ADDIS ABABA UNIVERSITY

COLLEGE OF NATURAL AND COMPUTATIONAL SCIENCES
SCHOOL OF INFORMATION SCIENCE

Predicting Quality of Service of ethiotelecom GSM Mobile Network using Machine Learning algorithms

By

Dereje Yihalem

January 2023

ADDIS ABABA UNIVERSITY

COLLEGE OF NATURAL AND COMPUTATIONAL SCIENCES

SCHOOL OF INFORMATION SCIENCE

Predicting Quality of Service of ethiotelecom GSM Mobile Network using Machine Learning algorithms

A Thesis Submitted to the School of Graduate Studies of
Addis Ababa University in Partial Fulfillment of the
Requirements for the Degree of
Masters of Science in Information Science and Systems (Information Systems)

By

Dereje Yihalem

Advisor

Dr. Solomon Teferra

January 2023

Predicting Quality of Service of ethiotelecom GSM Mobile Network using Machine Learning algorithms

By

Dereje Yihalem

Name and signature of Members of the Examining Board

Name	Title	Signature	Date
Solomon Tegegn (PhD)	Advisor		Feb 01/2023
Million Meshesha	Examiner		Feb 07/2023
Melkamu Beyene	Examiner		Feb 7, 2023

Declaration

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

Dereje Yihalem

January 2023

Dedication

I would like to dedicate this thesis work to my beloved family.

Acknowledgment

First and foremost, I would like to express my gratitude to the almighty God and to Saint Mary mother of God.

I would like to thank my research advisor Dr. Solomon Teferra for his extrovert guidance and support. He has shown me the right path of research and encouraged me to move forward throughout the study. He was always available to answer my questions, provide feedbacks and advises throughout the journey of this thesis work. Thank you Dr. Solomon!

My sincere thanks go to all my colleagues from the staff of the Ethiopian Communications Authority for providing the collected quality of service data and for their support during this work.

I am immensely indebted to my beloved family especially My Mother Asnakech Tesfaye, My brother Yehenew, and not mentioned here brothers & parents for giving me unconditional care, love, time, patient and support throughout my life.

Lastly, I offer my regards and blessings to all of those who supported me in any respect during the completion of the research.

Table of Contents

List of Tables

List of Figures

Acronyms

1G	First Generations
2G	Second Generations
3G	Third Generations
3GPP	Third Generation Partnership Project
4G	Fourth Generations
5G	Fifth Generations
ANSI	American National Standards Institute
ARIB	Alliance of Radio Industries and Business
AUC	Authentication Centre
BSC	Base Station Controller
BSS	Base Station Subsystem
BTS	Base Transceiver Station
CSSR	Call Set up Success Rate
CSV	Comma Separated Value
DCR	Drop Call Rate
EDGE	Enhanced Data rates in GSM Environment
EIR	Equipment Identity Register
ETSI	European Telecommunication Standard Institute
FDMA	Frequency Division Multiple Access
GPRS	General Packet Radio Service
GSM	Global Standard for Mobile Communication
HLR	Home Location Register

ICT	Information Communication Technology
IHSR	Inter cell Handover Success Rate
ISDN	Integrated Switch Digital Network
ITU	International Telecommunication Union
KNN	K-Nearest Neighbour
KPI	Key Performance Indicator
LR	Logistic Regression
LTE	Long Term Evolution
MS	Mobile System
MSC	Mobile Switching Centre
OMC	Operation Maintenance Centre
PSTN	Public Switch Telephone Network
QoS	Quality of Service
SDCCH	Stand-alone dedicated control channel
SMS	Short Message Service
SVM	Support Vector Machine
TCH	Traffic channels
VAS	Value Added Services
VLR	Visitor Location Register
WCDMA	Wide band Code Division Multiple Access

Abstract

Global System for Mobile Communication (GSM) is globally accepted standard for digital mobile communications. Ethiotelecom is one of the oldest telecom service providers in Africa, which offer telecommunication services in Ethiopia. GSM cellular mobile service is one of the various telecom services, which has millions of customers in Ethiopia. However, ethiotelecom has done many remarkable works in developing information and communications technology network, customers still have been complaining about the poor quality in GSM cellular mobile service. The prime objective of this study is building a predictive model using machine-learning techniques to determine the quality of service of GSM network for ethiotelecom Addis Ababa region, which helps to optimize quality of service in the area.

The data used in this study was obtained from Ethiopian Communications Authority quality of service department. For the aim of constructing the machine learning models, a total of 2294 data sets with 6 attributes are employed before preprocessing. After compilation of the primary dataset preprocessing task is undertaken to make suitable for the ML task like cleaning and attribute selection. Strictly, following the experimental research process, various experiments are conducted using python as a tool. This is done to find out the best model that classifies the KPI data by applying the best classification models by comparing the performance of the models developed using KNN, SVM, as well as the Logistic regression learning methods.

According to experimental results, logistic regression classification algorithm outperforms the other two classification algorithms with an accuracy of 99.854%. The finding indicates Call Setup Success Rate, Handover success rate, Dropped call rate and call attempt are the major determinant factors of QoS of GSM Network. The study also indicates that a GSM mobile network located in Addis Ababa and its surrounding are susceptible to failure in network quality.

Finally, the study is limited to build a predictive machine-learning model for classification of the dataset into the right level of QoS data. Through the results found in this study, we recommend ethiotelecom to implement such mobile network quality prediction techniques and avoid irregularities throughout the network that will improve customer satisfaction.

Key words: *Quality of service, Machine learning, Mobile network, GSM, Key Performance Indicators.*

CHAPTER ONE

INTRODUCTION

1.1 Background of the Study

The role and function of telecommunication is to provide an exchange of communication or information at a distance between people, satellites or computers. The remarkable development in telecommunication technology has made people to communicate instantly across a distance, share information and do business. The global system for mobile communication (GSM) is a worldwide-accepted standard for digital cellular communication. It is currently one of the accepted systems used for mobile telecommunications in all parts of the world. GSM is a second-generation (2G) digital cellular network developed to replace the first generation, which was analogue network [20].

All telecom Service Providers give National and International communication services on wired and wireless technologies. Cellular Network is one of them. There are different types of cellular network services too; voice call service is the common one. Most of the customers use voice service and this service generates the main source of income and profit for the operators [5].

Measuring service quality enables organization to know its position in the market and provides a strategic advantage to enhance its competitiveness [31]. Quality of Service (QoS) in the field of telecommunications can be defined as, "a set of specific requirements provided by a network to users, which are necessary in order to achieve the required functionality of an application or service" [31]. In GSM, Quality of Service is a critical issue of one's telecommunication service provider (Asoke k, 2007).

Ethiotelecom is the oldest public telecommunications operator in Africa. The services have started since 1894, when Minilik II, the King of Ethiopia, introduced telecommunication technology to the country & ethiotelecom has its current status in 29th November 2010. Since 1894, ethiotelecom was the only service provider In Ethiopia. However, recently on April 5, 2021, the government promises to be a milestone in Ethiopia's journey to become a digital economy for two new full-service telecom licenses to be award by the Ethiopian

Communications Authority (ECA). ECA has granted nationwide full-service unified telecommunication service license to "*Safaricom Telecommunication Ethiopia PLC*" effective from July 9, 2021. Moreover, bidding for the second telecom operator is in progress. The two new operators will compete with ethiotelecom in mobile communications, internet and other telecom services. Since, other operator does not start their service yet; this study will try to address the services provided by ethiotelecom only.

The main objective of quality of service is to maintain a basic minimum level of quality, to use competition, to improve quality, to promote consumer choice over quality vs. price and to ensure quality choices are available for all groups of consumers. Drive testis one of the method to evaluate the QoS operators. It is Assessing the coverage, capacity and Quality of Service (QoS) of a mobile radio network Benchmarking, Network optimization and Troubleshooting [11]. Performance experts in the telecom domain, specifically in ethiotelecom are expected to analyze the information in the measurements to manage and improve the quality of service (QoS). Sometimes it will be difficult to exhaustively produce essential information from this complex data by solely applying domain expertise and a prior knowledge as the performance experts do [2]. Machine learning techniques can assist and helpful to such tasks done by domain experts by automating the current manual method.

Machine learning is an application of AI that provides systems the ability to learn and improve automatically from experience without being explicitly programed, which are used for future predictions and identifying patterns in data [47]. Machine learning deals with algorithms that give computers the ability to learn, in much the same way as humans. This means that given a set of data, an algorithm infers information about the properties of the data, allowing it to make predictions about other data it may see in the future. It has found uses in different areas such as biotechnology, fraud detection, wireless networks, stock market analysis and national security [17].

Operators and regulatory authorities often report the performance of their network quality in terms of key performance indicators (KPIs). The major KPIs used for analyzing QoS are Call Set-up Success Rate (CSSR), Drop Call Rate (DCR), Handover Success Rate (HOSR), Traffic Channel (TCH) Congestion Rate (TCHCR) and Control Channel Set-up Failure (CCSF) [30].

This study tried to do the study by using the common KPIs and values of KPIs recommended by Ethiopian communication Authority for effective communication in Ethiopia.

Thus, in this study, an attempt has been made to propose machine learning model to analyze the data generated from GSM mobile network. So that, it can be possible to automatically analyze the quality of service of mobile network by applying machine learning techniques using the relevant KPIs extracted from the drive test dataset. Such a system can support the telecommunications service providers as well as regulatory authorities of the sector to take improvement measures based on the predictive information of the system.

1.2 Motivation

Currently in Ethiopia Communications Authority (ECA), mobile network data analysis was done by traditional simple statistical methods for network quality analysis and network performance analysis. However, the application of simple statistical techniques for data analysis is time consuming, error prone and tiresome activities. In this regard, to overcome the problem of simple statistical analysis an attempt has been made to apply machine-learning techniques, to show how to build a model from Ethiopian Communication Authority's GSM mobile network data and give effort to evaluate the Quality of Service given by ethiotelecom to customers.

According to Yared [4], in telecommunications sector there are many areas of active research interests. Some of these are telecom fraud detection, call drop analysis, prediction, and network element analysis for optimization.

In this research, the application of machine learning for 2G mobile networks is proposed specifically for predicting the QoS of GSM network. An attempt has been made to understand the behavior of quality of services (QoS) using machine learning techniques and design a predictive model that can determine QoS from Ethiotelecom GSM mobile network data.

1.3 Statement of the Problem

The International Telecommunication Union (ITU-T) defines QoS as, "The collective effect of service performance which determines the degree of service user satisfaction". QoS is the major techniques used for judging the performance of GSM services. Telecommunication regulatory authorities and the service providers are usually concerned in measuring quality of service to

protect the customers and service providers' interest respectively. Assessment of GSM services from service providers and regulators perspective is usually based on some Key Performance Indicators (KPIs). By using an extensive real world drive test dataset to show that, classical machine learning methods yield excellent prediction results.

Ethiotelecom is telecommunication service provider In Ethiopia. During the performance period from 01 July to 30 June 2020, the total subscribers of ethiotelecom reached 46.2 million. The annual report of the operator also indicates, among total number of subscribers, nearly half of them (22.4 million subscribers) are 2G users [63]. Even though, the mobile network generations has grown and currently reached fifth generation, peoples of developing countries like Ethiopia still utilizes second generation network because of economic and technological factors. Since key performance indicators of quality of service of mobile network are same for each generation of the network, the researcher believes that studying from 2G network is a milestone to proceed for other generation of mobile network for further study.

Even though, ethiotelecom has done many remarkable works in developing modern information and communications technology network, customers have been complaining about the poor quality in 2G GSM cellular mobile service [1]. Machine learning techniques can assist in these tasks by reducing the need for drive tests, and helping to predict and diagnose network failures even before they noticeably degrade the quality of service of the network users [14]. Performance evaluation of mobile networks can succeed via drive-test measurement systems as well as via network monitoring systems [9].

The developed automatic QoS prediction model for this study is thus suggested as a better replacement for the current manual method based on its accuracy and non-human involvement in predicting QoS of mobile network being investigated. This study utilizes the selected major KPIs to predict QoS of 2G GSM voice network.

There are several algorithms that can be used train predictive models using different data sources with different number of attributes. The learning natures of the algorithms vary based on the nature of data used. The is, therefore, to conduct series of experiments to select an appropriate algorithm and which set of the data attributes are relevant for the development of the best performing model for the given data set.

Hence, to this end, the study attempts to explore, investigate and answer the following main research questions.

1. Which set of data attributes are factors or relevant to determine the QoS of GSM network?
2. Which machine learning technique or model is suitable to predict and classify the KPI data in to the right level of QoS for 2G mobile voice network?

1.4 Objective

1.4.1 General objective

The general objective of this research is to construct an optimal model that predicts quality of service (QoS) of GSM mobile network using Machine Learning technique.

1.4.2 Specific Objectives

- To review literatures so as to understand the current sciences, technologies and standards related to this study.
- To conduct experiment and develop different predictive models and compare their performance.
- To select the best machine-learning algorithm which is more suitable to the data and the problem domain
- To evaluate the best performing model in different evaluation parameters.
- To report the result and forward recommendations for further studies.

1.5 Scope of the Study

Due to time limitation to cover each technology one by one and computational resource limitation to run the resulting data, the scope of this research work is limited to 2G GSM cellular network for voice only.

This thesis focuses on proposing machine learning based QoS predictive model using the proposed algorithms for 2G GSM voice network. The research followed experimental research approach to investigate the stated problem and undertaking the experimentation. The data for this

research was taken from Ethiopian Communication Authority. These data were collected around Addis Ababa in three different periods from February 2017 to January 17, 2019 by drive test machine. The original data contains a total of 2,294 records and 6 attributes.

The study did not examine the text message service and other mobile communication technologies such as 2.5G, 3G, 4G, LTE etc. experimentation has been done by supervised classification algorithms for selected and common key performance indicators (KPIs) of quality of service of mobile network.

1.6 Significance of the Study

Cellular mobile network operators throughout the world are grappling with the issues of enhancing quality of services [44]. The complex data collected from mobile network service providers and regulatory authorities shall be manage so that useful data may be extract and exposed. This research proposed an ML model to assess the data generated by mobile network from drive test systems. The findings of this study aid the network provider in improving the mobile network's service quality.

The objective of this research is to design a predictive model, so as, to determine QoS from Ethiopian communication Authority data using Machine-learning techniques, which further used to assist the Authority and operators in to optimize the ongoing GSM cellular mobile network.

This research also helps for the researchers as a ground root to study on mobile network quality. Generally, this study is important since its application can help to improve the quality of mobile network and satisfy mobile users or customers of the network. In addition, the paper will motivate other researches on field of GSM cellular network.

1.7 Organization of the Research

This research report organized into five chapters. The first chapter is discuss about the introduction part that introduces key points about the research including the background, the problem which the research paper want to address, the general and specific objective of the study, the scope and others. The second chapter devoted to review literatures on machine learning concepts and the domain of the research. Related research works that intersect machine learning and GSM mobile network specifically quality of service issues were reviewed. Chapter

3 contains research methodology and detail description of the dataset. The forth chapter deals with data preprocessing, experimentations and result interpretations. In this chapter, building of model with training dataset and validating the result with testing datasets and interpretation of the result of the experimentation were the major concern. Finally, a comparison of the algorithms used for reasonable accuracy was made. In the fifth chapter, conclusions and recommendations presented.

CHAPTER TWO

LITERATURE REVIEW AND RELATED WORKS

2.1 Overview of Telecommunication Network with QoS

This chapter has the background information of telecommunication networks and GSM technology. The architecture of each technology, description of the architecture, the technology difference and importance of QoS considered. Ethiotelecom provides these services for voice call. Most of the customers use voice service and this service generates the main source of income and profit for the operators [5]. That is exactly why this service mainly dealt with in this research.

2.1.1 Telecommunication

Telecommunication is a technology that eliminates distance between continents, between countries, between persons. The word communications, derived from the Latin word *communicatio*, the social process of information exchange, covers the human need for direct contact and mutual understanding. The word telecommunication, adding tele (distance), in which defined telecommunication as "information exchange in distance by various types of technologies over wire, radio, optical, or other electromagnetic systems" [32].

Telephone was invented by alexander Graham bell in March 1876. The introduction of telecommunication in Ethiopia dates back to 1894. Ethiopian Telecommunications Corporation is the oldest public telecommunications operator in Africa. The provision of mobile service has begun in 1999 with a capacity of 36,000 lines in Addis Ababa. By the end of December 2004, the number of subscribers reached about 207,000 [33].

2.1.2 Global System for Mobile Communication (GSM)

One of the most standard digital cellular telecommunications systems used all over the world is Global System for Mobile Communication (GSM) that is a Second Generation (2G) wireless access technology. It is the first cellular system to recognize digital modulation, network level architectures and services, Radio Frequency (RF) for GSM standard started at 1900 MHz It was

first used in Europe in 1991 and at present is one of the most popular digital cellular telecommunications systems amply used all over the world [5]. Voice service is the main and basic service provided by GSM.

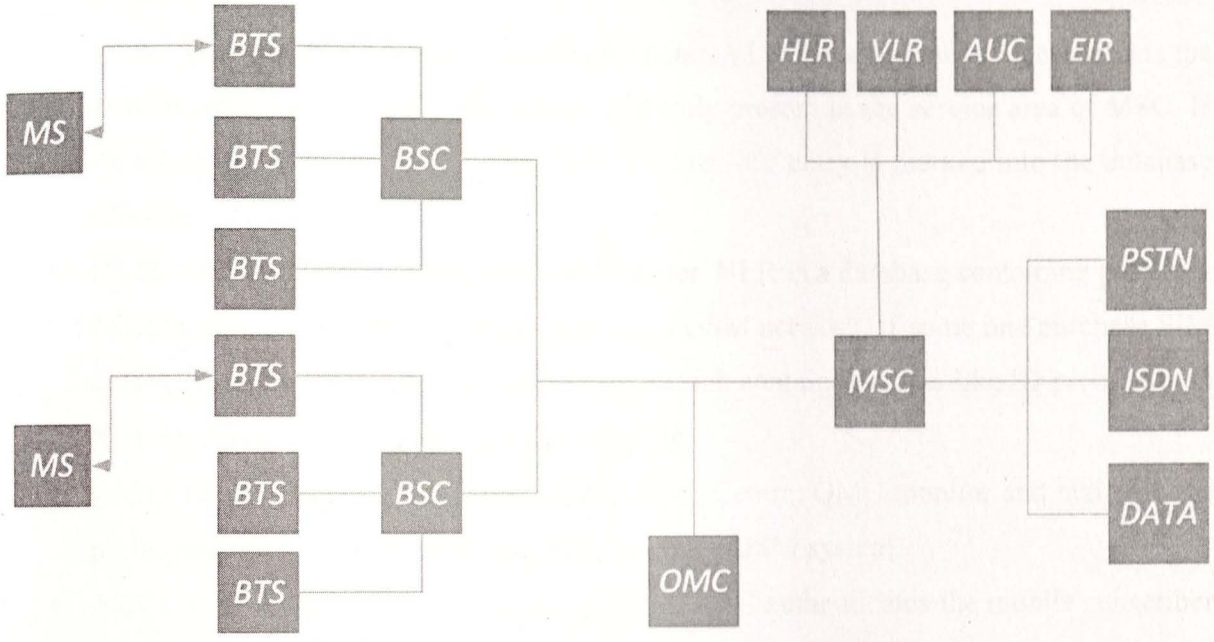


Figure 2.1, Global System for Mobile (GSM) Structure (Adopted from [5])

Mobile stations (MS) which are connected with the tower and that tower is connected to BTS through TRX, and then further connected to BSC and MSC. Let us understand the functionality of those different components.

1. MS: MS stands for Mobile System. MS comprises user equipment and software needed for communication with a mobile network. Mobile Station (MS) = Mobile Equipment (ME) + Subscriber Identity Module (SIM). These mobile stations are connected to tower and that tower connected with BTS through TRX. TRX is a transceiver that comprises transmitter and receiver. Transceiver has two performances of sending and receiving.

2. BTS: BTS stands for Base Transceiver Station, which facilitates wireless communication between user equipment and a network. Every tower has BTS.

3. BSC: BSC stands for Base Station Controller. BSC has multiple BTS. we can consider the BSC as a local exchange of our area, which has multiple towers and multiple towers have BTS.

4. MSC: MSC stands for Mobile Switching Center. MSC is associated with communication switching functions such as call setup, call release and routing. Call tracing, call forwarding all functions are performed at the MSC level. MSC is having further components like VLR, HLR, AUC, EIR and PSTN.

- **VLR:** VLR stands for Visitor Location Register. VLR is a database, which contains the exact location of all mobile subscribers currently present in the service area of MSC. If we are going from one state to another state then our entry is marked into the database of VLR.
- **HLR:** HLR stands for Home Location Register. HLR is a database containing pertinent data regarding subscribers authorized to use a GSM network. If some one purchase SIM card from in the HLR. HLR is like a home, which contains all data like ID proof, which plan are taking, which caller tune are using etc.
- **OMC:** OMC stands for Operation Maintenance Centre. OMC monitor and maintain the performance of each MS, BSC and MSC within a GSM system.
- **AUC:** AUC stands for Authentication Centre. AUC authenticates the mobile subscriber that wants to connect in the network.
- **EIR:** EIR stands for Equipment Identity Register. EIR is a database that keeps the record of all allowed or banned in the network. If we banned in the network them we cannot enter the network, and we cannot make the calls.
- **PSTN:** PSTN stands for Public Switched Telephone Network. PSTN connects with MSC. PSTN originally a network of fixed line analogue telephone systems. Now almost entirely digital in its core network and includes mobile and other networks as well as fixed telephones. The earlier landline phones which places at our home are PSTN.

