim, S. W. (2019). A Churn Prediction Model Using Random Forest: Analysis of Machine Learning Techniques for Churn Prediction and Factor Identification in Telecom Sector. *IEEE Access*, *7*, 60134–60149. https://doi.org/10.1109/ACCESS.2019.2914999

Vafeiadis, T., Diamantaras, K. I., Sarigiannidis, G., & Chatzisavvas, K. C. (2015). A comparison of machine learning techniques for customer churn prediction. *Simulation Modelling Practice and Theory*, *55*, 1–9. https://doi.org/10.1016/j.simpat.2015.03.003

Vujović, Ž. (2021). Classification Model Evaluation Metrics. *International Journal of Advanced Computer Science and Applications*, *12*(6), 599–606. https://doi.org/10.14569/IJACSA.2021.0120670

Wang, L. (2019). Research and Implementation of Machine Learning Classifier Based on KNN. *IOP Conference Series: Materials Science and Engineering*, *677*(5). https://doi.org/10.1088/1757-899X/677/5/052038

Zhou, H., Wang, X., & Zhu, R. (2022). Feature selection based on mutual information with correlation coefficient. *Applied Intelligence*, *52*(5), 5457–5474. https://doi.org/10.1007/s10489- 021-02524-x

Amin, A., Anwar, S., Adnan, A., Nawaz, M., Howard, N., Qadir, J., ... & Hussain, A. (2016). Comparing oversampling techniques to handle the class imbalance problem: A customer churn prediction case study. *Ieee Access*, *4*, 7940-7957.

Sana, J. K., Abedin, M. Z., Rahman, M. S., & Rahman, M. S. (2022). A novel customer churn prediction model for the telecommunication industry using data transformation methods and feature selection. *Plos one*, *17*(12), e0278095.

Osisanwo, F. Y., Akinsola, J. E. T., Awodele, O., Hinmikaiye, J. O., Olakanmi, O., & Akinjobi, J. (2017). Supervised machine learning algorithms: classification and comparison. *International Journal of Computer Trends and Technology (IJCTT)*, *48*(3), 128-138.

Tyagi, A. K. (2019, February). Machine learning with big data. In *Machine Learning with Big Data (March 20, 2019). Proceedings of International Conference on Sustainable Computing in Science, Technology and Management (SUSCOM), Amity University Rajasthan, Jaipur-India*.

Bhatnagar, A., & Srivastava, S. (2020, January). Performance Analysis of Hoeffding and Logisitic Algorithm for Churn Prediction in Telecom Sector. In *2020 International Conference on Computation, Automation and Knowledge Management (ICCAKM)* (pp. 377-380). IEEE.

Shirazi, F., & Mohammadi, M. (2019). A big data analytics model for customer churn prediction in the retiree segment. *International Journal of Information Management*, *48*, 238-253.

Liyanage, R. P. H., Kumara, B. T. G. S., Kuhaneswaran, B., & Prasanth, S. (2022). Deep Learning Approach for Detecting Customer Churn in Telecommunication Industry. In *Social Customer Relationship Management (Social-CRM) in the Era of Web 4.0* (pp. 196-215). IGI Global.

Rafiki, A., Hidayat, S. E., & Al Abdul Razzaq, D. (2019). CRM and organizational performance: A survey on telecommunication companies in Kuwait. *International Journal of Organizational Analysis*, *27*(1), 187-205.

Ray, S. (2019, February). A quick review of machine learning algorithms. In *2019 International conference on machine learning, big data, cloud and parallel computing (COMITCon)* (pp. 35-39). IEEE.

Rahmad, F., Suryanto, Y., & Ramli, K. (2020, July). Performance comparison of anti-spam technology using confusion matrix classification. In *IOP Conference Series: Materials Science and Engineering* (Vol. 879, No. 1, p. 012076). IOP Publishing.

Yacouby, R., & Axman, D. (2020, November). Probabilistic extension of precision, recall, and f1 score for more thorough evaluation of classification models. In *Proceedings of the first workshopon evaluation and comparison of NLP systems* (pp. 79-91).