2.1.2.1 Application of GSM

Access control devices: Access control devices can communicate with servers and security staff through SMS messaging. Complete log of transaction is available at the head-office Server instantly without any wiring involved and device can instantly alert security personnel on their mobile phone in case of any problem.

Transaction terminals: EDC machines, POS terminals can use SMS messaging to confirm transactions from central servers. The main benefit is that central server can be anywhere in the world. Today we need local servers in every city with multiple telephone lines. We can save huge infrastructure costs as well as per transaction cost.

Supply Chain Management: Today SCM require huge IT infrastructure with leased lines, networking devices, data centre, workstations and still you have large downtimes and high costs. We can do all this at a fraction of the cost with GSM M2M technology. A central server in our head office with GSM capability is the answer, we can receive instant transaction data from all branch offices, warehouses and business associates with nil downtime and low cost.

2.1.3 Standardization Organizations in Mobile Technologies

The major standardization bodies that play an important role in defining the specifications for the mobile technology as discussed in Mishra, 2004 are [35]:

ITU (International Telecommunication Union): The ITU, with headquarters in Geneva, Switzerland, is an international organization within the United Nations, where global telecom networks and services are coordinated in governments and the private sector. The ITU-T is one of the three sectors of ITU and produces the quality standards covering all the fields of telecommunications.

ETSI (European Telecommunication Standard Institute): This body was primarily responsible for the development of the specifications for the GSM. Owing to the technical and commercial success of the GSM, this body will also play an important role in the development of third-generation mobile systems. ETSI mainly develops the telecommunication standards throughout Europe and beyond.

ARIB (Alliance of Radio Industries and Business): this body is predominant in the Australasian region and playing an important role in the development of third generation mobile systems. ARIB basically serves as a standards developing organization for radio technology.

ANSI (American National Standards Institute): ANSI currently provides a forum for over 270 ANSI-accredited standards developers representing approximately 200 distinct organizations in

the private and public sectors. This body has been responsible for the standards development for the American networks.

3GPP (Third Generation Partnership Project): This body was created to maintain overall control of the specification design and process for third-generation networks. The result of the 3GPP work is a complete set of specifications that will maintain the global nature of the 3G networks.

2.1.4 Evolution of Mobile Network

In the last few decades, Mobile Wireless Communication networks have experienced a remarkable change. The mobile wireless Generation (G) generally refers to a change in the nature of the system, speed, technology, frequency, data capacity, latency etc. Each generation have some standards, different capacities, new techniques and new features, which differentiate it from the previous one. The first generation (1G) mobile wireless communication network was analog used for voice calls only. The second generation (2G) is a digital technology and supports text messaging. The third generation (3G) mobile technology provided higher data transmission rate, increased capacity and provide multimedia support. The fourth generation (4G) integrates 3G with fixed internet to support wireless mobile internet, which is an evolution to mobile technology and it overcome the limitations of 3G. It also increases the bandwidth and reduces the cost of resources. 5G stands for 5th Generation Mobile technology and is going to be a new revolution in mobile market which has changed the means to use cell phones within very high bandwidth. User never experienced ever before such high value technology, which includes all type of advance features, and 5G technology will be most powerful and in huge demand in near future.

Mobile network evolutions have been categorized into generations as shown in the following Figure 2.2. we took most of the contents regarding to evolution of mobile network, in the book titled "*Fundamentals of Cellular Network Planning and Optimization 2G/2.5G/3G... Evolution to 4G by Ajay R. Mishra*".

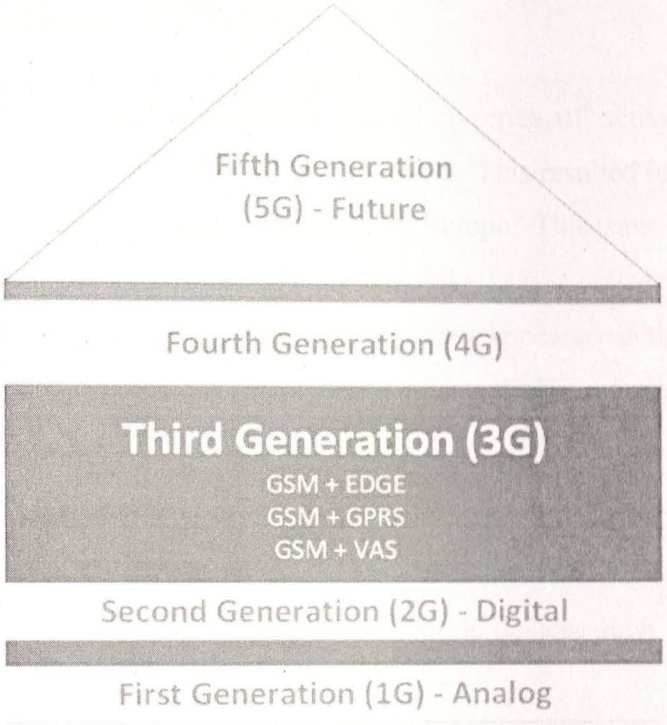


Figure 2.2, Evolution of mobile networks (Partially adopted from [35])

2.1.4.1 First-generation (1G) Network

The first-generation mobile system started in the 1980s was based on analogue transmission techniques. At that time, there was no worldwide (or even Europe-wide) coordination for the development of technical standards for the system. Nordic countries deployed Nordic Mobile Telephones or NMTs, while UK and Ireland went for Total Access Communication System or TAGS, and so on. Roaming was not possible and efficient use of the frequency spectrum was not there.

1G technology is a very basic voice analog phone system using circuited switched technology for the transmission of radio signals. All voice calls get Frequency modulated to higher frequency of 150MHz transmitted with frequency division multiple access (FDMA) technology in the frequency band of 824-894 MHz with channel capacity of 30 KHz, which was based on a technology called advance mobile phone service (AMPS) or total access communication system [51].

2.1.4.2 Second-generation (2G) Network

In the mid-1980s, the European commission started a series of activities to liberalize the communications sector, including mobile communications. This resulted in the creation of ETSI, which inherited all the standardization activities in Europe. This saw the birth of the first specifications, and the network based on digital technology; it was called the Global System for Mobile Communication or GSM. Since the first networks appeared at the beginning of 1991, GSM has gradually evolved to meet the requirements of data traffic and many more services than the original networks.

GSM and 2G Network: The Global System for Mobile Communications is a standard developed by the European Telecommunications Standards Institute to describe the protocols for second-generation digital cellular networks used by mobile devices such as mobile phones and tablets (Wikipedia).

The main elements of this system are BSS (Base Station Subsystem), in which there are the BTS (Base Transceiver Station) and BSC (Base Station Controllers); and the NSS (Network Switching Subsystem), in which there is the MSC (Mobile Switching Centre); VLR (Visitor Location Register); HLR (Home Location Register); AC (Authentication Centre), and EIR (Equipment Identity Register). This network is capable of providing all the basic services such as speech and data services up to 9.6 kbps, fax, etc. This GSM network also has an extension to the fixed telephony networks.

GSM and VAS (Value Added Services): The next advancement in the GSM system was the addition of two platforms, called Voice Mail System (VMS) and the Short Message Service Centre (SMSC). The SMSC proved to be incredibly commercially successful, so much so that in some networks the SMS traffic constitutes a major part of the total traffic. Along with the VAS, IN (Intelligent services), also made its mark in the GSM system, with its advantage of giving the operators the chance to create a whole range of new services. Fraud management and 'pre-paid' services are the result of the IN service.

GSM and GPRS (General Packet Radio Services)*:* As the requirement for sending data on the air-interface increased, new elements such as SGSN (Serving GPRS) and GGSN (Gateway

GPRS) were added to the existing GSM system. These elements made it possible to send packet data on the air-interface. This part of the network handling the packet data is also called the 'packet core network'. In addition to the SGSN and GGSN, it also contains the IP routers, firewall servers and DNS (domain name servers). This enables wireless access to the Internet and the bit rate reaching to 150kbps in optimum conditions.

GSM and EDGE (Enhanced Data rates in GSM Environment): With both voice and data traffic moving on the system, the need was felt to increase the data rate. This was done by using more sophisticated coding methods over the Internet and thus increasing the data rate up to 384 kbps.

2.1.4.3 Third-generation (3G) Network

In EDGE, high-volume movement of data was possible, but still the packet transfer on the air interface behaves like a circuit switches call. Thus, part of this packet connection efficiency is lost in the circuit switch environment. Moreover, the standards for developing the networks were different for different parts of the world. Hence, it was decided to have a network that provides services independent of the technology platform and whose network design standards are same globally.

Thus, 3G was born. In Europe it was called UMTS (Universal Terrestrial Mobile System), which is ETSI-driven. IMT-2000 is the ITU-T name for the third-generation system, while cdma2000 is the name of the American 3G variant. WCDMA is the air-interface technology for the UMTS. The main components include BS (base station) or node B, RNC (radio network controller) apart from WMSC (wideband CDMA mobile switching center) and SGSN/GGSN. This platform offers many Internet based services, along with video phoning, imaging, etc.

2.1.4.4 Fourth-generation (4G) Networks

The fundamental reason for the transition to the All-IP is to have a common platform for all the technologies that have been developed so far, and to harmonize with user expectations of the many services to be provided. The fundamental difference between the GSM/3G and All-IP is that the functionality of the RNC and BSC is now distributed to the BTS and a set of servers and

gateways. This means that this network will be less expensive and data transfer will be much faster.

4G offers a downloading speed of 100Mbps. 4G provides same feature as 3G and additional services like Multi-Media Newspapers, to watch T.V programs with more clarity and send Data much faster than previous generations [3]. LTE (Long Term Evolution) is considered as 4G technology. 4G is being developed to accommodate the QoS and rate requirements set by forthcoming applications like wireless broadband access, Multimedia Messaging Service (MMS), video chat, mobile TV, HDTV content, Digital Video Broadcasting (DVB), minimal services like voice and data, and other services that utilize bandwidth [45].

The main features of 4G are:

- Capable of provide 10Mbps-1Gbps speed
- High quality streaming video
- Combination of Wi-Fi and Wi-Max
- High security
- Provide any kind of service at any time as per user requirements anywhere
- Expanded multimedia services
- Low cost per-bit
- Battery uses is more
- Hard to implement
- Need complicated hardware
- Expensive equipment required to implement next generation network

2.1.4.5 Fifth-generation (5G) Networks

The eventual goal of the forthcoming 5G wireless networking is to have relatively fast data speeds, incredibly low latency, and substantial rises in base station's efficiency and major changes in expected Quality of Service (QoS) for customers relative to the existing 4G LTE networks. In order to deal with state-of-the art technologies and connectivity in the form of smart cell phones, internet of things (IoT) devices, autonomous vehicles, virtual reality devices and smart homes connectivity, the broadband data use has risen at a fast rate. Further, to meet the

latest applications, the bandwidth of the system needs to be increased widely [37]. In particular, the fifth generation (5G) mobile network seeks to resolve the shortcomings of previous telecommunication technologies and to be a possible primary enabler for future IoT applications [37].

2.1.5 Quality of Service in GSM Mobile Services

ITU defines Quality of Service as "the collective effect of service performance which determines the degree of satisfaction of the user of the service". In GSM cellular network quality of service described as the capability of the cellular service providers to provide a satisfactory service which includes voice quality, signal strength, low call blocking and dropping probability, high data rates for multimedia and data applications etc. This service is articulated based on the expected parameters value relative to user gained calculated over time parameter value [1].

Why Quality of Service?

The use of Cellular Mobile devices has become an important factor in human life as they are becoming more and more human companions. Massive use of mobile internet on social media, video streaming and other protocol applications is attracting everyone to have a mobile phone. Mobile phones use SIM cards, which are the main chip, provided by Mobile Network Operators to access different network services. Ensuring and anticipating user's Quality of Service needs, is the factor that draw the line between service providers and their competitions [29].

The quality of a service is a great separator in the Mobile business market. Today mobile users are switching from one operator to the other due to poor services complaining about poor internet speed, because of Quality of Service is not looked at efficiently [29].

2.1.5.1 Quality of Service Parameters

All GSM operators use quality of service parameters to judge their network performance and evaluate the Quality of Service regarding end user perspective [1]. The following two tables (table 2.1 and table 2.2) shows the standards of KPIs set by ethiotelecom and Ethiopian communication authority respectively.

No.	Quality of Service Parameters	Target/Set standards
1	Call Setup Success Rate	> 98%
2	Dropped Call Rate	<2%
3	Handover Success Rate	>95%
4	Blocked Call Rate	<2%
5	SDCCH Blocking Rate	< 0.5%
6	TCH Blocking Rate	< 5%

Table 2.1 GSM Mobile Service Set Standards Of Ethiotelecom [1]

No.	Quality of Service Parameters	Target/Set standards
1	Call Setup Success Rate	> 95%
2	Dropped Call Rate	<2%
3	Handover Success Rate	>96%

Table 2.2 GSM mobile service set standards of Ethiopian Communication Authority [12]

2.2 Overview of Machine learning

The paradigm of machine learning and artificial intelligence has pervaded our everyday life in such a way that it is no longer an area for esoteric academics and scientists putting their effort to solve a challenging research problem. The evolution is quite natural rather than accidental. With the exponential growth in processing speed and with the emergence of smarter algorithms for solving complex and challenging problems, organizations have found it possible to harness a humongous volume of data in realizing solutions that have far-reaching business values [49].

Machine learning is an application of AI that provides systems the ability to learn and improve automatically from experience without being explicitly programed. Which are used for future predictions (based on past data or Big Data) and identifying (discovering) patterns in data. Machine learning is itself a type of artificial intelligence that allows software applications to become more accurate in predicting outcomes without being explicitly programmed [47].

Machine learning deals with algorithms that give computers the ability to learn, in much the same way as humans. This means that given a set of data, an algorithm infers information about the properties of the data, allowing it to make predictions about other data it may see in the future. The focus of machine learning is the design of algorithms that recognize patterns and make decisions based on input data. Machine learning has found uses in areas like biotechnology, fraud detection, wireless networks, stock market analysis and national security [17].

2.2.1 Machine Learning Algorithms

Machine learning uses programmed algorithms that receive and analyses input data to predict output values within an acceptable range. As new data is fed to these algorithms, they learn and optimize their operations to improve performance, developing 'intelligence' over time.

The categories of Machine Learning Techniques are mainly divided into four categories: Supervised learning, unsupervised learning, Semi-supervised learning, and Reinforcement learning [48].

2.2.1.1 Supervised learning

In the supervised learning category of machine learning, the algorithms operates in such a way that it will develop a mathematical model of the data which comprises the inputs (data sent to a computer system) and the expected outputs (processed information sent out from a computer) [14]. In supervised learning, the machine is taught by example. The operator provides the machine-learning algorithm with a known dataset that includes desired inputs and outputs, and the algorithm must find a method to determine how to arrive at those inputs and outputs.

Supervised learning Algorithms

According to J E T Akinsola [52], the supervised machine learning algorithms which deals more with classification includes the following: Linear Classifiers, Logistic Regression, Naïve Bayes Classifier, Perceptron, Support Vector Machine; Quadratic Classifiers, K-nearest neighbor, Boosting, Decision Tree, Random Forest (RF); Neural networks, Bayesian Networks and so on.

One standard formulation of the supervised learning task is the classification problem: The learner is required to learn (to approximate the behavior of) a function which maps a vector into one of several classes by looking at several input output examples of the function. Inductive machine learning is the process of learning a set of rules from instances (examples in a training set), or more generally speaking, creating a classifier that can be used to generalize from new instances. Some supervised learning algorithms are discussed below.

Decision Trees

A decision tree is a non-parametric supervised learning algorithm, which is utilized for both classification and regression tasks. It has a hierarchical, tree structure, which consists of a root node, branches, internal nodes and leaf nodes.

A decision tree is a classifier expressed as a recursive partition of the instance space. The decision tree consists of nodes that form a rooted tree, meaning it is a directed tree with a node called "root" that has no incoming edges. All other nodes have exactly one incoming edge. A node with outgoing edges is called an internal or test node. All other nodes are called leaves (also known as terminal or decision nodes). In a decision tree, each internal node splits the instance space into two or more sub-spaces according to a certain discrete function of the input attributes values. In the simplest and most frequent case, each test considers a single attribute, such that the instance space is partitioned according to the attribute's value. In the case of numeric attributes, the condition refers to a range. Types of Decision Trees popular decision tree algorithms, such as the following.

ID3: ID3 algorithm is considered as a very simple decision tree algorithm (Quinlan, 1986). ID3 uses information gain as splitting criteria. The growing stops when all instances belong to a single value of target feature or when best information gain is not greater than zero. ID3 does not apply any pruning procedures nor does it handle numeric attributes or missing values.

C4.5: C4.5 is an evolution of ID3, presented by the same author (Quinlan, 1993). It uses gain ratio as splitting criteria. The splitting ceases when the number of instances to be split is below a certain threshold. Error–based pruning is performed after the growing phase. C4.5 can handle

numeric attributes. It can induce from a training set that incorporates missing values by using corrected gain ratio criteria.

CART: CART stands for Classification and Regression Trees (Breiman et al., 1984). It is characterized by the fact that it constructs binary trees, namely each internal node has exactly two outgoing edges. The splits are selected using the towing criteria and the obtained tree is pruned by cost–complexity Pruning. When provided, CART can consider misclassification costs in the tree induction. It also enables users to provide prior probability distribution. An important feature of CART is its ability to generate regression trees. Regression trees are trees where their leaves predict a real number and not a class. In case of regression, CART looks for splits that minimize the prediction squared error (the least–squared deviation). The prediction in each leaf is based on the weighted mean for node.

Logistic Regression

This algorithm is suited for binary classifications. Logistic regression algorithm calculates the class membership probability for one of the two categories in a dataset. It is best suited for data clearly separated by a single, linear boundary (Dreiseitl and Ohno-Machado, 2002; Smola and Vishwanathan, 2008)

Support Vector Machines

The algorithm aims at classifying data by finding linear decision boundary (called hyperplane) which separates the data classes. The algorithm aims at finding the hyperplane that has the largest margin between two classes. For nonlinear situations, the algorithm considers a loss function that penalizes the points on the wrong side of the hyperplane. Sometime this algorithm uses a kernel to transform nonlinearly separable data into higher dimensions where a linear decision boundary can be found. The SVM is suitable for binary data, but also discrete data can be used as input. High dimensional data can be managed easily. The algorithm performance decreases in presence of noise (Kotsiantis et al.,2007);

K-Nearest Neighbor

This algorithm categorizes an object depending on the classes of the nearest neighbours in the dataset. Consequently, the algorithm assumes that objects that are close to each other are similar. The algorithm can be trained using different distance metrics (e.g. Euclidean, Chebyshev, etc.). The algorithm can work with binary and discrete variables, but its performance is strongly affected by the data size and the presence of outliers and noise (Kotsiantis et al., 2007).

Neural Network

This algorithm consists of a set of simple, interconnected computation units called neurons, organized into layers with different roles called input, output and hidden layer, respectively. The number of hidden layers depends upon the model complexity. The neurons are connected via weighted links, and the way the neurons are connected defines different types of Neural Network. A Neural Network is trained iteratively to find the right weights for links. It best fits the modeling of highly nonlinear systems, when data are available incrementally and there is a constant need to update the model. Neural Networks algorithm can deal with noise and outliers in the dataset (Singh et al., 2016).

Advantages and Disadvantages of Supervised learning

The foremost advantage of Supervised Learning is that all classes or analog outputs manipulated by the algorithm of this paradigm are meaningful to humans. Moreover, it can be easily used for discriminative pattern classification, and for data regression. However, it also has several disadvantages. The first one is caused by the difficulty of collecting supervision or labels. When there is a huge volume of input data, it is prohibitively expensive, if not impossible, to label all of them. For example, it is not a trivial task to label a huge set of images for image classification. Second, as not everything in the real world has a distinctive label, there are uncertainties and ambiguities in the supervision or labels. For example, the margin for separating the two concepts of "Success" and "Fail" is not distinct. These difficulties may limit the applications of the Supervised Learning paradigm in some scenarios. To overcome these limitations in practice, other learning paradigms, such as Unsupervised Learning, Semi supervised Learning, Reinforcement Learning, Active Learning, or some mixed learning approaches can be considered [55].

Applications

Supervised Learning enables a machine to learn the human behavior or object behavior in certain tasks. The machine to perform similar actions on these tasks can then use the learned knowledge. Since the computing machinery may perform some input-output mappings much faster and more persistent than the human, machines equipped with a good supervised learner can perform certain tasks much faster and accurate than the human. On the other hand, because of the limitation in hardware, software, and algorithm designs, existing Supervised Learning algorithms still cannot match human is learning ability on many complicated tasks.

Supervised Learning have been successfully used in areas such as Information Retrieval, machine learning, Data Mining, Computer Vision, Speech Recognition, Spam Detection, Bioinformatics, Cheminformatics, Market Analysis etc. (Wikipedia, 2010).

2.2.1.2 Unsupervised Learning Algorithms

Unsupervised learning algorithms operate in such a way that it takes set of data and detect the patterns in it for grouping or clustering purpose. Unsupervised learning algorithms identify resemblance in the data and react based on the presence or absence of such identity in each new piece of data. The algorithms learn from test data that is not labeled, classified, or categorized. Unsupervised learning analyzes unlabeled datasets without the need for human interference, i.e., a data driven process [48]. The main purpose of these algorithms is to find the common patterns in previously unseen data. Clustering is the most popular form of unsupervised learning.

2.2.1.2.1 Advantages and Disadvantages Unsupervised Learning

There are some reasons why we sometimes choose unsupervised learning in place of supervised learning. Here are some of the advantages: It does not require a training data to be labeled. The automatic labeling of the training data set saving the time spent in hand classification and Classification task is fast. This type of learning is similar to human intelligence in some way as the model learns slowly and then calculates the result.

The cons of unsupervised learning algorithm are: There are no notions of the output along the learning process. It does not allow estimating or mapping the results of a new sample. Results vary considerably in the presence of outliers. It only performs classification tasks. It is also a time-consuming process. The learning phase of the algorithm might take a lot of time, as it analyses and calculates all possibilities. For some projects involving live data, it might require continuous feeding of data to the model, which will result in both inaccurate and time-consuming results.

Application: The applications for this learning are quite limited. However, the following are the ones, which are widely popular. It is mainly useful in fraud detection in credit cards, Useful for genome analysing and so on.

2.2.1.3 Semi-Supervised learning

As the name indicates, these algorithms combine labeled and unlabeled data to generate an appropriate mapping function or classifier. Artificial neural networks are extremely popular in the field of prediction in wireless networks. The other techniques like decision tress and unsupervised learning are used much lesser. Experiments prove that using a combination of techniques instead of a single one provides the best results [17].

The semi-supervised learning is situated between unsupervised learning (with no labeled training data) and supervised learning (with labeled training data). It is a hybrid form of machine learning techniques because it operates on labeled and unlabeled data, which brings a better accuracy. The major aim of unsupervised learning is to give great outcomes for prediction than the ones done with labeled data. The application areas of semi-supervised learning are text classification, fraud detection, machine translation, etc. [48].

2.2.1.4 Reinforcement learning

Reinforcement learning is used in operation research, game theory, information theory, swarm-intelligence, and genetic algorithms, etc. The learning uses the reward or penalty system, and the primary goal is to use leading obtained from environmental parameters to validate the reward or to minimize the risk involved. The algorithms are used in autonomous vehicles or in learning to play a game against a human opponent, it is an effective tool in training AI models for increase

automation which is used in robotics, autonomous driving tasks, manufacturing, and supply chain logistics [48].

2.2.2 Applications of ML in Telecommunication

Automating the optimization and management of wireless mobile networks has the potential to significantly reduce the operational costs for network operators, as well as to improve the quality of user experience. Currently, much of network management involves human interaction ranging from conducting drive tests in order to evaluate the network coverage and performance to diagnosing customer complaints. We argue that machine-learning techniques can assist in these tasks by reducing the need for drive tests, and helping to predict and diagnose network failures even before they noticeably degrade the quality of service of the network users [14].

Performance evaluation of mobile network can be succeeding via drive-test measurement systems as well as via network monitoring systems. Live measurements data are usually imported, stored and post-processed by database management system [15]. Quality analysis is proposed to be performed using machine learning methods.

Machine learning has proven to be the next disruptive force that every business needs and the telecommunication industry is not getting left behind. Telco's are at the forefront of the tech revolution and digital transformation to widen their services while improving consumer quality. In a world where consumers increasingly demand top-quality products and services, communication service providers turn to artificial intelligence and machine learning to help the companies deliver and fulfill their customers' expectations. Here are some application areas of machines learning in the telecom sector (Ronald Jaime, 2022):

2.2.2.1 Customer Service and Operational Support

Customer support has always been the Achilles of telecom companies. Companies often make it difficult for users to access customer support platforms like online chat, phone numbers, and contact forms to reduce user complaints. If the customer is lucky and connects with a customer care representative, they don't get the answers or assistance they seek. Customer support challenges start with having a limited number of people operating the phones and chats compared to the influx of customer complaints and requirements.

Using machine learning-based chat bots, telcos can comprehensively solve this serial problem. Chat bots are available 24/7 and can help customers quickly access the information they require using a ticketing system. NLP-based chat bots can take customer service further by interpreting the meaning behind the customer's words. For example, using the customer's tone and word choice, the bot can determine if the customer is frustrated or angry. Modern chat bots can also use machine learning algorithms and NLP to analyze historical information, networking logs, server ticketing data, and real-time customer input to deliver an amazing customer experience by effectively solving the customer's problems. Chat bots can also play a role in on-site maintenance, reducing the need for technical visits while cutting back on business costs.

2.2.2.2 Network Automation and Optimization

Modern communication networks are complicated and challenging to manage. The deployment of ML technologies and others like SDN can help operators leverage advanced automation in their network operations to optimize network architecture and improve management and control.

ML systems can predict and identify possible network-related issues and apply fixes that optimize reliability using the network and device data. AI, NLP, and ML can use various data parameters collected from the customer and their devices like requests, complaints, and service logs and analyze them to help telecoms uncover trends and performance issues in different demographics, time zones, devices, and locations.

2.2.2.3 Predictive Maintenance

Predictive maintenance is another top area where machine learning could help telecommunications companies improves service quality and reliability. Using sophisticated algorithms and machine learning ability, the companies can forecast future results by building historical data. ML systems can then use various data-driven techniques to monitor the current equipment condition and predict possible failure based on previous patterns. Using this information, telecoms can proactively fix various issues, including data center services, cell devices, and even devices placed in their customers' homes.

Telecoms can use AI and machine learning to use these capabilities across hardware, cloud, open-source frameworks, and neural networks. Ultimately, that means providing customers with a more reliable and stable network that guarantees improved customer experience and retention. As machine learning and AI develops in the telecom sector, their applications and benefits will broaden.

2.2.3 ML and Quality of Service use cases

Machine learning use cases in telecom have shown great potential in assisting with anomaly detection, root cause analysis, managed services, and network optimization. However, to work effectively, they require specific computational, pipeline and support infrastructure [57].

In the area of system monitoring, anomaly detection systems are crucial for identifying performance issues and problematic network behavior. Proactively predicting the degradation of key performance indicators, and identifying the likely root cause, can help reduce and prevent outages. In the area of managed services, ML models can improve trouble ticket management by effectively classifying, prioritizing, and escalating incidents. Capacity planning and customer retention can be improved through explainable churn prediction. Furthermore, in the area of intelligent networks, the incorporation of ML tools can enable self-healing radio networks, which automatically detect issues and take corrective actions [75].

2.3 Related Works

With the aim of searching for literature to understand the background of QoS of GSM network, the researcher attempted to retrieve a group of articles. The search parameters and synonyms that were used to logically guide the search engines included GSM network, 2G, Telecommunication and machine learning, cellular mobile communication, Drive test and so on. The researcher also identified literatures that are related to QoS of GSM network. The articles were searched in Google, Google Scholar, Research gate, IEEE etc. using keywords listed in the following Table 2.3.

Keywords Used for Searching Related Works
Telecommunications
Global system for mobile communications/ GSM
Evolution of mobile network
Quality of service of 2G network
Quality of service of network voice network
Application of Machine learning
Machine learning for wireless communication
Drive test
KPIs for Determining QoS of GSM network

Table 2.3, Key words for searching related works

Numerous telecommunications related works are using machine-learning techniques in order to raise the quality and efficiency of telecomm service. However, the contribution of local researches on ML is not enough. A number of researches and articles published in telecommunication service literature shown in the analysis of these service related works. Here are some of them, which reviewed along those lines.

2.3.1Local related research works

Muluken Tigabu [1], has conducted a research on how to build a predictive model using DM techniques to determine the occurrence of BTS failure for ethiotelecom North West Region

Gondar branch, which helps to optimize quality of service problems in the area. The primary source of his data is ethiotelecom North West Region Gondar Branch BSC database from September 2013 - August 2014 a total of 17,560 records before preprocessing. The research identifies the ongoing network problem and provides timely solution for the existing quality of service problems in the area.

The research used and followed the six steps CRISP-DM methodology to investigate the stated problem. J48, PART and SMO algorithms were implemented in WEKA 3.6.0.To build and compare classifier models. Finally, a model developed with J-48 classifier was taken as the final best working classification model

Lulu Deyu [2], addressed the applicability of data mining techniques to analyze the mobile telecommunication network QoS based on the selected KPIs. She tried to analyze the QoS for mobile telecommunications network applying different data mining techniques using the relevant KPI data extracted from the ethiotelecom network management system. On her study, a sample data indicating QoS KPIs has been taken from the live network of ethiotelecom.

Strictly, following the KDD process, she conducted experiments using the Weka open source data-mining tool. The classification model built on MLP has the best accuracy relative to Naïve Bayes classification algorithm.

Zenebe Kassaw [3], built radio coverage map for the purpose of UMTS network coverage prediction and hole detection using spatial interpolation techniques such as IDW and OK methods, using georeferenced RSCP measurement data collected from drive test and network topology information taken from ethiotelecom Addis Ababa, Ethiopia. According to Zenebe, the cellular coverage estimation performed through drive tests traditionally, which consist of geographically measuring different network coverage metrics with a motor vehicle equipped with mobile radio measurement facilities.

The collected coverage measurements through drive test are accurate but limited to roads and other regions accessible by motor vehicles. Drive tests cannot be conducted in the whole region of the network due to many obstacles such as buildings, lakes, and vegetation. Therefore, the drive test is quite inefficient means to solve the coverage problems and cannot offer a complete

and reliable picture of the network situation. In his study, experimental analysis was performed on a sample data collected from drive test UMTS network in Addis Ababa Ethiopia. Two general interpolation methods were employed with different parameters. The first method is IDW with various powers and number of neighbors and the second method is OK with Gaussian, Spherical and Exponential semivariogram models with different numbers of neighbors [3].

Yared Alibo Ayiza [4], designed a predictive model that can determine mobile call drops from ethiotelecom mobile network data. According to Yared, Mobile call drop is the main problems of all telecom operators. He selected around 20,000 records of one year and six months collection of Fault Management data. After eliminating irrelevant and unnecessary data, a total of 16996 datasets with 8 attributes are used for the purpose of conducting his study.

The study follows experimental research. In order to apply this experimental research, he used the six-step process of Hybrid Model. J48 decision tree algorithm with 10-fold cross validation registered better performance and processing speed of 95.43% and 0.06 sec respectively.

Teweldebrhan Mezgebo [6], present performance comparison and implementation of machine learning algorithms for automatic detection of anomaly cells for Addis Ababa LTE (4G) network. KNN based anomaly detection algorithms such as KNN classification, local outlier factor (LOF) and connectivity outlier factor (COF) anomaly detection models is implemented, and their comparative evaluation are made for Addis Ababa LTE cells.

The methodology adopted by the research is experimental research. According to Teweldebrhan, Experimentation results show that COF provides slightly better performance than the other models with negligible performance difference.

Addisu Shiferaw Fite [8], On his paper tries to evaluate the performance of two machine learning algorithms; Multivariate Linear Regression (MLR) and Support Vector Regression (SVR) in predicting video streaming perceived quality by end-users through the already measured video IFs, in Long-Term Evolution (LTE or 4G) for Addis Ababa scenario.

The methodology adopted by the researcher is also experimental research. From the results, the proposed video streaming QoE model shows a high correlation and low MSE between the

measured and the predicted QoE. The spatial distribution and density of the estimated QoE presented with the help of the preferred model.

2.3.2 Abroad related works

Charalampos N. Pitas, Konstantina E Chourdaki, Athanasios Panagopoulos and Philip Constantinou [9], Present data mining methods for speech and video quality analysis and prediction for GSM and UMTS mobile communication networks. Quality of speech and video telephony services is proved that can be discovered applying algorithms like the k nearest neighbor classifier, decision trees and artificial neural networks. The methodology they adopted is experimental and their results suggest that, learning from QoS measurements is suitable for building evaluation and prediction models.

Hassan Abdulkareem, Abdoulie Momodou Sunkary T ekanyi, Abduljalal Yushau Kassim and Ziyaulhaq Muhammad Zakariyya [10] Tries to assess the QoS of MTN GSM network in four geographical areas of Kaduna State (Kaduna south, Kaduna North, Zaria and Kafanchan), Nigeria. The data collated from the management center of MTN network was used for the evaluation of the measured KPI parameters using the data management tool.

According to them, the result of research paper was compared with those specified by the Nigerian Communications Commission (NCC). The quality of service of the KPI results for in the four locations during the three months' period of January, February, and March, 2016 showed considerable good performance by the MTN network in all the locations in terms of call drop rate because they all performed well within the benchmark.

Janne Riihijärvi and Petri Mähönen [14] Discusses the application of machine learning techniques for performance prediction problems in wireless networks. These problems often involve using existing measurement data to predict network performance where direct measurements are not available. They explore the performance of existing machine learning algorithms for these problems and propose a simple taxonomy of main problem categories.

As an example, they use an extensive real-world drive test data set to show that classical machine learning methods such as Gaussian process regression, exponential smoothing of time series and random forests can yield excellent prediction results. Applying these methods to the

management of wireless mobile networks has the potential to significantly reduce operational costs while simultaneously improving user experience.

Kabir Kadiri and Oluwaseun Samuel Lawal [16] Evaluates voice quality of four Global System for Mobile (GSM) Communication providers in five selected cities in Kwara State with thoughtfulness of network performance evaluation and the quality of service (QoS) improvement of GSM network system. Three assessment components/parameters which are network accessibility, service retainability and connection quality for evaluating QoS on the network were mainly adopted. The parameters were applied on four GSM networks in the studied areas using customers' complaints method. In addition, a standard method known as Perceptual Evaluation of Speech Quality (PESQ) - (International Telecommunication Union-Telecommunication Standardization Sector) ITU-T standard P.862, used for measuring call voice quality and Mean Opinion Score (MOS) is adopted. The two methods were therefore compared to assess call voice quality of the four GSM networks.

The Key Performance Indicators (KPIs) on which the GSM networks were tested include call set-up success rates (CSSR), call drop rate (CDR), call completion success rates (CCSR), handover success rates (HSR) and traffic channel congestion rate (TCHR).

The result of the study shows that the Quality of Service of GSM system in the selected cities is unreliable. The study also shows that the GSM network accessibility and retainability in the country are unsatisfactory. However, the call voice quality was observed to be on the peak in these cities across the four network providers. At the end of this manuscript, suggestions are given by them on how to advance both the Quality of Service and the positive impact of GSM network in the selected areas and the country as a whole.

Jide Julius Popoola and Adewale Enoch A reo [20] Developed automatic QoS prediction model thus suggested as a better replacement for the current manual method based on its accuracy and non-human involvement in predicting QoS of GSM network being investigated. Acording to Jide and Adewale, Comparison of some key performance indicators with standard threshold values has been a major approach for determining QoS of Global System for Mobile Communication (GSM) in Nigeria. This comparative approach, which usually involves human involvement, is prone to error.

Thus, an automatic artificial neural network (ANN) predictive QoS model was developed in their study and presented in the paper. In carrying out the study, five key performance indicators (KPIs) data were collected from the GSM operator used. The collected KPIs parameters were used to develop a mathematical model that was transformed into the proposed automatic QoS predicted model using ANN. The developed QoS prediction model, when evaluated was found to be accurate and could perform favorably well when compared with the manual approach being used by the Nigerian Communications Commission.

Rajesh Ganesan, B. Vinayagasundaram and X. Mercilin Raajin [22] Tries to achieving QoS in GSM Network by Efficient Anomaly Mitigation and Data Prediction Model. On the paper, the Call Detail Record (CDR) of the real-time data set is analyzed to find the traffic intense region over a spanning area.

The researcher adopted experimental research approach for his research work. In the study, the anomalies in the GSM network spread over an area are analyzed. Efficient bandwidth allocation results in the organized load balancing in the network which on a prolonged time frame will improve the Quality of Service (QoS) in the network.

Anna Corazza, Francesco Isgrò and Roberto Prevete [24] Describes a preliminary study in predicting failures in a mobile phones networks based on the analysis of real data. A ridge regression classifier has been adopted as machine learning engine, and interesting and promising conclusion were drawn from the experimental data.

In the paper, they described a possible strategy to tackle a problem of alarm prediction in a domain where time series of features are available. The alarm prediction has been defined as a classification problem that was solved by ridge regression. From the experimental results, the researchers conclude that geographical localization is important for the performance and that it is only possible to preview alarms occurring in a few hours.

Pasi Lehtim¨aki [25] Uses of measurement information in selection of most useful optimization action have been studied. In order to obtain good network performance efficiently, the expected performance of the alternative optimization actions must be possible to evaluate. In the thesis, methods to combine measurement information and application domain models are presented in

order to build predictive regression models that can be used to select the optimization actions providing the best network performance.

In the study, expert-based methods have been presented for the monitoring and analysis of multivariate cellular network performance data. These methods allow the analysis of performance bottlenecks having an effect in multiple performance indicators. In addition, methods for more advanced failure diagnosis have been presented aiming in identification of the causes of the performance bottlenecks. The use of measurement information in selection of most useful optimization action has been studied. According to [25], In order to obtain good network performance efficiently, the expected performance of the alternative optimization actions must be possible to evaluate. In the study, methods to combine measurement information and application domain models are presented in order to build predictive regression models that can be used to select the optimization actions providing the best network performance.

Segun I. Popoola, Aderemi A. Atayero, Nasir Faruk b and Joke A. Badejo [28] In the paper, the Key Performance Indicators (KPIs) for Quality of Service (QoS) of Global System for Mobile Communications (GSM) networks in Nigeria are provided and analyzed. The data provided in this study contain the Call Setup Success Rate (CSSR), Drop Call Rate (DCR), Stand-alone Dedicated Channel (SDCCH) congestion, and Traffic Channel (TCH) congestion for the four GSM network operators in Nigeria (Airtel, Etisalat, Glo, and MTN). These comprehensive data were obtained from the Nigerian Communications Commission (NCC).

On the paper, significant differences in each of the KPIs for the four quarters of each year were presented based on Analysis of Variance (ANOVA). The values of the KPIs were plotted against the months of the year for better visualization and understanding of data trends across the four quarters. Multiple comparisons of the mean-quarterly differences of the KPIs were also presented using Tukey's Post Hoc test.

Muwawa Jean Nestor Dahj [29] The study focuses on the application models of Data Mining and Machine Learning covering cellular network traffic, in the objective to arm Mobile Network Operators with full view of performance branches (Services, Device, and Subscribers). The purpose is to optimize and minimize the time to detect service and subscriber patterns behavior. Different data mining techniques and predictive algorithms have been applied on real cellular

network datasets to uncover different data usage patterns using specific Key Performance Indicators (KPIs) and Key Quality Indicators (KQI).

Prediction algorithms and models including Classification Tree, Random Forest, Neural Networks and gradient boosting have been used with an exploratory Data Analysis, determining relationship between predicting variables. The data is segmented in to two, a training set to train the model and a testing set to test the model. The evaluation of the best performing model is based on the prediction accuracy, sensitivity, specificity and the Confusion Matrix on the test set. With increase in Smart phone adoption, access to mobile internet services, applications such as streaming, interactive chats require a certain service level to ensure customer satisfaction. As a result, an SQM framework is developed by the researcher, with Service Quality Index (SQI) and Key Performance Index.

According to [30], Quality of Service (QoS) and Quality of end-user Experience (QoE) are the two major techniques used for judging the performance of GSM services. While the former is adjudged by service providers, the latter is determined by reactions from end-users (subscribers). Assessment of GSM services from service providers' perspective is usually based on some Key Performance Indicators (KPIs). The paper utilizes the five major KPIs (CCSF, TCH-CR, CSSR, DCR, HOSR) to assess the GSM QoS provided by MTN.

The research is exploratory research and results of the study show that the KPIs deviate from the recommended values both during the event and non-event period thus the QoS requires an uncompromising improvements in order to curtail further degradation in the services derived by the rapidly increasing subscribers' rate.

Betelehem Alemayehu [47] Proposed ML models that predict mobile network congestion using machine-learning algorithms in the Ethiotelecom cellular network. The researcher adopted experimental research methodology for her research work.

The results of the study showed that performance analysis of Multilayer Perception Neural Network models is a crucial process in model implementation of Multilayer Perception Neural Network for mobile network congestion prediction and a multilayer perceptron having 15 layers can give a comparable prediction of the real mobile network congestion situation.

Generally, the main intensions of all literatures are to investigate the quality of service issues in GSM cellular network. However, in our country there is no enough Machine learning research works on GSM Quality of service. Hence, this research paper is try to address the application of machine learning techniques on telecommunication especially on quality of services in GSM mobile network.