Annexes

A. A letter submitted to Computer Science and Engineering department to request a formal letter for data collection purpose.

ቀን 16/05/2015

ጉዳዩ፦ የትብብር ደብዳቤ እንዲጻፍልኝ ስለመጠየቅ ይሆናል ከላይ በርዕሱ ለመግለጽ እንደተሞከረዉ እኔ ተማሪ የሮሳን ብርሃኑ ዘውዴ የተባልኩ የኮምፒዩተር ሳይንስ እና ምህንድስና ት/ት ክፍል የሁለተኛ ዓመት የድህረ ምረቃ ተማር ስሆን በአሁኑ ጊዜ የጥናታዊ ጽሁፌን በመስራት ላይ እገኛለሁኝ። ርዕሱም Telecom customer churn prediction model: The case of Ethio-telecom ሲሆን ለዚህ ጥናታዊ ጽሁፍ እንዲረዳኝ ዳታ ከኢትዮ ቴሌኮም መሰብሰብ ይኖርብኛል። ስለሆነም ለኢትዮ ቴሌኮም ዋና መስሪቤት የድጋፍ ደብዳቤ እንዲጻፍልኝ ስል በትህትና አጠይቃለሁኝ።

ከሰላምታ ጋር

የሮሳን ብርሃኑ

B. Ethio-telecom customer dataset

A	B	C	D	E	F	G	H
SERVICE_NUMBER	CUSTOMER_CATEGORY	CATEGORY	ADMINISTRATIVE_REGION	ZONE_SUB_CITY	GENDER	AGE	TOTAL_TRAFFIC_MIN
92757800	Residential	prepaid	Addis Ababa	Bole Sub City	Male	41	104.3666667
92724500	Residential	prepaid	Somali Region	Gode Zone	Male	22	202.65
92724500	Residential	prepaid	Somali Region	Gode Zone	Female	28	735.4166667
92757800	Residential	prepaid	Oromiya Region	Adama Special Zone	Male	53	241.3333333
92757800	Residential	prepaid	Oromiya Region	East shewa zone	Male	37	287.4166667
92724500	Residential	prepaid	Somali Region	Gode Zone	Male	20	628.3
92724500	Residential	prepaid	Somali Region	Gode Zone	Female	24	158.4333333
92724500	Residential	prepaid	Somali Region	Gode Zone	Male	26	75.73333333
92763800	Residential	prepaid	SNNP	GEDEO ZONE	Male	28	498.8333333
92724500	Residential	prepaid	Oromiya Region	West Aris Zone	Male	39	134.35
92724500	Residential	prepaid	Oromiya Region	Arsi zone	Female	31	32.93333333
92724500	Residential	prepaid	Oromiya Region	Bale Zone	Male	30	597.35
92724500	Residential	prepaid	Oromiya Region	Bale Zone	Male	31	14.35
92724500	Residential	prepaid	Oromiya Region	Bale Zone	Male	33	256.1333333
92724500	Residential	prepaid	Oromiya Region	Bale Zone	Male	33	88.18333333
92724500	Residential	prepaid	Oromiya Region	North shewa zone	Male	22	20.9
92724500	Residential	prepaid	Oromiya Region	Bale Zone	Male	22	51.5
92724500	Residential	prepaid	Oromiya Region	Bale Zone	Male	29	247.75
92763800	Residential	prepaid	Oromiya Region	West shewa zone	Female	26	36.16666667
92724500	Residential	prepaid	Oromiya Region	Bale Zone	Male	31	0