2.3.3 Summary on Related Works

Author & Year	Title	Methodology/ Approaches	Algorithms Used	Key Findings	Gap/ Remarks
Muluken Tigabu, 2015 E.C [1]	Application Of Data Mining Techniques to Predict Base Transceiver Stations (Bts) Failure Rate: The Case Of Ethiotelecom North West Region Gondar District	Cross-Industry Standard Process for Data Mining (CRISP-M)	J48, PART and SMO	• Bts that registered low Handover success rate and located in Gondar town and its surrounding are more susceptible to failure occurrences. • J-48 decision tree algorithm was taken as a final model for this particular study based on its performance relative to the other two classifiers.	Focuses only on network quality In rural areas. Urban areas which are the most congested network not included in the study
Lulu Deyu 2014 E.C [2]	Data Mining Approach to Analyze Mobile Telecommunications Network Quality Of Service: The Case Of Ethiotelecom	Knowledge Discovery in Databases (KDD) process	Multilayer perceptron and the Naïve Bayes	• The classification model built on MLP has the best accuracy relative to Naïve Bayes classification algorithm. •Among the KPIs studied, Call Success Rate (CSR) could measure an end-to-end performance of a mobile network.	Focuses only clustering approach.
Abdulkareem, Tekanyi	Analysis Of A GSM Network Quality Of	Quantitative Research	No	•The quality of service of the KPI results for in the four locations during the three months'	Use a limited number of

and Muhammad, 2020 [10]	Service Using Call Drop Rate And Call Setup Success Rate As Performance Indicators	Approach		period of January, February, and March, 2016 showed considerable good performance by the MTN network in all the locations in terms of call drop rate because they all performed well within the benchmark.	quality of service parameters
Janne and Petri 2018 [14]	Machine Learning for Performance Prediction in Mobile Cellular Networks	Experimental research	Random forests, Gaussian process regression and exponential smoothing of time series	•Using an extensive real-world drive test data set, the researchers tried to show that classical machine learning methods such as Gaussian process regression, exponential smoothing of time series and random forests can yield very good prediction results for drive test data.	Does not clarify the QoS in different telecom services also voice service quality not included in the study.
Kadiri1, Samuel and Olawale, 2019 [16]	Assessment of Call Voice Quality of GSM Network Operators in 5 Cities in Kwara State	Quantitative research approach	No	•None of the GSM network providers has up to 90% call completion success rate (CCSR). This is an indication that the service retainability of all GSM networks in the selected areas is very low •Furthermore, the result shows that there is better performance of all the networks in terms of service integrity, most especially in call voice quality.	Use traditional statistics method

				• The QoS and overall performance of the GSM network operation in the selected cities is poor, undependable and displeased.	
Jide Julius Popoola and Adewale Enoch A reo, 2020 [20]	Modeling and Development of a Novel Quality of Service Prediction Model for Global System for Mobile Communications Network using Artificial Neural Networks	Experimental Research Approach	scaled conjugate gradient (SCG) and Levenberg-Marquardt (L-M)	• Scaled conjugate gradient (SCG) training algorithm outperforms that of Levenberg-Marquardt (L-M)	Focuses only in Artificial Neural Networks
Dem ostenes, Renata and Grac, 2013 [27]	Predicting the Quality Level of a VoIP Communication through Intelligent Learning Techniques	Experimental Research Approach	Decision Trees, Multilayer Perceptron and Naives-Bayes	Decision Trees algorithm reach higher test results concordant with the results obtained.	The GSM network not included in the study
Muwawa Jean Nestor Dahj,	Data Mining And Predictive Analytics Application On Cellular Networks To Monitor	CRISP-DM (Cross-Industry Standard Process for Data	Classification Tree, Random Forest, Neural Networks and	Random Forest perform better than other algorithms	He is not including a hidden knowledge

2018, [29]	And Optimize Quality Of Service And Customer Experience	Mining)	Gradient boosting		
Betelehem Alemayehu, 2022 [47]	Mobile Network Congestion Prediction Using Machine Learning: The Case of Ethiotelecom	Experimental Research Approach	Multilayer Perceptron (MLP) having 10 and 15 layers	The study found that an average loss value of 1.192 and mean absolute error of 0.345 for the three sites using a multilayer perceptron having 10 hidden layers and average loss value of 1.2781 and mean absolute error of 0.272 for the three sites using a multilayer perceptron having 15 hidden layers.	The study is limited in MPL two layers only.

Table 2.4 Summery of related research works

CHAPTER THREE

RESEARCH METHODOLOGY AND DESCRIPTION OF DATASET

3.1 Research Methodology

Research methods in machine learning play a pivotal role since the accuracy and reliability of the results are influenced by the research methods used. On [38], the researchers analyzed a total of 100 articles published since 2019 in IEEE journals and the study revealed that Machine learning uses quantitative research methods with experimental research design being the de facto research approach.

However, most of the researchers used an experimental research design. This involved the design of an experiment and conducting the experiment to obtain results [38]. The general approach used in conducting experiments in this research is applied the phases that should be followed in any experimental design which are: Data collection, Data pre-processing, Model training, model testing, and model evaluation.

Experimental Design in Computational Intelligence is one of the most important aspects in every research process, thus it is crucial to correctly define all the steps that should be taken to ensure obtaining good results. This study used and followed an experimental research methodology to study the problem posed. The methodology was followed while the experiments were performed. The general research architecture applied for this research work is showed as the following on the figure 3.4.

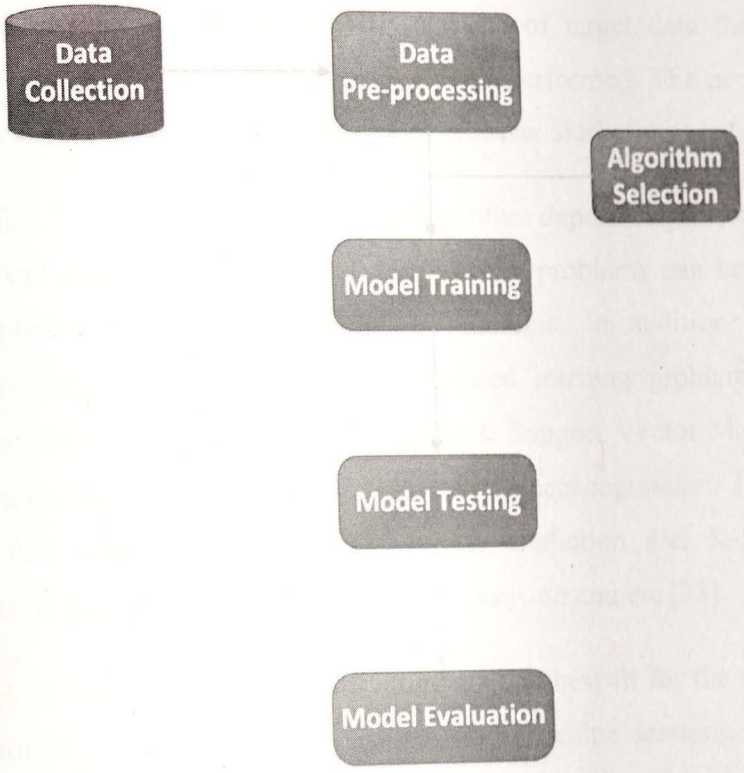


Figure 3.1: General Research Approach (Adopted from [38])

Data collection: In this stage, two main sources of data may use secondary data and primary data. The secondary data set was obtained from data repositories, government entities and Non-governmental organizations. Primary data was mainly sourced from recording data over a long period.

The data used for this study is collected from Ethiopian Communication Authority Quality of Service department. It consist of 2G GSM network collected in three different periods. Measurements collected from the drive test for voice only around Addis Ababa. To enhance the validity and reliability of the model this research used datasets, which are collected in three different seasons.

Data Preprocessing: this is a very critical stage in machine learning since this stage can influence the performance of models and consequently influence the results of the model. This is so because machine learning algorithms learn from the data provided to them and use the acquired knowledge to make decisions. Data pre-processing involved a series of activities. The activities are data cleanup, data normalization, and noise reduction etc.

In this study, from the initial original dataset, selection of target data then preprocess and transform into relevant data so as machine learning tasks performed. The performed major data preprocessing tasks has been explained in the next chapter (Chapter four) of this study.

Algorithm Selection: It is noted that the choice of algorithm depends mainly on the nature of the problem to be solved. According to the analyzed articles, problems can be grouped as either classification problems or regression/ prediction problems. In addition, problems can be classified as either supervised learning or unsupervised learning problems. The main pure classification algorithms used are K-Nearest Neighbors, Support Vector Machine and Logistic regression. The main purely prediction algorithm used is Linear regression. The most commonly used algorithms that support both classification and prediction are; K-Nearest Neighbors, Support Vector Machine, Artificial Neural Networks, Decision and etc.[38].

Since, the research is supervised learning this study applied best fit for the dataset used for this study and the problem domain and also, by adopting machine learning algorithm selection framework. After all, supervised machine algorithms K-Nearest Neighbors, Support vector machine and Logistic regression are used for model building.

Model Training: The data set was broken down into a training set and a testing set. The researcher used percentage split and 10-fold cross validation used to train and test the model. In this approach, the data set was divided into 10 equal parts. 9 parts or segments were used to train the model and then the remaining 1 segment was used to test the model. This was repeated 10 times with the testing segment being alternated among the 10 equal parts. Therefore, each part was used for testing once and for training 9 times.

Model Evaluation: The evaluation matrix that was used by this study is the confusion matrix and accuracy as the main evaluation metric for developed models. The researcher used more than one algorithm and evaluated the performance of each algorithm in the accuracy.

Tools and Techniques used: Python is delivering an easier and more efficient and effective way of doing machine-learning research. In addition, the vast libraries in python help researchers to perform numerous activities on the data or models with many conveniences [38].

This study used python programming language to develop the model, train the model, and test the model. The utilized python libraries are Pandas for data preprocessing, Matplotlib for Plotting and visualization, and the Scikit-Learn library for statistical modeling including classification tasks.

3.2 Data set Description

The original dataset: has 5 attribute and 1254 instances for 2009 E.C, 5 attribute and 444 instances for 2010 E.C and 5 attribute and 596 instances for 2011 E.C. totally, the data used for this research consists 5 attributes and 2294 instances before preprocessing.

Source: Ministry of innovation and technology, Telecom standard and Quality of service team. Currently this directorate reestablish as Ethiopian communications Authority.

3.2.1 Drive Test for 2009 E.C (DS-1)

The drive test is conducted on 6th February, 2017 to 10th February, 2017 between 10:00 AM and 3:30 PM. various areas of Addis Ababa are covered including major roads and some residential areas.

The drive test route covered various major roads and mostly the north-western and south-western part of the city including the central area. The total drive test route covered approximately 470 km.

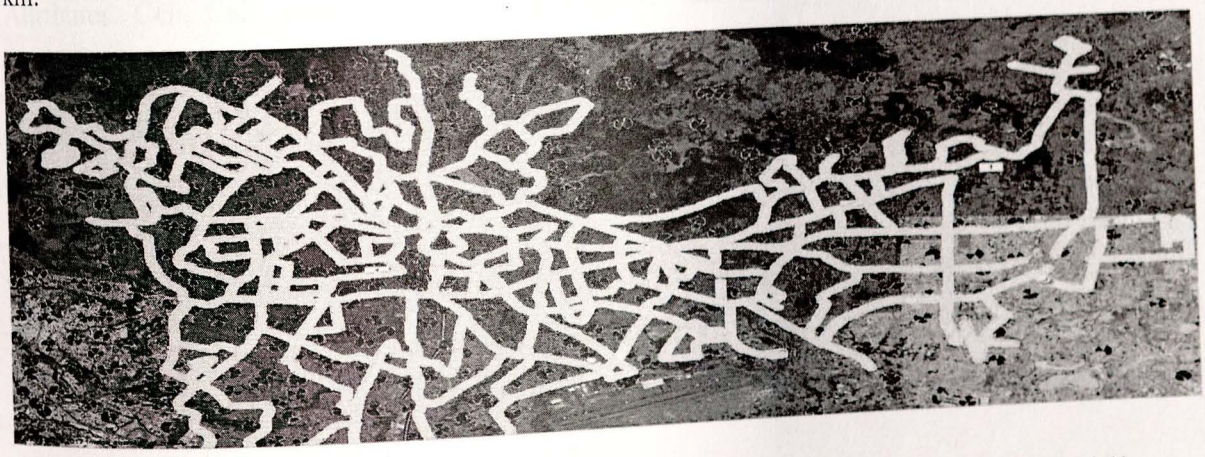


Figure 3.2, the route that was followed during the drive test (February 2009 E.C, AA)

3.2.2 Drive Test for 2010 E.C (DS-2)

The drive test is conducted on 21st – 22nd, 25th, 27th and 29th June, 2018 between 9:30 AM and 1:00 AM. Various areas of Addis Ababa (Nifas silk, Yeka and Gulele sub cities) are covered including major roads and some residential areas. The total drive test route covered approximately 153 km.

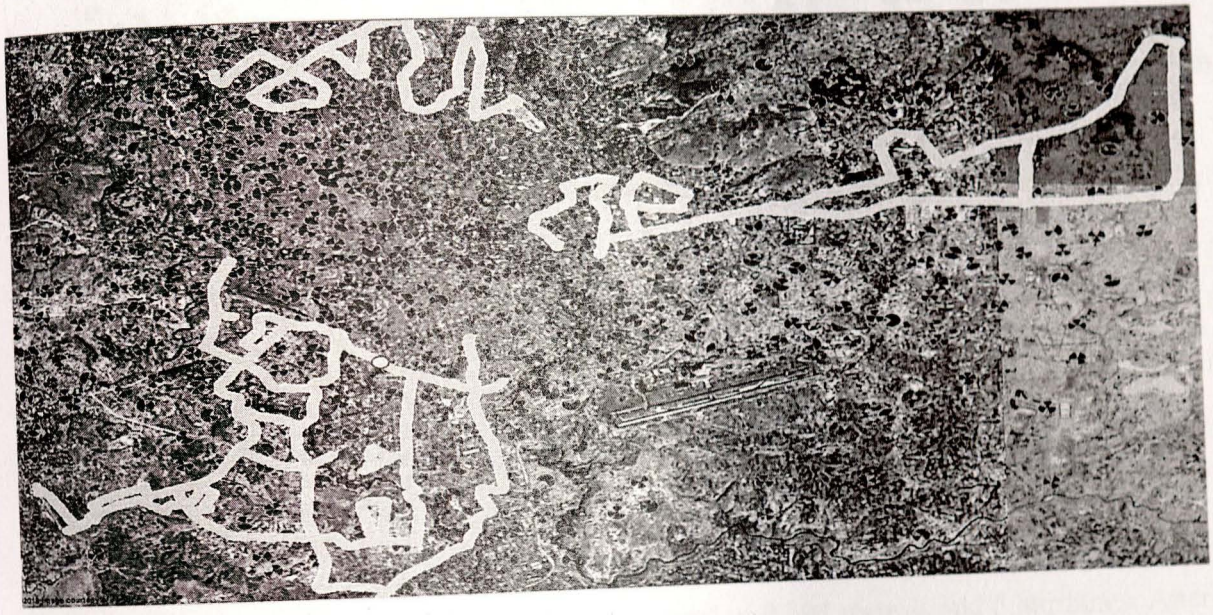


Figure 3.3, the route that was followed during the drive test (June 2010 E.C, AA)

Attributes: Cell, Call Attempts, Call Setup Failures, Call Setup ok, Call Setup Success Rate (CSSR, %), Dropped Call, Dropped Call Rate (DCR, %), Inter-cell Handover OK, Inter-cell Handover Fail, HO Success Rate (%) and finally the class (QoS).

3.2.3 Drive Test for 2011 E.C (DS-3)

The drive test is conducted from December 28, 2018 to January 17, 2019 between 9:30 AM and 1:00 PM. Most of the major roads of Addis Ababa are covered in the drive test. The total drive test route covered approximately 300km [36].

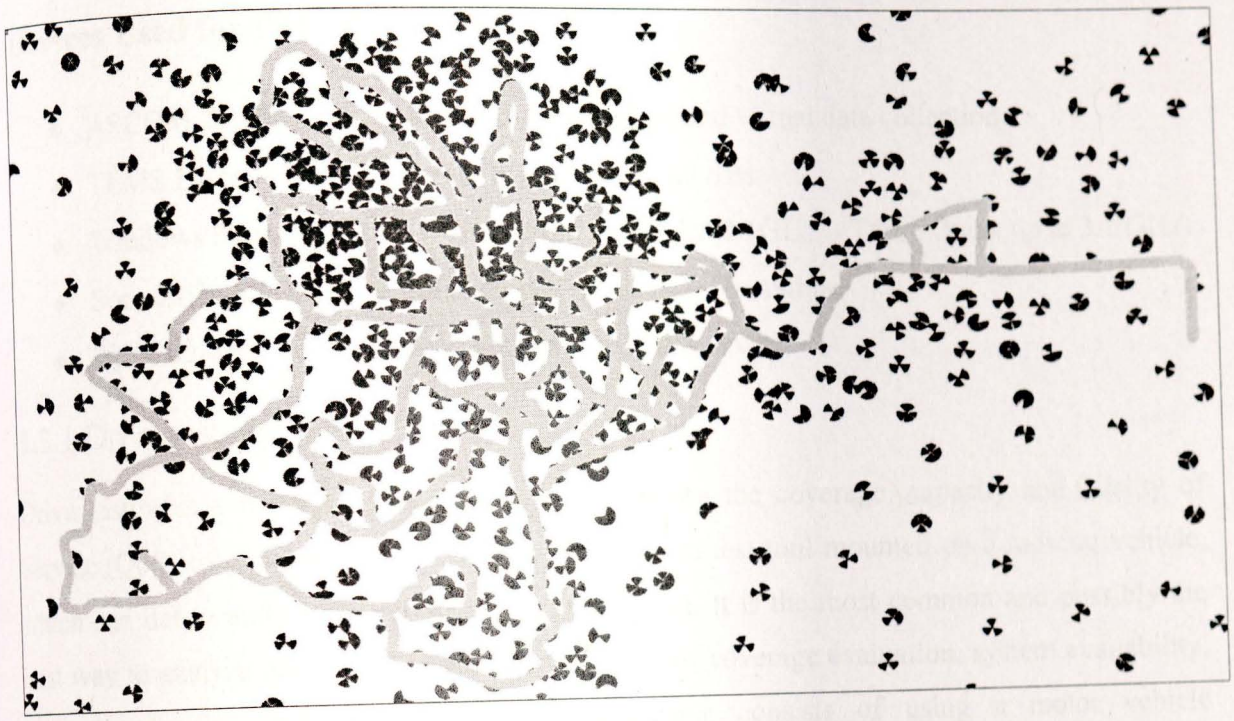


Figure 3.4 the route that was followed during the drive test (January 2011 E.C, AA)

In general, the data used for this research was taken from Ethiopian Communication Authority (previously MCIT) in quality of service department. These data are collected in three different periods from February, 2017 to January 17, 2019 a total of 2,294 records and 5 attributes. After preprocessing 2294 records and 4 attributes and 1 class attribute were used for the entire experiment and building predictive models.

3.3 Business Understanding

Ethiopian Communication Authority collected the data from drive test. Drive test is conducted on a moving vehicle by using ASCOM TEMS platform (software) for data collection for service quality monitoring purpose on GSM (2G) networks. The test software platform automatically generates test calls on mobile handsets. The test avoids any human-related impact of subjectivity as it is totally automated. The call duration is set to 120 seconds whereas the interval between calls is set to 2 seconds. The test was made on daytime as it best fits subscriber experience and the vehicle speed was kept between 30-50 KPH.

Devices Used for Data Collection

- ASCOM TEMS Investigation 17.2 for physical and virtual data collection
- TEMS Discovery 11.1.2 for analysis of collected data.
- Windows 8 based Laptop, (Acer core i7, 3612QM 2.1GHz, - Turbo Boost up to 3.1 GHz)
- Sony Ericsson (LT18i/LT18a) mobile handsets - 4 in number
- 900V Su-Kam Power Inverter and External GPS device

3.3.1 Drive testing

Drive testing is a method of measuring and assessing the coverage, capacity and Quality of Service (QoS) of a cellular mobile radio network by a test tool mounted on a moving vehicle, which can detect and record physical and virtual data. It is the most common and possibly the best way to analyze network performance by means of coverage evaluation, system availability, network capacity, and call quality. The technique consists of using a motor vehicle containing mobile network quality measurement equipment that can detect and record a wide variety of physical and virtual parameters of mobile cellular service in a given geographical area.

3.3.2 Ethiopian Communications Authority

Ethiopian Communications Authority (ECA) serves as the country's communications services and postal services regulatory authority in Ethiopia. ECA has the authority and mandate to monitor and measure the quality of Service (QoS) of Telecommunication Services provided by the operators (Ethiotelecom). Also it responsible for setting the standards of telecom Quality of Service (QoS), ensuring the Quality of the Service by conducting periodical drive test to protect the interests of the consumers of mobile networks in Ethiopia.

QoS is defined as the overall service performance which is particularly seen from the user of the service point of view. It refers to the capability of a network to provide better service to selected network traffic over various technologies. It is quantitatively determined by means of parameters or performance indicators referred to as key performance indicators (KPIs). Evaluation of QoS is determined by comparing some parameters against standard KPI values [12].

3.3.3 Ethiotelecom

Telecommunication service was introduced in Ethiopia by Emperor Menelik II in 1894 when the construction of the telephone line from Harar to the capital city, Addis Ababa, was commenced. Then the interurban network was continued to expand satisfactorily in all other directions from the capital. Many important centers in the Empire were interconnected by lines, thus facilitating long distance communication with the assistants or operators at intermediate stations frequently acting as verbal human repeaters between the distant calling parties [33].

The telecommunications sector was restructured and two separate independent entities namely, Ethiopian Telecommunications Authority (ETA) and the Ethiopian Telecommunications Corporation (ETC) were established by Proclamation No. 49/1996 on November 1996.

3.3.3.1 Establishment of Ethiotelecom

As a continuation of the 2005/06 to 2009/10, five-year plan and after concentrating its efforts on education, health and agriculture, the Ethiopian government has decided to focus on the improvement of telecommunication services, considering them as a key lever in the development of Ethiopia. Ethiotelecom was born, on Monday 29th November 2010, from this ambition of supporting the steady growth of our country, within the Growth Transformation Plan (GTP), with ambitious objectives for 2015 (https://www.ethiotelecom.et/).

3.3.4 GSM Network Analysis

The analysis of the network has been done in terms of accessibility, coverage, mobility, and retainability measurements [12].

I. Accessibility

Accessibility is the ability of a service to be obtained within specific tolerances and other given conditions, when requested by the user. In other words, it is the measure of the ability of a user to obtain the requested service from the system. Accessibility is monitored by measuring Call Setup Success Rate (CSSR), which is defined as the ratio of Established Calls to Call Attempts. Accessibility for voice is determined by the number of Call Setup Failures and their classification.

II. Retainability

Retainability is the ability of a service, once obtained, to continue to be provided under given conditions for a requested duration. During the Drive test, Retainability is measured by making a call from user A to user B and after the call is successfully setup, holding the call for duration of 120 sec. If the call drops during this period, it is considered to be a dropped call. The Call Drop Rate (CDR) parameter gives a reliable measure of the ability of the mobile network to maintain a call once it has been correctly established. The target set by ECA for **call drop rate** is to be less than **2%.**

III. Mobility

In a cellular system, a base station has only a limited coverage area. Hence, it is possible for a moving subscriber to be out of range of a base station while making a call. The process by which a mobile telephone call is transferred from one base station to another as the subscriber passes the boundary of a cell is called a handover. The Handover success rate (HOSR) more than 95% is considered to be good.

3.3.5 Key Performance Indicators

Key performance indicators are measurable parameters, which help to quantitatively represent quality of service of service providers. Below are some of the parameters that are used during the test to indicate the performance of the network operator, Ethiotelecom [36].

I. Call Setup Success Rate (CSSR)

Call setup success rate assesses the percentage of originating calls that were successfully established by customers. The target set by the Regulator for this Key Performance Indicator is >98%. Mathematically it is computed as:

*Call setup success rate (%) = (Total Calls successfully established/ Total number of call attempts)*100*

II. Handover Success Rate

Handover success rate measures the ability of a customer to talk on the cell phone for a long distance without getting disconnected. It is the ability of a call connection to be handed over from one cell to another cell. This parameter is directly linked to Call drop rate because a handover failure normally results in a dropped call. The target set by the Regulator for this Key Performance Indicator is >95%.

*Handover success rate (%) = (1- (Total calls dropped during handover/ Total handover attempts)) *100%*

III. Dropped Call Rate

Dropped call rate is a measure of calls that are prematurely disconnected before end of conversation against the number of all successfully established calls. Call drops may be experienced due to network problems such as handover failure or equipment faults. The target set by the Regulator for this Key Performance Indicator is <2%. Mathematically it is computed as:

*Dropped call Rate (%) = (Total calls dropped after being established/ Total Calls successfully established)*100%*

IV. Blocked call rate

Blocked call rate is the rate at which call attempts are blocked or unsuccessful because of a lack of resources for connection due to congestion. The target set by the Regulator for this Key Performance Indicator is <2%. Mathematically it is computed as:

Blocked call rate (%) = (100% - Call setup success rate)

V. Call Setup Time

The call set up time can be defined as the time interval from the instant the user initiates a connection request until the complete message indicating call disposition is received by the calling terminal.

Latency of the network is measured by the call setup delay. The target set by the Regulator for this Key Performance Indicator is < 25 seconds.

VI. Received Signal Strength (Power level)

Signal strength is expressed as the magnitude of the received signal by a mobile phone from a cellular network, commonly measured dBm. Various factors affect signal strength and areas which have signal strength between 0 dBm to -85dBm are considered to have good coverage, whereas areas with signal strength less than -120dBm are considered to be poorly covered.

Coverage - Signal Strength is computed as:

*Coverage rate - Signal strength 0 to -85dBm (%) = (Number of Calls with signal strength between 0 to -85dBm /Total number of calls made)*100%*

VII. Received signal quality

Voice quality is determined by the received signal quality. It refers to the clearness of a speaker's voice as perceived by a listener. The parameter used to classify the level of quality of voice is RxQual. It is a value between 0 and 7, where each value corresponds to an estimated number of bit errors in a number of bursts (DataStream transmitted in one timeslot). Normally, RxQual value 0-5 are considered to be good and RxQual greater than 5 are considered to be bad.

RxQual	Bit Error Rate (BER)
0	BER < 0.2%
1	0.2% < BER < 0.4%
2	0.4% < BER < 0.8%
3	0.8% < BER < 1.6%
4	1.6% < BER < 3.2%
5	3.2% < BER < 6.4%
6	6.4% < BER < 12.8%
7	12.8% < BER

Voice Quality is computed as: *Samples with good voice quality (%) = (Rx Qual samples with 0-3 value/Total Rx Qual Samples)*100%*

VIII. Chip Energy to Interference Ratio

Signal energy to interference ratio is the ratio of energy per chip (code bit) to the overall power density including of the interfering signals. In case no true interference is present, the

interference level is equal to the noise level. In other words, Ec/Io equals Ec/No. Technically Ec/Io should be the correct measurement but, due to equipment capability, Ec/No is actually measured. In UMTS, Ec/No and Ec/Io are often used interchangeably. It worsens as sector traffic load increases. It is usually measured in decibels (db) and basically it should be -16 db or higher in 95% of the coverage area.

The percentage in which the signal energy to interference ratio is greater than -16 db is calculated as: E_c/I_o *(%) = (E_c/I_o samples with 0 to -16db/Total E_c/I_o Samples)*100%*

IX. Received Signal Code Power

Received Signal Code Power (RSCP) denotes the power measured by a receiver on a particular physical communication channel. It is used as an indication of signal strength, as a handover criterion, in downlink power control, and to calculate path loss. While RSCP can be defined generally for any CDMA system, it is more specifically used in UMTS. Also, while RSCP can be measured in principle on the downlink as well as on the uplink, it is only defined for the downlink and thus presumed to be measured by the UE (User Equipment) and reported to the Node B (a term used in UMTS equivalent to BTS in GSM).RSCP is commonly measured in dBm and values less than -100dBm are considered to be poor.

The analysis of network quality is done in terms of accessibility (call set up success rate), mobility (Inter cell handover rate) and retainability (Call drop rate) measurements [12]. Finally, the study focuses on proposing machine learning based Quality of service of predictive model using the proposed algorithms for GSM voice network by using the main determinants key performance indicators of QoS of GSM mobile network based on literature suggestion and the recommendations of local and international standardization & regulatory authorities which are, International telecommunication union (ITU) and Ethiopian Communication Authority (ECA).

CHAPTER FOUR

DATA PREPROCESS, EXPERIMENTAL RESULTS AND ANALYSIS

4.1 Overview

This chapter discusses the data preprocessing, experimental results and model building conducted during the research. As stated in previous chapters, the domain experts in Ethiopian communication Authority have been doing the analysis using simple statistical method on the data, which is extracted from ASCOM TEMS system/software. This simple statistical method is unable to utilize the whole data and reach at good analysis result. To overcome this problem, this study proposes the machine learning techniques, algorithms and methods.

4.2 Data Preparation

The main purpose of data preparation is to make the data more suitable for the next step, which is the modeling phase. There are different methods for data preprocessing such as, data cleaning, data integration, data transformation, data reduction and data formatting and so on. For this research work, the following data preprocessing tasks has been performed.

4.2.1 Data cleaning

Data cleaning refers to the pre-processing of data in order to handling noise and missing values [1]. It is the process of ensuring that all values in a dataset are consistent and correctly recorded. To do so all the data that are available on the database was cleaned to make suitable for the model-building task. For this purpose, the researcher makes use of the python tool. Moreover, MS - Excel application used for labeling the data. In this subsection, different data cleaning tasks was carried out.

4.2.1.1 Filling Missing Value

Missing values refers to one or more fields of an attribute, which have no value in it. The existence of many such cases makes datasets incomplete and building models of any type whether descriptive or predictive with incomplete data makes the resulting model non-

representative of the reality [1]. Since, the data type used this study is continuous, the researcher tried to handle missing values of the attribute and replace with mean value. The researcher replaced the missing values through a data analysis tool, python. Among five attributes 2294 (11470 / 3883 missing) = 29% missing values.

Cell	Call attempts	(CSSR, %)	(DCR, %)	IHSR (%)
BCCH-111	24	100	0	100
BCCH-111_octBSIC-12	8	100	0	N/A
BCCH-111_octBSIC-13	4	100	0	100
BCCH-111_octBSIC-14		N/A	N/A	100
BCCH-111_octBSIC-17	8	100	0	100
BCCH-111_octBSIC-2		N/A	N/A	100
BCCH-111_octBSIC-20		N/A	N/A	100
BCCH-111_octBSIC-21		N/A	N/A	100
BCCH-111_octBSIC-22	8	100	0	100
BCCH-111_octBSIC-23	4	100	0	100
BCCH-111_octBSIC-25	24	100	0	100
BCCH-111_octBSIC-3		N/A	N/A	100
BCCH-111_octBSIC-30	4	100	0	100
BCCH-111_octBSIC-31		N/A	N/A	100
BCCH-111_octBSIC-33		N/A	N/A	100
BCCH-111_octBSIC-34		N/A	N/A	100
BCCH-111_octBSIC-35	4	100	0	100
BCCH-111_octBSIC-37	20	100	0	100
BCCH-111_octBSIC-4	12	100	0	N/A

Table 4.1, sample original dataset before with missing value

Cell	Call attempts	(CSSR, %)	(DCR, %)	(IHSR, %)
BCCH-111	24	100	0	100
BCCH-111_octBSIC-12	8	100	0	99
BCCH-111_octBSIC-13	4	100	0	100
BCCH-111_octBSIC-14	6	0	1.5	100
BCCH-111_octBSIC-17	8	100	0	100
BCCH-111_octBSIC-2	6	0	1.5	100
BCCH-111_octBSIC-20	6	0	1.5	100
BCCH-111_octBSIC-21	6	0	1.5	100
BCCH-111_octBSIC-22	8	100	0	100
BCCH-111_octBSIC-23	4	100	0	100
BCCH-111_octBSIC-25	24	100	0	100
BCCH-111_octBSIC-3	6	0	1.5	100
BCCH-111_octBSIC-30	4	100	0	100
BCCH-111_octBSIC-31	6	0	1.5	100
BCCH-111_octBSIC-33	6	0	1.5	100
BCCH-111_octBSIC-34	6	0	1.5	100
BCCH-111_octBSIC-35	4	100	0	100
BCCH-111_octBSIC-37	20	100	0	100
BCCH-111_octBSIC-4	12	100	0	99

Table 4.2, sample dataset after replaced missing values with mean

4.2.2 Feature Selection

Feature selection, as a dimensionality reduction technique, aims to choose a small subset of the relevant features from the original ones by removing irrelevant, redundant, or noisy features. Feature selection usually leads to better learning performance, i.e., higher learning accuracy, lower computational cost, and better model interpretability. Removing irrelevant features will not affect learning performance. In fact, the removal of irrelevant features may help learn a better model, as irrelevant features may confuse the learning system and cause memory and computation inefficiency [39].

According to N. Hoque, D. K. Bhattacharyya and J. K. Kalita [60], The most common feature selection methods are: Filter methods (ANOVA, Pearson correlation, variance thresh holding), Wrapper methods (forward, backward, and stepwise selection) and Embedded methods (Lasso, Ridge, Decision Tree).

4.2.2.1 Filter Methods

Filter methods select features based on a performance measure regardless of the employed data-modeling algorithm. Only after the best features are found, the modeling algorithms can use them. Filter methods can rank individual features or evaluate entire feature subsets [40].

For filter models, features are selected based on the characteristics of the data without utilizing learning algorithms. This approach is very efficient. However, it does not consider the bias and heuristics of the learning algorithms. Thus, it may miss features that are relevant for the target-learning algorithm. A filter algorithm usually consists of two steps. In the first step, features are ranked based on certain criterion. In the second step, features with the highest rankings are chosen [40].

4.2.2.2 Wrapper Methods

In wrapper methodology, selection of features is performed by considering it as a search problem, in which different combinations are made, evaluated, and compared with other combinations. It trains the algorithm by using the subset of features iteratively.

The major disadvantage of the filter approach is that it totally ignores the effects of the selected feature subset on the performance of the clustering or classification algorithm. The optimal feature subset should depend on the specific biases and heuristics of the learning algorithms. Based on this assumption, wrapper models use a specific learning algorithm to evaluate the quality of the selected features [39]. Thus, for classification tasks, a wrapper will evaluate subsets based on the classifier performance (e.g. KNN, SVM etc.).

4.2.2.3 Embedded Models

Filter models are computationally efficient, but totally ignore the biases of the learning algorithm. Compared with filter models, wrapper models obtain better predictive accuracy estimates, since they take into account the biases of the learning algorithms. However, wrapper models are very computationally expensive. Embedded models are a trade-off between the two models by embedding the feature selection into the model construction [39].

Thus, embedded models take advantage of both filter models and wrapper models. They are far less computationally intensive than wrapper methods, since they do not need to run the learning models many times to evaluate the features, and they include the interaction with the learning model. The biggest difference between wrapper models and embedded models is that wrapper models first train learning models using the candidate features and then perform feature selection by evaluating features using the learning model, while embedded models select features during the process of model construction to perform feature selection without further evaluation of the features [39].

4.2.2.4 Key Features

Feature selection is a very complicated and vast field of machine learning; there is no fixed rule of the best feature selection method. However, choosing the method depend on a machine learning engineer who can combine and innovate approaches to find the best method for a specific problem. One should try a variety of model fits on different subsets of features selected through different statistical Measures (*www.javapont.com).*

Finally, the researcher were selected significant attributes by using wrapper and filter method through python software, Sequential Forward Selection (SFS) method in linear regression model, together with their Average cv_score (Cross – Validation Score) value.

	feature_idx	cv_scores	avg_score	feature_names
1	(2,)	[0.9139382524335267]	0.913938	((CSSR, %),)
2	(2, 3)	[0.9409010182925642]	0.940901	((CSSR, %), (DCR, %))
3	(2, 3, 4)	[0.9522319962155864]	0.952232	((CSSR, %), (DCR, %), (IHSR, %))
4	(1, 2, 3, 4)	[0.9540643398209644]	0.954064	(Call attempts, (CSSR, %), (DCR, %), (IHSR, %))
5	(0, 1, 2, 3, 4)	[0.9541328577224827]	0.954133	(Cell, Call attempts, (CSSR, %), (DCR, %), (IH...

Table 4.3 Ranked features in using linear regression.

After attributes are ranked, the researcher tried to remove the list rank attribute and build a model until significance change is not happening on the accuracy results. As result, the ranked attributes listed from top to bottom as follows: (CSSR, %), (DCR, %), (IHSR, %), Call attempts and Cell.

It clearly shows that attribute "Cell" ranked on the last row and the researcher tries to build a model before and after attribute "Cell" removed. The result described as the following.

Cell	Call attempts	(CSSR, %)	(DCR, %)	(IHSR, %)	Class (QoS)
001	24	100	0	100	1
002	8	100	0	99	1
003	4	100	0	100	1
004	6	0	1.5	100	0
005	8	100	0	100	1
006	6	0	1.5	100	0
007	6	0	1.5	100	0
008	6	0	1.5	100	0
009	8	100	0	100	1
010	4	100	0	100	1
011	24	100	0	100	1
012	6	0	1.5	100	0
013	4	100	0	100	1
014	6	0	1.5	100	0

Table 4.4, Sample KPI values before feature selection

For example, the following shows the accuracy results before attribute selection.

Experiment 1:

```
0.9298245614035088
              precision    recall  f1-score   support

           0       1.00      0.83      0.91        47
           1       0.89      1.00      0.94        67

    accuracy                           0.93       114
   macro avg       0.95      0.91      0.93       114
weighted avg       0.94      0.93      0.93       114

[[39  8]
 [ 0 67]]
```

Figure 4.1 accuracy results SVM before attribute selection

Call attempts	(CSSR, %)	(DCR, %)	(IHSR, %)	Class (QoS)
24	100	0	100	1
8	100	0	99	1
4	100	0	100	1
6	0	1.5	100	0
8	100	0	100	1
6	0	1.5	100	0
6	0	1.5	100	0
6	0	1.5	100	0
8	100	0	100	1
4	100	0	100	1
24	100	0	100	1
6	0	1.5	100	0
4	100	0	100	1
6	0	1.5	100	0

Table 4.5, Sample KPI v1alues after feature selection

For example, the following shows the accuracy results after attribute selection.

Experiment 2:

```
0.9736842105263158
              precision    recall  f1-score   support

           0       1.00      0.94      0.97        47
           1       0.96      1.00      0.98        67

    accuracy                           0.97       114
   macro avg       0.98      0.97      0.97       114
weighted avg       0.97      0.97      0.97       114

[[44  3]
 [ 0 67]]
```

Figure 4.2, accuracy result in SVM after attribute selection

The second attribute ranked on the list was "Call attempts" and the experimental result after those attribute removed is shown as following figure 4.3.

```
1.0
Classification Report:
              precision    recall  f1-score   support

           0       1.00      1.00      1.00        47
           1       1.00      1.00      1.00        67

    accuracy                           1.00       114
   macro avg       1.00      1.00      1.00       114
weighted avg       1.00      1.00      1.00       114

Confusion Matrix:
[[47  0]
 [ 0 67]]
```

Figure 4.3, accuracy result in SVM after the second attribute (call attempt) removed

The experimental result in support vector machine after the second attribute was removed shows that the accuracy report too high which is 100%. As a result, the researcher decides stop feature selection process before call attempt attribute is selected. Because the accuracy report is, 100% and this may over fit the model. However, the removal of irrelevant features may help learn a better model; removal of some features may lead to over fit the model and leads to affect the performance of the model.

Finally, the best predictors selected for building the model are (CSSR, %), (DCR, %), (IHSR, %), Call attempts.

4.2.3 Detecting Noisy data and Outliers

The data stored in a database may reflect outlier - noise, exceptional case, or incomplete data object and random error in a measure of variable. These incorrect attribute values may be due to data entry problems, faulty data collection, inconsistency in naming convention or technology limitation [1]. Since, the data recorded automatically on the drive test machine; the data objects used in this research paper do not have any outlier or incomplete data.

4.2.4 Data Reduction

Data reduction techniques are used to minimize the volume of the data set in order to make suitable for analysis process [1]. On this research paper, also some data reduction methods are applied including feature selection.

In feature selection, relevant subset of variables was selected by eliminating features with little or no predictive information, which can significantly improve the comprehensibility of the resulting classifier models and often build a model that generalizes better to unseen points [1]. Accordingly, the researchers select variables, which are believed to have direct relationship with the quality of services, total five candidate attributes (4 independent attributes and one class attribute have been selected). The details were listed on the above attribute selection section.

4.2.5 Data Integration

Machine learning and data mining often requires data integration or the merging of data from multiple data sources. In this study there are three data's which collected in different time intervals which are: from 6th February to 10th February 2017, 21st June to 29th June 2018 and 28th December 2018 to 17th January 2019. Merging and integration was performed by using Microsoft excel 2016.

The original dataset has, 4 attribute and 1254 instances for 2009 E.C, 4 attribute and 444 instances for 2009 E.C and 4 attribute and 596 instances for 2011 E.C. By integrating those data sets, the data used for this research consist 4 attributes and 2294 instances.

4.2.6 Setting Class Attribute

In order to classify records into different classes the target attribute selected in this research was Quality of Service "QoS" attribute which has two distinct values called **"Fail"** and **"Success".** Consequently, the attribute was a class/dependent attribute for this particular study.

4.2.7 Data formatting

In data, formatting step after the preprocessing of the initial data the final dataset is in Xls Microsoft Excel format and changed into comma delimited CSV file format to make suitable for the selected python tool. The original and target data set number of records, data type and description of the attributes are presented in the following table 4.6.

No	Attribute name	Data Type	Description
1	(CSSR, %)	Numeric	Assesses the percentage of originating calls that were successfully established by customers
2	(DCR, %)	Numeric	Measure of calls that are prematurely disconnected before end of conversation
3	(IHSR, %)	Numeric	Measures the ability of a customer to talk on the cell phone for a long distance without getting disconnected
4	Call attempts	Numeric	A Demand by a user for a connection to another user
5	Cell	Numeric	Cell site (Base Station) IDs

Table 4.6, Description of predictors and data formatting

After preprocessing tasks has done, the study used 5 attributes and 2294 instances for the entire experimentations of this study. The data size of the data set is 45kb with .csv file extension. The following table 4.7 shows partial view of the dataset used for model building.

Call attempts	(CSSR, %)	(DCR, %)	(IHSR, %)	Class (QoS)
24	100	0	100	1
8	100	0	99	1
4	100	0	100	1
6	0	1.5	100	0
8	100	0	100	1
6	0	1.5	100	0
6	0	1.5	100	0
6	0	1.5	100	0
8	100	0	100	1
4	100	0	100	1
24	100	0	100	1
6	0	1.5	100	0
4	100	0	100	1
6	0	1.5	100	0
6	0	1.5	100	0
6	0	1.5	100	0
4	100	0	100	1
20	100	0	100	1
12	100	0	99	1
4	0	1.5	100	0
6	0	1.5	100	0
8	100	50	90.91	0
6	0	1.5	100	0

Table 4.7, Sample data used for experimentation (labeled "class QoS" 1 & 0 is for success & failed respectively)

4.3 Experiment Design

Before building a model, we need to generate a procedure or mechanism to test the model's quality and validity. For instance, in supervised machine learning tasks such as classification, it is common to use classification accuracy measure as quality measures for machine learning models. Besides, other standard measure including precision, recall and F1-measure are available. Therefore, the test design specifies that the dataset should be separated into training, test set, builds the model on the training set, and estimates its quality on the separate test set.