C. Customer dataset continued

I	J	K	L	M	N	O
TOTAL_FEE_ETB	SMS_USAGE/BIRR	SMS_FEE_ETB	DATA_USAGE_MB	DATA_REVENUE_ETB	VOICE_USAGE/MIN	churn
0.8775	5	0	27783.04	5.23	528.5867193	No
17.0145	11	0	0	0	528.5867193	No
7.0368	75	0	50216.17	6.96	528.5867193	No
31.7293	3	0	5.77	1.21	528.5867193	No
17.1146	2	0	0	0	528.5867193	No
11.3921	133	1.2	260.81	1.53	528.5867193	No
0.4953	38	0	0.04	0.09	528.5867193	No
34.4009	6	0.72	0	0	528.5867193	No
8.6405	19	0.12	858.33	0.81	528.5867193	No
0	0	0.48	0	0	528.5867193	Yes
13.4185	0	0.48	6.98	1.61	528.5867193	No
32.4938	2	0	92.21	6.39	528.5867193	No
6.6321	1	0.12	3.48	0.79	528.5867193	No
24.3874	16	0	0	0.01	528.5867193	No
7.6144	72	2.64	134.16	28.66	528.5867193	No
7.8831	9	0.84	0	0	528.5867193	No
24.8706	9	1.08	0	0	528.5867193	No
13.333	95	4.32	0	0	528.5867193	No
11.8895	4	0.24	0	0.05	528.5867193	No
0	0	0	192.74	0	528.5867193	Yes

D. sample code for feature selection (Information gain)

```
from sklearn.model_selection import train_test_split
X_train,X_test,y_train,y_test=train_test_split(df.drop(labels=['churn'], axis=1),
    df['churn'],
    test_size=0.3,
    random_state=0)

X_train.head()
```

```
from sklearn.feature_selection import mutual_info_classif
# determine the mutual information
mutual_info = mutual_info_classif(X_train, y_train)
mutual_info

mutual_info = pd.Series(mutual_info)
mutual_info.index = X_train.columns
mutual_info.sort_values(ascending=False)

#let's plot the ordered mutual_info values per feature
mutual_info.sort_values(ascending=False).plot.bar(figsize=(20, 8))
```

E. Sample code for building and evaluating the model

Support Vector Machine

SVM

```
# import required libraries
from sklearn.metrics import confusion_matrix, cohen_kappa_score, accuracy_score, precision_score, recall_score, f1_score
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.svm import SVC
from sklearn.metrics import confusion_matrix
from sklearn.preprocessing import LabelEncoder

from google.colab import files
uploaded = files.upload()
# read data from csv file
data = pd.read_csv('telecom_churn_dataset_4.csv')
#print(data)

# splitting data into training and test set
training_set,test_set = train_test_split(data,test_size=0.2,random_state=1)
#print("train:",training_set)
#print("test:",test_set)

# prepare data for applying it to svm
x_train = training_set.iloc[:,0:10].values # data
y_train = training_set.iloc[:,10].values # target
x_test = test_set.iloc[:,0:10].values  # data
y_test = test_set.iloc[:,10].values # target
#print(x_train,y_train)
#print(x_test,y_test)

# fitting the data (train a model)
classifier = SVC(kernel='rbf',random_state=1,C=10000000000,gamma='auto',)
classifier.fit(x_train,y_train)

# perform prediction on x_test data
y_pred = classifier.predict(x_test)
#test_set['prediction']=y_pred
#print(y_pred)
```

Random forest

RF

```
import pandas as pd
import matplotlib.pyplot as plt
from sklearn.preprocessing import MinMaxScaler
from sklearn.model_selection import train_test_split
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import confusion_matrix, cohen_kappa_score, accuracy_score, precision_score, recall_score, f1_score
from google.colab import files

uploaded = files.upload()
df = pd.read_csv('telecom_churn_dataset_4.csv')
scaler = MinMaxScaler()
scaler.fit(df)
df_norm = scaler.transform(df)
df_norm = pd.DataFrame(df_norm)

clf = RandomForestClassifier(
    n_estimators=2,
    max_features="sqrt",
    min_samples_split=10,
    min_samples_leaf=1,
    bootstrap=True,
    oob_score=True,
    verbose=0,
    criterion="entropy",
    random_state=42,
    class_weight={0: 0.5, 1: 0.5},
    n_jobs=-1,
  )



  clf.fit(x_train, y_train)
```

```
y_pred = clf.predict(x_test)
accuracy = accuracy_score(y_test, y_pred)
if acc_high < accuracy:
  acc_high = accuracy
high_tree = n
accuracies.append(accuracy)
```