The process of building predictive models requires a well-defined training and validation protocol in order to ensure that most accurate and robust prediction. In this research, 2294 instances are used for training and testing. Python, jupyter notebook software has used to set up and measure the quality, validity and test of the selected model.

For purpose of this study, k-fold (10-folds) cross validation and percentage split test options are used because of its relatively low bias and variations. Accordingly, the datasets are randomly partitioned equally into ten parts. Hence, 90% of the dataset is for training and 10% for testing for former and the dataset are partitioned in to percentages (20-80) splits option meaning 80% of the dataset for training and remaining for testing). Moreover, to build the model of this research 4 independent and 1 dependent variables or attributes are used.

4.3.1 Selecting Modeling Technique

Selecting appropriate model depends on machine learning goals. Consequently, to attain the objectives of these research three classification techniques has been selected for model building. The analysis was performed using python. Among the different available classification algorithms KNN, SVM and Logistic Regression were used for experimentation of this study. The researcher selected the above algorithms, easy of understanding and interpretation of the result of the model. These algorithms become best matches for this study not only by their pros and cons but also by different referring the algorithm selection framework to choose the best algorithm.

In the selection of a suitable ML algorithm for data analysis many aspects should be taken into account, and most of the times selecting an algorithm only on the base of the promised accuracy or computational speed leads to unsatisfactory results [56].

[56], Proposed a selection framework that works on two different layers, each one linked to a different aspect of the analysis. This would guide the user in the selection of the ML algorithms suitable for the analysis of a specific dataset. The first layer of the ML algorithm selection is based on the presence of labels in the dataset and on the scope of analysis (i.e. learning activity). In this way, the user is guided towards Supervised or Unsupervised Learning algorithms. In the second layer, four more drivers guide the user in the identification of proper ML algorithms. The drivers have been identified after a literature review of ML application cases.

Algorithm selection process also associated to multiple drivers and dataset characteristics. Some of the drivers are listed as the following

Data Type: The ML algorithms are built under specific assumptions; each one is usually informed to work with specific types of data (binary, discrete, categorical, and continuous).

Scalability: The scalability of an algorithm measures the growth of its time complexity in relation to the growth of the problem size. It measures the capacity of an algorithm to handle big inputs.

Robustness to Outliers: It is defined as the ability of a ML algorithm to deal with the presence of data not belonging to the analyzed sample. If an algorithm is robust to outliers and noise, its performance is not affected by their presence.

Response Type: It is defined as the outcome of the analysis. As for the Data Type driver, there are different possible types of response for the analysis. As for the Data Type driver, the possible Response Types are binary, discrete, categorical and continuous.

The researcher identified three ML algorithms suitable for their scope (drivers Learning and Learning Activity) and the dataset characteristics (drivers Data Type, Scalability, Robustness to Outliers/Noise and Response Type).

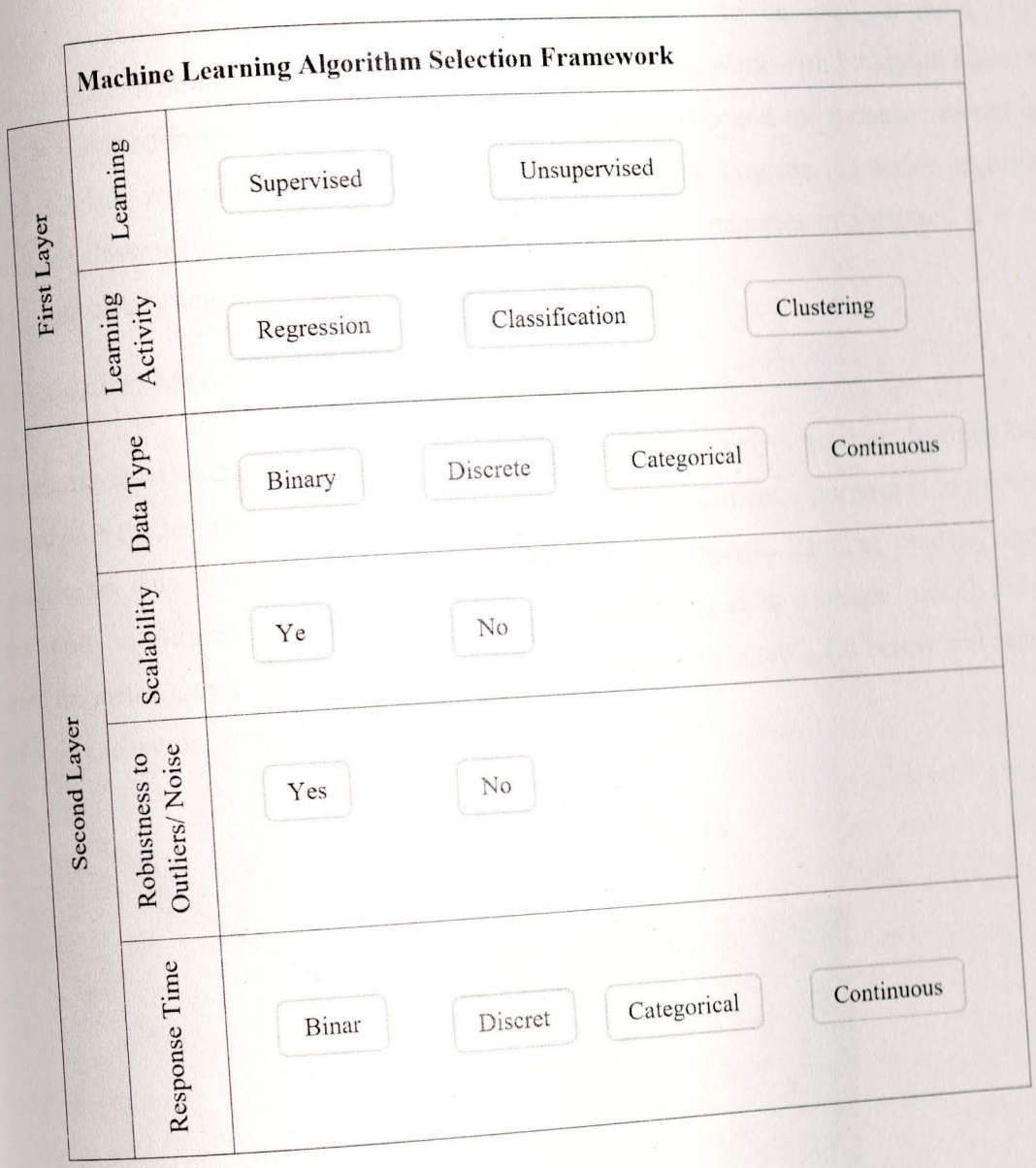


Figure 4.8, Machine Learning Algorithm Selection Framework (Adopted [56])

There are various Supervise machine learning algorithms like Decision trees, Support vector machine (SVM), K-nearest neighbor, K-means clustering, Naive Bayes, Random Forest etc. The performance of these algorithms is depending on the data set provided. So before deploying any model, selection of particular algorithm play very important role in the performance of the model. Algorithm selection processes was done based on the best fit for the problem, best fit for the data and by referring related literatures suggestions.

The SVM is suitable for binary data, but also discrete data can be used as input. High dimensional data can be managed easily. K-Nearest Neighbor can work with binary and discrete variables, but its performance is strongly affected by the data size and the presence of outliers and noise. Logistic regression is suited for binary classifications. Logistic regression algorithm calculates the class membership probability for one of the two categories in a dataset. It is best suited for data clearly separated by a single, linear boundary [56].

4.3.2 Evaluation Metrics

Performance evaluation metrics are used to determine how effectively a machine learning model performed with the test data was given. A learning model's fundamental purpose is to generalize successfully on data that has never been seen before. On specific learning models, specific metrics must be utilized, and not all metrics may be employed in a single model [58]. To compute the metrics, this study use a confusion matrix as shown in table 4.7 below and accuracy as the main evaluation for the developed models.

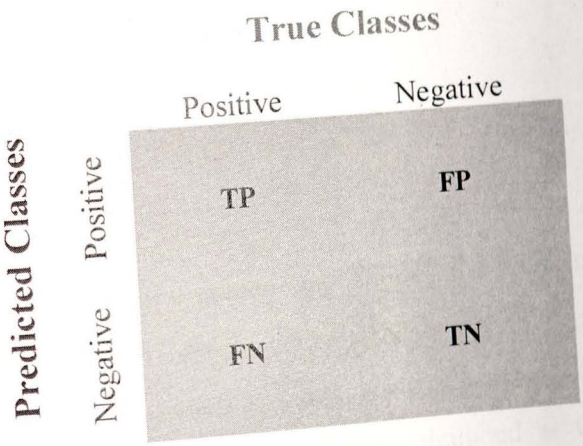


Table 4.7: Confusion matrix (Adopted from [58])

- **TP (True Positive):** The actual value was positive and the model predicted a positive value
- **FP (False Positive)**: Your prediction is positive, and it is false.
- **FN (False Negative):** Your prediction is negative, and result it is also false
- **TN (True Negative):** The actual value was negative and the model predicted a negative value

Accuracy: Accuracy is a measure for how many correct predictions our model. Made for the complete test dataset. Accuracy is a good basic metric to measure the performance of the model [58]. This study used accuracy as the main metric to evaluate the developed models. Accuracy is calculated by dividing the correct predictions to the overall prediction value.

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}$$

Precision: Precision tells us how many of the correctly predicted case actually turned out to be positive. This would determine whether our model is reliable or not. Precision is the ratio of correct prediction to the sum of true and false positive prediction.

$$\text{Precision} = \frac{TP}{TP+\text{FP}}$$

Recall: Recall tells us how many of the actual positive cases we were able to predict correctly with our model. Recall is the ratio of correct prediction to the sum of true positive and false negative prediction.

$$\text{Recall} = \frac{TP}{TP+FN}$$

F1-Score: When we try to increase the precision of model, the recall grows down and vice-versa. F1-Score is a harmonic mean of Precision and Recall and so it gives a combined idea about these two metrics. It is maximum when precision is equal to recall. The F-score is a way of combining the precision and recall of the model. It is calculated as the harmonic mean of precision and recall.

$$\text{F1-Score} = \frac{1}{\frac{1}{recall}+\frac{1}{precision}}$$

4.4 Running Experiments

As it is stated above, the study applied two methods for running different kinds of experiments namely, 10 fold cross validation and percentage splits. Based on the experimental design

establishing a model to be developed is very important to see the model result and analysis of each result, to compare the result of one model with the previous one and finally help us to find out the outperforming model based on criteria of evaluation. Consequently, for both of the methods (10 fold cross validation and percentage splits) following scenarios has been done for each of three-selected model with python tool.

Scenario #1: *Performance result before feature selection*

Scenario #2: *Performance result after feature selection*

Scenario #3: *Performance result with 30 - 70 Percentage split (in KNN, SVM & LR)*

Scenario #4: *Performance result with 20 - 80 Percentage split (in KNN, SVM & LR)*

Scenario #5: *Performance result with 10-Fold cross validation (in KNN, SVM & LR)*

The researcher has been selected the above five scenarios by considering their easy building, understanding and interpretation of the model they generate, suitable for the experimental data set and literature supports.

4.4.1 Model building using K-Nearest Neighbors

K - Nearest Neighbor is a supervised learning algorithm where the result of new instance query is classified based on majority of k-nearest neighbor category [41]. The K-Nearest Neighbor (KNN), is one of the very powerful algorithm of machine-learning algorithms is used in this research.

KNN classification used to classify data sets in to two labels based on the label of majority of its neighbors. The KNN algorithm assumes that similar things exist in proximity. The quote "birds of the same feather flock together" better explains the KNN classification. A nonparametric classification method classifies data sets based on learning from training data sets [6].

4.4.1.1 How KNN works

KNN is a supervised learning algorithm. A labeled training dataset is provided where the data points are categorized into various classes, so that class of the unlabeled data can be predicted. In

Classification, different characteristics determine the class to which the unlabeled data belongs. KNN is mostly used as a classifier. It is used to classify data based on closest or neighboring training examples in a given region. This method is used for its simplicity of execution and low computation time [61]. According to the nearest neighbor technique, the new unlabeled data is classified by determining which classes its neighbors belong to. KNN algorithm utilizes this concept in its calculation. In case of KNN algorithm, a particular value of K is fixed which helps us in classifying the unknown tuple.

When a new unlabeled tuple is encountered in the dataset, KNN performs two operations: First, it analyzes the K points closest to the new data point, i.e. the K nearest neighbors. Second, using the neighbour's classes, KNN determines as to which class the new data should be classified into. When some new data is added, it classifies the data accordingly.The KNN working can be explained on the basis of the below algorithm:

Step-1: Select the number K of the neighbours

Step-2: Calculate the distance (Euclidean) between the query-instance and all the training samples

Step-3: Sort the distance and determine nearest neighbors based on the K^{th} minimum distance

Step-4: Gather the category of the nearest neighbors

Step-5: Use simple majority of the category of nearest neighbors as the prediction value of the query instance

There is no particular way to determine the best value for "K", so we need to try some values to find the best out of them. The most preferred value for K is 5. A very low value for K such as K=1 or K=2, can be noisy and lead to the effects of outliers in the model. Large values for K are good, but it may find some difficulties (Antony Christopher [62]).

Advantage	Disadvantage
Simple to implement and intuitive to understand	High prediction complexity for large datasets: Not great for large datasets, since the entire training data is processed for every prediction.

n learn non-linear decision boundaries when used r classification and regression. Can came up with a ighly flexible decision boundary adjusting the value K.	Higher prediction complexity with higher dimensions: The prediction complexity in supervised learning gets higher for higher dimensional data
Constantly evolves with new data: Since there is no explicit training step, as we keep adding new data to the dataset, the prediction is adjusted without having to retrain a new model.	KNN Assumes equal importance to all features: Since KNN expects points to be close in ALL dimensions, it might not consider points that are close in several dimensions, though farther away in a few favorably.
Single Hyper-parameters: There is a single hyper-parameter, the value of K. This makes hyper-parameter tuning easy.	Sensitive to outliers: A single mislabeled example can change the class boundaries. This could specially be a bigger problem for larger dimensions
Choice of distance metric: There are many distance metrics to choose from. Some popular distance metrics used are Euclidean, Manhattan, Minkowski, hamming distance and so on.	

Table 4.8, Pros and cons of KNN

Experiment 3:

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

In this experiment, we have done preprocessing and experimental works using python to use dataset and optimize the data. An analytical task has been done using KNN algorithm. The objective of this work is to predict quality of service of 2G GSM network based on different

parameters. After experiment has been done through KNN, result shows as on the following table 4.9.

```
0.9869281045751634
Classification Report:
              precision    recall  f1-score   support

           0       1.00      0.97      0.98       195
           1       0.98      1.00      0.99       264

    accuracy                           0.99       459
   macro avg       0.99      0.98      0.99       459
weighted avg       0.99      0.99      0.99       459

Confusion Matrix:
[[189   6]
 [  0 264]]
```

Figure 4.4, Performance result for KNN algorithm with 20- 80 Percentage split

K-Nearest Neighbor (Algorithm)	Accuracy	Precision	Recall	F1 - Score
Experiment 3: 70% train, 30% test	95.930	96	96	96
Experiment 4: 80% train, 20% test	99.692	99	99	99
Experiment 5: 10 - fold Cross Validation	99.825			

Table 4.9, Performance result for KNN

As shown in the above table 4.9, the experimentation has performed in three scenarios by using percentage split test mode at the split of 70% - 30%, 80% - 20% train test and 10-cross validation. Hence, out of the total three experimental models we have different results in the accuracy of the model. As a result, model 5 had the greatest performance, which selected as a best model for k- nearest neighbor algorithm.

The overall performance of the selected model is measured using confusion matrix evaluation technique. Accordingly, from three experiments of KNN experiment 5 perform better which has scored an accuracy of 98.825%. This shows that out of 2296 instances 2,291 (99.825%) instances were correctly classified while 5 (0.8%) instances were incorrectly classified.

4.4.2 Model Building Using Support Vector Machine (SVM)

SVMs find the hyper plane having one dimension less than the original dimensionality of the vectors separating the two classes. The target of the separation is to maximizing the distance of the elements of the classes from the hyper plane on the different sites of it. The closest class element from the two classes is called support vectors (Hamel, et al., 2009) [53].

One of the most effective tools of the SVM is using kernel functions. The idea to ensure higher-class separation capability is to transform the input space into another space having usually higher dimensionality. This space is called as feature space. When an appropriate transformation is found for the problem analyzed, typically, it results better modeling accuracy and usually it results no significant increase in computational time. Another very important feature of SVMs is that the target function of their training for building up its kernel is quadratic and convex having no local but a global extreme [54]. These features and their promising applications result that SVMs are very popular in machine learning applications.

SVM were further developed and extended to handle much more complex assignments, e.g., multiclass classification even if when the classes are not linearly separable. In this case the target of the SVM is to minimize the number of misclassified class elements together with the maximization of the distance between the separating hyper plane and the support vectors [53].

Based on the promising results in classification assignments SVMs were extended to realize also estimation tasks. Similar to the classification their estimation capabilities are considered also successful even if the dimensionality of the input space is very high (Hamel, et al., 2009). The input-output configuration strongly influences the accuracy of the developed model especially if dependencies between parameters are non-invertible [42].

4.4.2.1 Haw SVM works

The best way to understand the SVM algorithm is by focusing on its primary type, the SVM classifier. The idea behind the SVM classifier is to come up with a hyper-lane in an N-dimensional space that divides the data points belonging to different classes. However, this hyper-pane is chosen based on margin as the hyperplane providing the maximum margin between the two classes is considered. These margins are calculated using data points known as Support Vectors. Support Vectors are those data points that are near to the hyper-plane and help in orienting it.

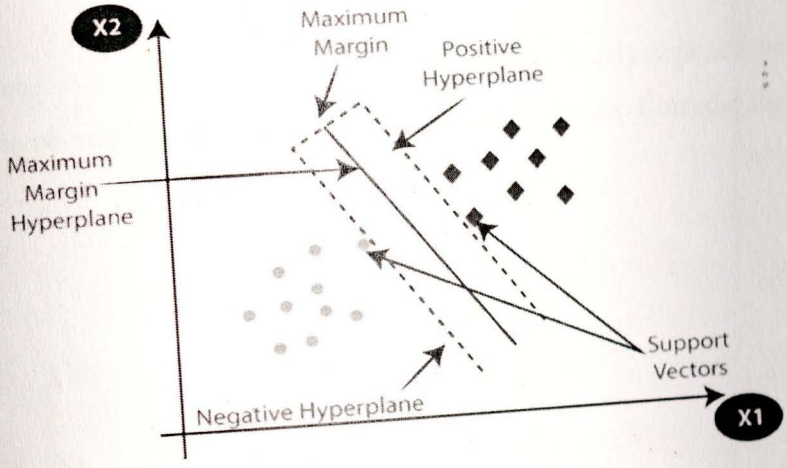


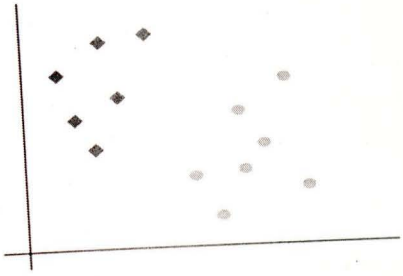
Figure 4.5, linearly separable data points

Support vector machine can be one of the twos

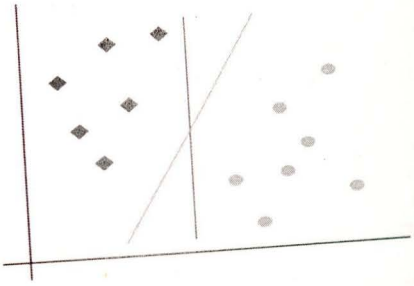
- **Linear SVM:** Linear SVM is used for linearly separable data, which means if a dataset can be classified into two classes by using a single straight line, then such data is termed as linearly separable data, and classifier is used called as Linear SVM classifier.
- **Non-linear SVM:** Non-Linear SVM is used for non-linearly separated data, which means if a dataset cannot be classified by using a straight line, then such data is termed as non-linear data and classifier used is called as Non-linear SVM classifier.

If the functioning of SVM classifier is to be understood mathematically then it can be understood in the following ways

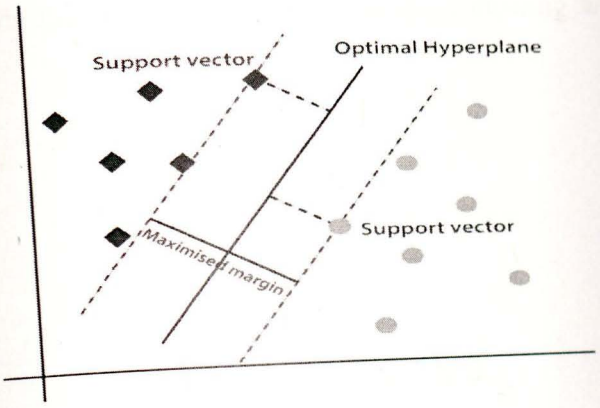
Linear SVM: The working of the SVM algorithm can be understood by using an example. Suppose we have a dataset that has two tags (green and blue), and the dataset has two features x1 and x2. We want a classifier that can classify the pair(x1, x2) of coordinates in either green or blue. Consider the below image:



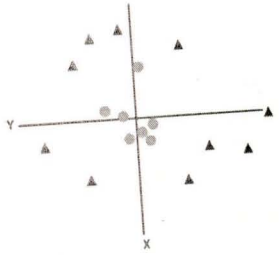
So as it is 2-d space so by just using a straight line, we can easily separate these two classes. However, there can be multiple lines that can separate these classes. Consider the below image:



Hence, the SVM algorithm helps to find the best line or decision boundary; this best boundary or region is called as a hyperplane. SVM algorithm finds the closest point of the lines from both the classes. These points are called support vectors. The distance between the vectors and the hyperplane is called as margin. Moreover, the goal of SVM is to maximize this margin. The hyperplane with maximum margin is called the optimal hyperplane.

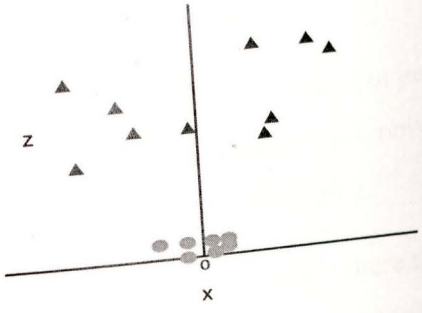


Non-Linear SVM: If data is linearly arranged, then we can separate it by using a straight line, but for non-linear data, we cannot draw a single straight line. Consider the below image:

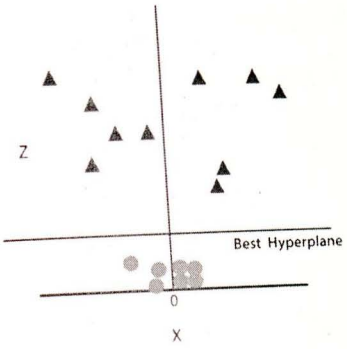


So to separate these data points, we need to add one more dimension. For linear data, we have used two dimensions x and y, so for non-linear data, we will add a third dimension z. It can be calculated as: $z=x^2+y^2$

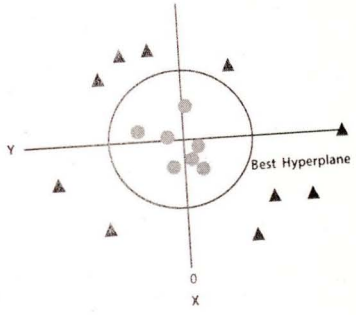
By adding the third dimension, the sample space will become as below image:



So now, SVM will divide the datasets into classes in the following way. Consider the below image:



Since we are in 3-d Space, hence it is looking like a plane parallel to the x-axis. If we convert it in 2d space with z=1, then it will become as:



Hence, we get a circumference of radius 1 in case of non-linear data.

Advantage	Disadvantage
SVM works relatively well when there is a clear margin of separation between classes.	SVM algorithm is not suitable for large data sets.
SVM is more effective in high dimensional spaces.	SVM does not perform very well when the data set has more noise i.e. target classes are overlapping.
SVM is effective in cases where the number of dimensions is greater than the number of samples.	In cases where the number of features for each data point exceeds the number of training data samples, the SVM will underperform.

SVM is relatively memory efficient	As the support vector classifier works by putting data points, above and below the classifying hyperplane there is no probabilistic explanation for the classification.

Table 4.10, pros and cons of SVM algorithm

```
0.9803921568627451
              precision    recall  f1-score   support

           0       1.00      0.95      0.98       195
           1       0.97      1.00      0.98       264

    accuracy                           0.98       459
   macro avg       0.98      0.98      0.98       459
weighted avg       0.98      0.98      0.98       459

[[186   9]
 [  0 264]]
```

Figure 4.5, Performance result for SVM algorithm with 20 - 80 Percentage split

After experiments done through SVM, result shows as the following table 4.11.

Support Vector Machine (Algorithm)	Accuracy	Precision	Recall	F1 - Score
Experiment 6: 70% train 30% test	98.2558	98	98	98
Experiment 7: 80% train 20% test	98.0392	98	98	98
Experiment 8: 10 - fold Cross Validation	99.2013			

Table 4.11, experimental result in SVM

As shown in the above table 4.11, the experimentation has performed in three scenarios by using percentage split test mode at the split of 70% - 30%, 80% - 20% train test and 10-cross validation. Hence, out of the total three experimental models we have different results in the

accuracy of the model. As a result, model 8 had the greatest performance, which selected as a best model for Support Vector Machine algorithm.

The overall performance of the selected model is measured using confusion matrix evaluation technique. Accordingly, the model has scored an accuracy of 99.2013% on experiment 8. This shows that out of 2296 instances 2,277 (99.2013%) instances were correctly classified, while 19 (0.83%) instances were incorrectly classified.

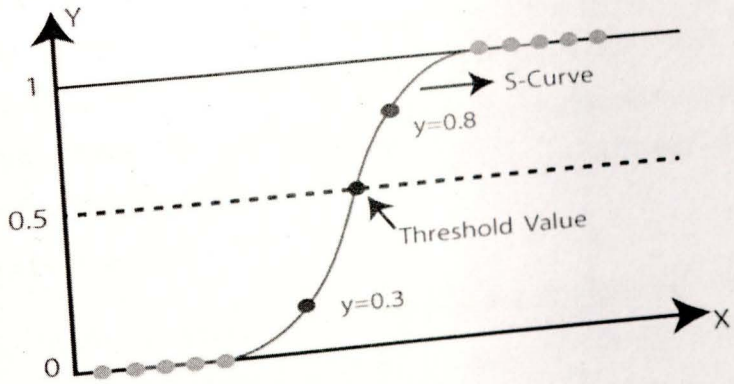
4.4.3 Model Building using Logistic Regression

Logistic regression is another example of supervised learning. It is used to calculate or predict the probability of a binary (yes/no) event occurring. Since we have two possible outcomes to this question - "Success" in network quality or "Fail" quality of mobile network - this is called binary classification. In logistic regression, the outcome is continuous and can be any possible value. However, in the case of logistic regression, the predicted outcome is discrete and restricted to a limited number of values [50].

Logistic regression is a powerful tool, allowing multiple explanatory variables being analyzed simultaneously, meanwhile reducing the effect of confounding factors [50].

4.4.3.1 Haw Logistic Regression works

Logistic regression predicts the output of a categorical dependent variable. Therefore, the outcome must be a categorical or discrete value. It can be either Yes or No, 0 or 1, true or False, etc. but instead of giving the exact value as 0 and 1, it gives the probabilistic values which lie between 0 and 1.



Logistic regression uses the concept of predictive modelling as regression; therefore, it is called logistic regression, but is used to classify samples; therefore, it falls under the classification algorithm.

Logistic regression models are generally used for predictive analysis for binary classification of data. However, they can also be used for multi-class classification. Logistic regression models can be classified into three main logistic regression analysis categories. They are:

Binary Logistic Regression Model: This is one of the most widely used logistic regression models, used to predict and categorize data into either of the two classes. For example, a patient can have cancerous cells, or they cannot. The data can't belong to two categories at the same time.

Multinomial Logistic Regression Model: The multinomial logistic regression model is used to classify the target variable into multiple classes, irrespective of any quantitative significance. For instance, the type of food an individual is likely to order based on their diet preferences - vegetarians, non-vegetarians, and vegan.

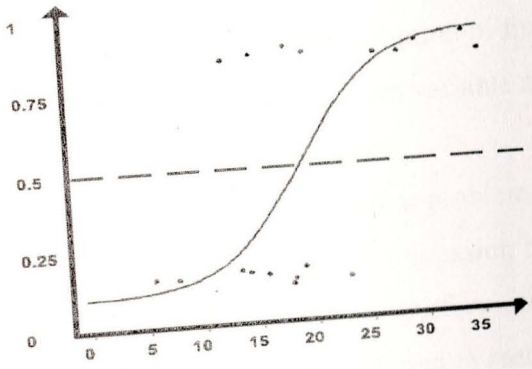
Ordinal Logistic Regression Model: The ordinal logistic regression model is used to classify the target variable into classes and also in order. For example, a pupil's performance in an examination can be classified as poor, good, and excellent in a hierarchical order. Thus, we can see that the data is not only classified into three distinct categories, but each category has a unique level of importance.

Machine learning generally involves predicting a quantitative outcome or a qualitative class. The former is commonly referred to as a regression problem. In the scenario of linear regression, the input is a continuous variable, and the prediction is a numerical value. When predicting a qualitative outcome (class), the task is considered a classification problem. Examples of classification problems include predicting what products a user will buy or if a target user will click on an online advertisement.

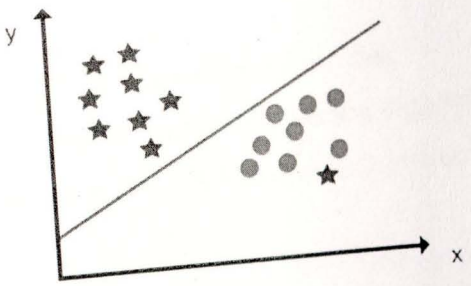
Not all algorithms fit cleanly into this simple dichotomy, though, and logistic regression is a notable example. Logistic regression is part of the regression family as it involves predicting

outcomes based on quantitative relationships between variables. However, unlike linear regression, it accepts both continuous and discrete variables as input and its output is qualitative. In addition, it predicts a discrete class such as "Yes/No" or "Customer/Non-customer".

In practice, the logistic regression algorithm analyzes relationships between variables. It assigns probabilities to discrete outcomes using the sigmoid function, which converts numerical results into an expression of probability between 0 and 1.0. Probability is either 0 or 1, depending on whether the event happens or not. For binary predictions, you can divide the population into two groups with a cut-off of 0.5. Everything above 0.5 is considered to belong to group A, and everything below is considered to belong to group B.



A hyperplane is used as a decision line to separate two categories (as far as possible) after data points have been assigned to a class using the Sigmoid function. The class of future data points can then be predicted using the decision boundary.



Advantage	Disadvantage
Logistic regression is easier to implement, interpret, and very efficient to train.	If the number of observations is lesser than the number of features, Logistic Regression should not be used; otherwise, it may lead to overfitting.
It makes no assumptions about distributions of classes in feature space.	It constructs linear boundaries.
It is very fast at classifying unknown records.	The major limitation of Logistic Regression is the assumption of linearity between the dependent variable and the independent variables.
Good accuracy for many simple data sets and it performs well when the dataset is linearly separable.	Non-linear problems can't be solved with logistic regression because it has a linear decision surface. Linearly separable data is rarely found in real-world scenarios.
It can interpret model coefficients as indicators of feature importance.	It is tough to obtain complex relationships using logistic regression. More powerful and compact algorithms such as Neural Networks can easily outperform this algorithm.

Table 4.12, Pros and cons of Logistic regression algorithm

Model building was performed using the logistic regression algorithm with Percentage split test mode 80% train and 20% test option in each experiment is shown as the following table 4.13.

```
0.9985486211901307
              precision    recall  f1-score   support

           0       1.00      1.00      1.00       377
           1       1.00      1.00      1.00       312

    accuracy                           1.00       689
   macro avg       1.00      1.00      1.00       689
weighted avg       1.00      1.00      1.00       689

[[376   1]
 [  0 312]]
```

Figure 4.6 Example performance result for Logistic Regression algorithm

Logistic Regression (Algorithm)	**Accuracy**	**Precision**	**Recall**	**F1 - Score**
Experiment 9: 70% train 30% test	99.8248	100	100	100
Experiment 10: 80% train 20% test	99.8248	100	100	100
Experiment 11: 10 - fold Cross Validation	99.8548			

Table 4.13, experimental result in LR

As shown in the above table 4.13, the experimentation has performed in three scenarios by using percentage split test mode at the split of 70% - 30%, 80% - 20% train test and 10-cross validation. Hence, out of the total three experimental models we have different results in the accuracy of the model. As a result, model 11 had the greatest performance, which selected as a best model for Logistic regression Machine algorithm.

The overall performance of the selected model is measured using confusion matrix evaluation technique. Accordingly, the model has scored an accuracy of 99.8548% in experiment 11. This shows that out of 2296 instances 2,292 (99.854%) instances were correctly classified, while 4 (0.17%) instances were incorrectly classified.

4. 5 Result Discussion and Comparison

In order to select a machine-learning model for classification tasks in the context of this study, it is necessary to evaluate the selected best model from KNN, SVM and logistic regression. Each model basically evaluated based on their classification accuracy results.

The researcher created predictive models for quality of service of 2G GSM mobile network, by using Drive test dataset collected from Ethiopian Communications Authority (ECA) with three machine-learning algorithms. The best results were achieved through logistic regression. Generally, most of the algorithms used achieved model accuracy greater than 95%. The best algorithm (logistic regression) produced an Accuracy of 99.854%.

The methods and concepts for building predictive models for use in quality of services have been well described. In comparison with results from logistic regression on the same dataset, the models created using support-vector machines and k-nearest neighbor produced slightly better results. In the first, models created using logistic regression on data from QoS produced better results (Accuracy = 99%) than models created using support-vector machines (Accuracy = 97.36%) and k-nearest neighbor (Accuracy = 98.24%).

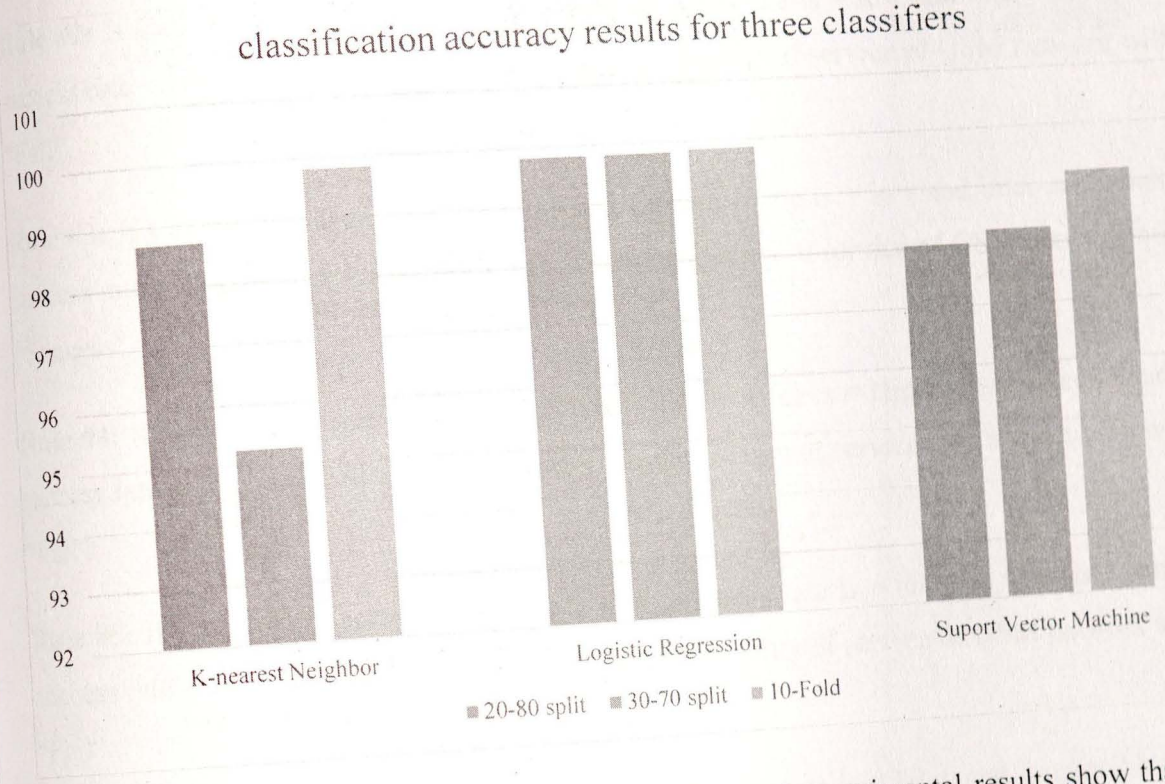
	K-nearest Neighbor	Support Vector Machine	Logistic Regression
Percentage split test mode 70% train			
Accuracy	95.930	98.255	99.824
Precision	96	98	100
Recall	96	98	100
F1 - Score	96	98	689
Percentage split test mode 80% train			

Accuracy	98.692	98.039	99.824
Precision	99	98	100
Recall	99	98	100
F1 - Score	459	459	689
10 –fold cross validation test mode			
Accuracy	99.825	99.2013	99.8548

Table 4.14, Summery and Comparison of classification accuracy for the three classifiers

As indicated in Table 4.16, based on the classification accuracy, Logistic regression classifier shows the highest classification accuracy. It accurately classifies 99.854% of KPI records in to their right QoS label. Consequently, the model built on logistic regression is the best model to classify KPI instances in to their correct class when it is compared to the other two classifiers.

On the above, experimentations either 10 fold or percentage split test options or their respective models were built, evaluated and compared. Finally, a model built using logistic regression mode was taken as a final model for this particular study based on its performance relative to the other two (SVM and KNN) classifiers.



The findings the study based on the literatures review and experimental results show that Call Setup Success Rate, Handover success rate, dropped call rate and call attempt are the major determinant factor of Quality of Service of GSM network. On the experiment, the research can observed on the performance results that the three key performance indicators are relevant for model development.

4.3.5 Evaluation

The selected model generates different rules. Among them, a sample of discovered rules was presented for discussion is as follows. The semantics of these rules with the real environment is confirmed by domain experts.

Rule #1: If Call Setup Success Rate = High and Call drop rate is = low and inter-cell handover success rate = High and call attempt = Low, Then the Quality of service of GSM network will be "Success".

Rule #2: If Call Setup Success Rate = Low and Call drop rate is = low and inter-cell handover success rate = High and call attempt = Low, Then the Quality of service of GSM network will be "Fail".

Rule #3: If Call Setup Success Rate = High and Call drop rate is = low and inter-cell handover success rate = High and call attempt = High, Then the Quality of service of GSM network will be "Success".

Rule #4: If Call Setup Success Rate = High and Call drop rate is = High and inter-cell handover success rate = High and call attempt = Low, Then the Quality of service of GSM network will be "Fail".

Rule #5: If Call Setup Success Rate = High and Call drop rate is = low and inter-cell handover success rate = Low and call attempt = Low, Then the Quality of service of GSM network will be "Fail".

Rule #6: If Call Setup Success Rate = Low and Call drop rate is = High and inter-cell handover success rate = Low and call attempt = High, Then the Quality of service of GSM network will be "Fail".

Rule #7: If Call Setup Success Rate = High and Call drop rate is = Low and inter-cell handover success rate = Low and call attempt = High, Then the Quality of service of GSM network will be "Success".

Rule #8: If Call Setup Success Rate = High and Call drop rate is = High and inter-cell handover success rate = Low and call attempt = High, Then the Quality of service of GSM network will be "Success".

Rule #9: If Call Setup Success Rate = Low and Call drop rate is = High and inter-cell handover success rate = High and call attempt = High, Then the Quality of service of GSM network will be "Fail".

Rule #10: If Call Setup Success Rate = Low and Call drop rate is = Low and inter-cell handover success rate = High and call attempt = Low, Then the Quality of service of GSM network will be "Fail".

CHAPTER FIVE

CONCLUSION AND RECOMMENDATIONS

The previous chapter discussed about issues in the building of the Machine-learning model. In this chapter, the researcher concludes the overall work of the study and provides recommendation for other problems to be investigated.

5.1 Conclusion

The aims of this project were to investigate towards the development of an optimal model that predicts quality of service (QoS) of 2G GSM mobile network through machine learning techniques by using drive test dataset. The research used and followed experimental research methodology to investigate the stated problem. The methodology has been followed while undertaking the experimentation. The data for this research was taken from Ethiopian Communication Authority in quality of service department. These data are collected in three different periods from February, 2017 to January 17, 2019 a total of 2,294 records and 6 attributes. After preprocessing 2294 records and 4 attributes and one class attribute were used for the entire experiment and building predictive models.

This study tries to analyze the KPI data indicating the general quality of service (QoS) in a mobile network by applying different machine learning technologies. Among the KPIs studied, Call Success Rate (CSR), Call Drop Rate (CDR) and Inter cell Handover Success Rate (IHSR) could measure an end-to-end performance of a mobile network. KNN, SVM and Logistic Regression algorithms were implemented in python to build and compare classifier models. Classifiers were tested on 10 fold cross validation and percentage split test options by varying percentage of train and test splits.

The finding of this research indicates that a 2G mobile network located in Addis Ababa city and its surrounding are susceptible to failure in network quality. Low call set up success rate is more exposure to failure occurrences. Moreover, the finding interpreted above Call Setup Success Rate, Handover success rate, Dropped call rate and call attempt are the major determinant factor of Quality of service of 2G GSM Network. Following research question number one (RQ1), the

study indicates Call Setup Success Rate, Handover success rate, Dropped call rate and call attempt attributes are relevant predictors to determine the QoS of GSM network.

However, in three algorithms KNN, SVM and Logistic Regression slightly similar result was attained. As far the researcher try to compare the test options which more suitable for the experimental data by using the those three algorithms, logistic regression with 10 fold cross validation test option results accuracy of 99.854% which performs a better result than the rest experimental results. Following research question number two (RQ2), the study indicates a model build in logistic regression is more suitable to predict and classify the KPI data in to the right level of QoS for 2G mobile voice network.

In general, the study is limited to building a predictive machine learning model to classify the dataset. Based on the results found in this study, we recommend ethiotelecom and ECA to implement such techniques to predict mobile network quality and avoid anomalies throughout the network, which will improve customer satisfaction.

5.2 Recommendation

This Research has been conducted mainly for an academic purpose. However, it revealed the potential applicability of machine learning technique to predict the quality of mobile network for optimizing the ongoing network performance. This research work can contribute a lot towards a comprehensive study in this area in the future, in the context of our country. The results of this study have also shown that machine learning technology particularly the classification technique is well applicable in the efforts of improving Quality of Services.

Hence, based on the findings of this study, the following recommendations forwarded.

- The learning model with the largest size of training sets appears to be the most accurate and consistently delivers a much better and stable results [59]. Performance measured through the study is promising. However, this research was conducted for academic purpose. To deploy in the company with little modification and to come up with more comprehensive models, it is recommended that experimental test to be conducted by the organization with inclusion of many dataset by using large training and testing datasets.

- This research has been attempted to determine the QoS of 2G GSM Network with limited data set and 6 attributes collected from Addis Ababa city and its surroundings. Further researches can also be conducted a research by using datasets collected from all other telecommunication regional districts in Ethiopia.
- Nowadays, business problems become complex and diverse. Thus, the researcher believes that, application of other machine learning techniques (rather than classification) with different algorithms is also a potential research area in GSM cellular network. These might leave a room to conduct further studies.
- This research attempted to determine the QoS of 2G GSM Network. Further researches can also be conduct in other generations (Like 3G and 4G) mobile voice network to determine quality of services.

References

[1]. **Muluken Tigabu,** Application of Data Mining Techniques to Predict Base Transceiver Stations (Bts) Failure Rate: The Case Of ethiotelecom North West Region Gondar District, Master of Science in Information Technology, University of Gondar, 2015.

[2]. **Lulu Deyu**, Data Mining Approach to Analyze Mobile Telecommunications Network Quality of Service: The Case of ethiotelecom, Master of Science in Information science, Addis Ababa University, 2014.

[3]. **Zeneb Kassaw**, Coverage Prediction Based on Spatial Interpolation Techniques: The Case of UMTS Network in Addis Ababa, Masters of Science in Telecom Network Engineering, Addis Ababa University, 2020.

[4]. Yared Alibo Ayiza, Identifying the Reason for Mobile Call Drops Using Data Mining Technology, Master of Science in Computer Science , St. Mary's University , 2018.

[5]. **Menbere Asfaw**, Quality of Experience Model for Addis Ababa Voice Service Using Adaptive Neuro Fuzzy Inference Approach, Masters of Science in Telecom Network Engineering, Addis Ababa University, 2019.

[6]. **Teweldebrhan Mezgebo**, Anomaly Detection of LTE Cells using KNN Algorithms: The Case of Addis Ababa, Masters of Science in Telecom Network Engineering, Addis Ababa University, 2019.

[7]. **Abdulkerim Seid**, Quality of Service Assessment on Fixed-Wireless Broadband Internet Service the case of ethiotelecom, Masters of Science in Telecom Network Engineering, Addis Ababa University, 2019.

[8]. **Addisu Shiferaw**, QoE Assessment Model for Addis Ababa LTE Video Streaming Service Using Machine Learning Techniques, Masters of Science in Telecom Network Engineering, Addis Ababa University, 2020.

[9]. **Charalampos N. Pitas, Konstantina E Chourdaki, Athanasios Panagopoulos** and **Philip Constantinou,** QoS Mining Methods for Performance Estimation of Mobile Radio Networks, Conference Paper , 2010.

[10]. **Hassan Abdulkareem, Abdoulie Momodou Sunkary T ekanyi, Abduljalal Yushau Kassim** and **Ziyaulhaq Muhammad Zakariyya,** Analysis Of A GSM Network Quality Of

Service Using Call Drop Rate And Call Setup Success Rate As Performance Indicators, Article in European Journal of Electrical Engineering, 2020.

[11]. **International Communication Union (ITU)**, Quality of Service and Quality of Experience Regulation, 2017.

[12]. **Ethiopia Communications Authority (ECA)**, Telecommunications Quality of Service Directive, 2021.

[13]. **Agnieszka Ławrynowicz** and **Volker Tresp,** Introducing Machine Learning, Article oznan University of Technology, Poland, 2014.

[14]. **Janne Riihijärvi** and **Petri Mähönen**, Machine Learning for Performance Prediction in Mobile Cellular Networks, IEEE computational intelligence magazine, 2018.

[15]. **Charalampos N. Pitas, Konstantina E Chourdaki, Athanasios Panagopoulos** and **Philip Constantinou,** QoS Mining Methods for Performance Estimation of Mobile Radio Networks, Conference Paper, 2010.

[16]. **Kabir Kadiri** and **Oluwaseun Samuel Lawal**, Assessment of Call Voice Quality of GSM Network Operators in 5 Cities in Kwara State, Article in Journal of Scientific Research and Reports, 2019.

[17]. **Gitanjali Bhutani**, Application of Machine-Learning Based Prediction Techniques in Wireless Networks, Int. J. Communications, Network and System Sciences, 2014.

[18]. **Y aohua Sun, Mugen Peng and Shiwen Mao**, Application of Machine Learning in Wireless Networks: Key Techniques and Open Issues, Article on Institute of Electrical and Electronics Engineers (IEEE), 2018.

[19]. **Fabricio Carvalho de Gouveia** and **Thomas Magedanz,** Quality of Service in Telecommunication Networks, Telecommunication Systems and Technologies - Vol. II. 2014.

[20]. **Jide Julius Popoola** and **Adewale Enoch A reo,** Modeling and Development of a Novel Quality of Service Prediction Model for Global System for Mobile Communications Network using Artificial Neural Networks, Journal of Applied Science & Process Engineering Vol. 7, No. 2, 2020.

[21]. **Mikko Multanen, Kimmo Raivio** and **Pasi Lehtim¨aki**, Hierarchical analysis of GSM network performance data, Helsinki University of Technology Laboratory of Computer and Information Science, 2015.

[22]. **Rajesh Ganesan, B. Vinayagasundaram** and **X. Mercilin Raajin**, Achieving QoS in GSM Network by Efficient Anomaly Mitigation and Data Prediction Model, Conference Paper on IEEE, 2018.

[23]. **Nasser Kimbugwe, Tingrui Pei** and **Moses Ntanda Kyebambe,** Application of Deep Learning for Quality of Service Enhancement in Internet of Things: A Review, Energies School of Computer Science, Xiangtan University, Xiangtan 411105, China, 2021.

[24]. **Anna Corazza, Francesco Isgrò** and **Roberto Prevete**, A machine learning approach for predictive maintenance for mobile phones service providers, Conference Paper, 2017.

[25]. **Pasi Lehtim¨aki**, Data Analysis Methods for Cellular Network Performance Optimization, Dissertations in Information and Computer Science Espoo, Helsinki University of Technology, 2008.

[26]. **Aroussi Sana** and **Abdelhamid Mellouk**, Survey on machine learning-based QoE-QoS correlation models, Article on IEE, 2014.

[27]. **Dem ostenes Zegarra Rodr ıguez, Renata Lopes Rosa** and **Grac¸a Bressan,** Predicting the Quality Level of a VoIP Communication through Intelligent Learning Techniques, Conference paper, The Seventh International Conference on Digital Society, 2013.

[28]. **Segun I. Popoola, Aderemi A. Atayero, Nasir Faruk b and Joke A. Badejo,** Data on the key performance indicators for quality of service of GSM networks in Nigeria, Journal, 2018.

[29]. **Muwawa Jean Nestor Dahj**, Data Mining and Predictive Analytics Application on Cellular Networks to Monitor and Optimize Quality of Service and Customer Experience, Master of degree in Electrical Engineering, University of South Africa November, 2018

[30]. **Yusuf Babatunde Lawal, Kingsley Eghonghon Ukhurebor, Mathew Adefusika Adekoya and Efosa Aigbe**, Quality of Service and Performance Analysis of A GSM Network In Eagle Square, Abuja and Its Environs, Nigeria. Article, International journal of scientific and engineering research, 2016

[31]. **Muhammad Asif Khan**, an Empirical Assessment of Service Quality of Cellular Mobile Telephone Operators in Pakistan, Article in Asian Social Science. 2010.

[32]. **Anton A. Huurdeman,** the Worldwide History of Telecommunications, 2003.

[33]. **Worku Bogale**, A Background Paper on Telecom & Telecom Statistics in Ethiopia, Article, 2005.

[34]. Global System for Mobile Communication (GSM), Article in Asian Social Science. www.iec.org

[35]. **Ajay R. Mishra**, Fundamentals of Cellular Network Planning and Optimization 2G/2.5G/3G... Evolution to 4G, Book volume 1, 2004.

[36]. ECA, Drive Test Report for Addis Ababa, 2011 E.C

[37]. **Meer Zafarullah Noohani** and **Kaleem Ullah Magsi,** A Review Of 5G Technology: Architecture, Security and wide Applications, Article International Research Journal of Engineering and Technology (IRJET), 2020.

[38]. **Jackson Kamiri** and **Geoffrey Mariga,** Research Methods in Machine Learning: A Content Analysis, Article, International Journal of Computer and Information Technology, 2021.

[39]. **Suhang Wang, Huan Liu** and **Jiliang Tang,** Feature Selection, January 2016.

[40]. **A. Jović, K. Brkić** and **N. Bogunović,** A review of feature selection methods with applications, Article Springer Science Business Media, 2016

[41]. **Devendra Prasad, Sandip Kumar Goyal, Dr. Amit Kumar Bindal** and **Avinash Sharma** System Model for Prediction Analytics Using K-Nearest Neighbors Algorithm, Article in Journal of Computational and Theoretical Nanoscience, 2019.

[42]. **Dr. Zsolt János Viharos** and **Krisztián Balázs Kis**, Support Vector Machine (SVM) based general model building algorithm for production control, conference paper, 2011.

[43]. **Zhongheng Zhang,** Model-building strategy for logistic regression: purposeful selection, Jan, Jinhua Municipal Central Hospital, Jinhua Hospital of Zhejiang University, china, 2016.

[44]. **André Rodrigues OliveraI, Valter RoeslerII, Cirano IochpeII, Maria Inês SchmidtIII, Álvaro VigoIV, Sandhi Maria BarretoV, Bruce Bartholow Duncan,** Comparison of machine-learning algorithms to build a predictive model for detecting undiagnosed diabetes - ELSA-Brasil: accuracy study, MSc. IT Analyst, Postgraduate Computing Program, Universidade Federal do Rio Grande do Sul (UFRGS), Porto Alegre (RS), Brazil, 2016.

[45]. **Ms. Lopa J. Vora**, Evolution of Mobile Generation Technology: 1g to 5g and Review of Upcoming Wireless Technology 5g, International Journal of Modern Trends in Engineering and Research (IJMTER) Volume 02, 2015.

[46]. **Adnan Mohsin Abdulazeez** Classification Based on Decision Tree Algorithm for Machine Learning, Article in Journal of Applied Science and Technology. 2021.

[47]. Betelehem Alemayehu Hailu, Mobile Network Congestion Prediction Using Machine Learning: The Case of Ethio Telecom, Master of Science in Computer Science, St. Mary's University, 2022

[48]. Jaydip Sen, Machine Learning - Algorithms, Models and Applications, Book volume 7, 2018.

[49]. Machine learning, www.tutorials.com

[50]. **Sandro Sperandei**, Understanding logistic regression analysis, Article in Biochemia Medica, 2014.

[51] **M.Benisha, R.Thandaiah Prabu, Thulasi Bai** Evolution of Mobile Generation Technology, International Journal of Recent Technology and Engineering (IJRTE), 2019.

[52]. **J E T Akinsola**, Supervised Machine Learning Algorithms: Classification and Comparison, Article, in International Journal of Computer Trends and Technology (IJCTT), 2017

[53]. **Lutz Hamel**, Knowledge Discovery with Support Vector Machines, Book volume 1, 2009.

[54]. **Terrence S Furey** and **David Haussler,** Support Vector Machine Classification and Validation of Cancer Tissue Samples Using Microarray Expression Data, Article in Bioinformatics, 2010.

[55]. **Qiong Liu** and **Ying Wu** Supervised Learning, Article, 2012

[56]. **Sala R, Zambetti M, Pirola F and Pinto R**, How to select a suitable machine-learning algorithm: a feature-based, scope-oriented selection framework, Department of Management, Information and Production Engineering, University of Bergamo, Viale Marconi, 5, Dalmine (BG), 24044, Italy, 2014.

[57]. **Nikita Butakov, Loren Jan Wilson, Wenting Sun** and **Angel Barranco,** Machine learning use cases: how to design ML architectures for today's telecom systems, Article, 2021.

[58]. **Zohreh Karimi**, Confusion Matrix, Article, 2021.

[59]. **A. R. Ajiboye, R. Abdullah-Arshah, H. Qin** and **H. Isah-Kebbe,** Evaluating The Effect of Dataset Size on Predictive Model Using Supervised Learning Technique, Article in International Journal of Computer Systems & Software Engineering, 2015.

[60]. **N. Hoque, D.K. Bhattacharyya** and **J.K. Kalita** MIFS-ND: A mutual information-based feature selection method, Journal, Science Direct 2014.

[61]. **Srishti Verma** and **Aleena Swetapadma** A Brief Review of Nearest Neighbor Algorithm for Learning and Classification, Conference Paper, 2019.

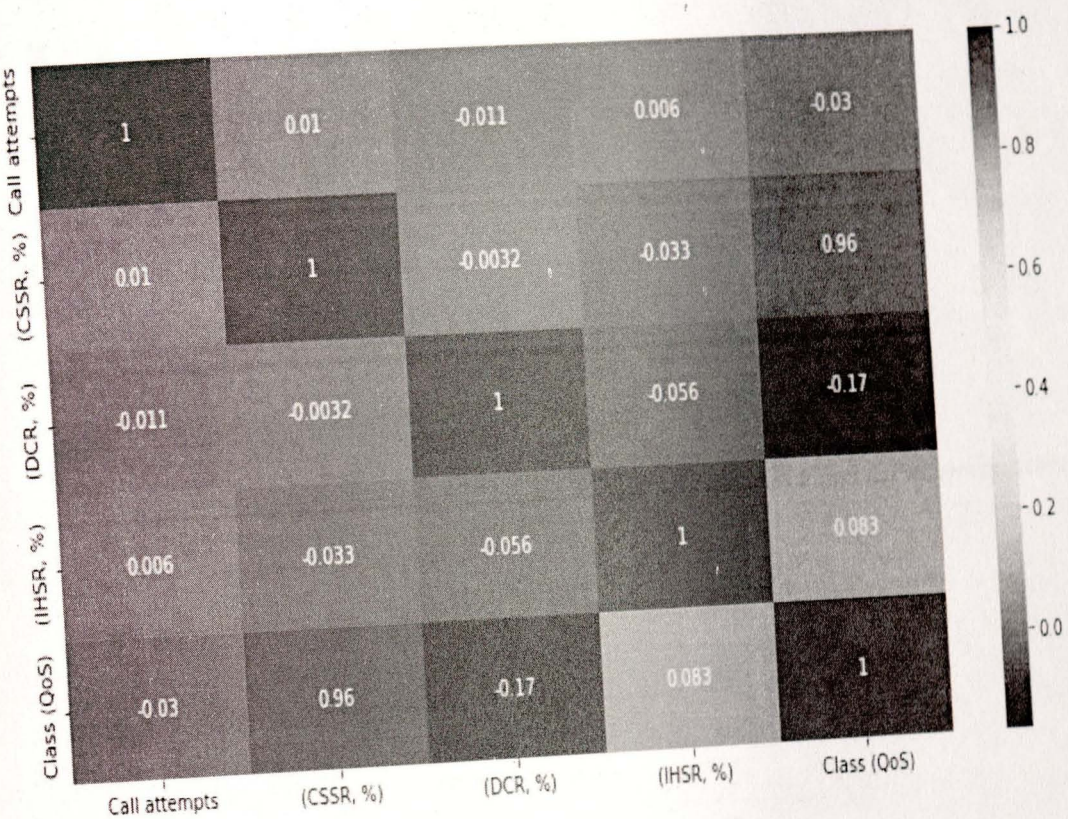
[62]. Antony Christopher, K-Nearest Neighbor, Article, 2021

[63]. Ethio Telecom 2012 EFY (2019/20) Annual Business Performance Summary Report, www.ethiotelecom.et, 2020

Appendix A: Sample original data extracted from drive test machine

Cell	Call attempts	Call Setup Failures	Call Setup OK	Call Setup Success Rate (CSSR, %)	Dropped Call	Dropped Call Rate (DCR, %)	Intercell Handover OK	Intercell Handover Fail	Intercell Handover Success Rate (%)
BCCH-111	24		24	100		0	28		100
BCCH-111_octBSIC-12	8		8	100		0			N/A
BCCH-111_octBSIC-13	4		4	100		0	8		100
BCCH-111_octBSIC-14				N/A		N/A	12		100
BCCH-111_octBSIC-17	8		8	100		0	8		100
BCCH-111_octBSIC-2				N/A		N/A	4		100
BCCH-111_octBSIC-20				N/A		N/A	4		100
BCCH-111_octBSIC-21				N/A		N/A	4		100
BCCH-111_octBSIC-22	8		8	100		0	8		100
BCCH-111_octBSIC-23	4		4	100		0	20		100
BCCH-111_octBSIC-25	24		24	100		0	28		100
BCCH-111_octBSIC-3				N/A		N/A	8		100
BCCH-111_octBSIC-30	4		4	100		0	12		100
BCCH-111_octBSIC-31				N/A		N/A	12		100
BCCH-111_octBSIC-33				N/A		N/A	4		100
BCCH-111_octBSIC-34				N/A		N/A	4		100
BCCH-111_octBSIC-35	4		4	100		0	32		100
BCCH-111_octBSIC-37	20		20	100		0	36		100
BCCH-111_octBSIC-4	12		12	100		0			N/A
BCCH-111_octBSIC-41	4			0		N/A	32		100
BCCH-111_octBSIC-43				N/A		N/A	36		100
BCCH-111_octBSIC-45	8		8	100	4	50	40	4	90.91
BCCH-111_octBSIC-46				N/A		N/A	16		100

Appendix B: Correlation and description between predictors



	Call attempts	(CSSR, %)	(DCR, %)	(IHSR, %)	Class (QoS)
count	2294.000000	2294.000000	2294.000000	2294.000000	2294.000000
mean	6.041412	46.409747	1.490549	99.264303	0.440279
std	3.633268	49.727913	7.906753	5.542967	0.496529
min	1.000000	0.000000	0.000000	0.000000	0.000000
25%	4.000000	0.000000	0.000000	100.000000	0.000000
50%	6.000000	0.000000	1.500000	100.000000	0.000000
75%	6.000000	100.000000	1.500000	100.000000	1.000000
max	48.000000	100.000000	200.000000	100.000000	1.000000

Appendix C: Sample python code snapshots for model development

KNN

In [46]:
```
from sklearn.neighbors import KNeighborsClassifier
classifier = KNeighborsClassifier(n_neighbors = 2)
classifier.fit(x_train, y_train)
```

Out[46]:
```
KNeighborsClassifier
KNeighborsClassifier(n_neighbors=2)
```

In [47]:
```
y_pred = classifier.predict(x_test)
```

In [48]:
```
from sklearn.metrics import classification_report, confusion_matrix, accuracy_score
print(accuracy_score(y_test, y_pred))
print ('Classification Report:')
print (classification_report(y_test, y_pred))
print ('Confusion Matrix:')
print (confusion_matrix(y_test, y_pred))
```

```
0.9824561403508771
Classification Report:
              precision    recall  f1-score   support

           0       1.00      0.96      0.98        47
           1       0.97      1.00      0.99        67

    accuracy                           0.98       114
   macro avg       0.99      0.98      0.98       114
weighted avg       0.98      0.98      0.98       114

Confusion Matrix:
[[45  2]
 [ 0 67]]
```

SVM

In [17]:
```
from sklearn.metrics import classification_report, confusion_matrix, accuracy_score
from sklearn.svm import SVC
classifier = SVC()
classifier.fit(x_train, y_train)
y_pred = classifier.predict(x_test)
print("Accuracy score:")
print(accuracy_score(y_test, y_pred))
print (classification_report(y_test, y_pred))
print (confusion_matrix(y_test, y_pred))
```

```
Accuracy score:
0.997384481255449
              precision    recall  f1-score   support

           0       1.00      1.00      1.00      1284
           1       0.99      1.00      1.00      1010

    accuracy                           1.00      2294
   macro avg       1.00      1.00      1.00      2294
weighted avg       1.00      1.00      1.00      2294

[[1278    6]
 [   0 1010]]
```

In [ ]:

LR

In [6]:
```
from sklearn.model_selection import train_test_split
x_train, X_test, y_train, y_test = train_test_split(x, y, test_size=0.3, random_state=0)
from sklearn.linear_model import LogisticRegression
from sklearn import metrics
logreg = LogisticRegression()
logreg.fit(x_train, y_train)
```

Out[6]:
```
LogisticRegression()
```

In [7]:
```
y_pred = logreg.predict(X_test)
```

In [ ]:

In [8]:
```
from sklearn.metrics import classification_report, confusion_matrix, accuracy_score
from sklearn.metrics import classification_report, confusion_matrix
print(accuracy_score(y_test, y_pred))
print (classification_report(y_test, y_pred))
print (confusion_matrix(y_test, y_pred))
```

```
0.9985486211901307
```