Addis Ababa University

Addis Ababa Technology Institute

School of Electrical and Computer Engineering

Comparative Study of Machine Learning Algorithms for Smishing SMS detection model From CDR Dataset

By

Samson Akele

Advisor: Dr. Yalemzewd Negash

A thesis submitted to the School of Electrical and Computer Engineering in partial fulfillment of the requirements for the Degree of Master of Science in Computer Engineering

ADDIS ABABA, ETHIOPIA

May 2023

ADDIS ABABA UNIVERSITY

ADDIS ABABA INSTITUTE OF TECHNOLOGY

SCHOOL OF ELECTRICAL and COMPUTER ENGINEERING

By

Samson Akele

Approved and signed by board of Examiners:

	Name	Signature	Date
Dean, School of Electrical And Computer Engineering	Dr.Bisrat Derebssa	_____	_____
Advisor	Dr.Yalemzewd Negash	_____	_____
External Examiner	_____	_____	_____
External Examiner	_____	_____	_____

Declaration

I, the undersigned, declare that this research is my original work and has not been presented for a degree in any other university and that any source of material used for the research has been acknowledged.

Declared by:

Name	**Signature**	**Date**
Samson Akele	_____	_____

Approved By

Name	**Signature**	**Date**
Dr.Yalemzewd Negash	_____	_____

ACKNOWLEDGMENTS

First of all, my gratitude above all goes to the creator and governor of the two worlds, the almighty God, and all his Angels and saints for his incalculable and marvelous gifts to me. I would like to express my special thanks to my advisor Dr.Yalemzewd Negash for his unreserved supervision, prudent guidance, and suggestions in responding to my question and making valuable discussions that we had. Last but not least I would like to thank my friend, Getahun T., for his advice, and support at any juncture of my stay in the process specified in the domain area.

Abstract:

Phishing is becoming a significant threat to online security, and it spreads through a variety of channels like email and SMS or even a phone calls to gather crucial profile data about the victims. Although numerous anti-phishing measures have been created to halt the spread of phishing, it remains an unresolved issue. Smishing is a phishing attack that uses a mobile device's Short Messaging Service (SMS) to obtain the victim's credentials.

Employing an automated detection system will help improve identification and stop it before affecting targeted companies and third parties to alleviate this critical problem. A Smishing SMS detection-based CDR data framework is important to early monitoring experts and service providers in screening this kind of phishing attack, provides more accuracy, automates detection time, and keeps safe individuals. Many mobile phone users have been victimized yearly due to mistakenly interpreting the lures. Developing Accurate Smishing detecting system is helpful for organizations and related third parties who are highly affected due to smishing.

This paper compares machine learning algorithms for the smishing SMS detection model. In this thesis, six supervised machine learning algorithm classifiers K-nearest Neighbor (KNN), Support vector machine (SMV), decision tree (DT), Naive Bayer (NB), Random Forest (RF), and logistic regression (LR) are compared for the performance of detecting Smishing SMS which is more recommended by scholars and the result obtained prove that these algorithms are much efficient in detecting Smishing problem. 10-fold cross-validation based on correlation algorithms is used for classification and implementation.

The research collected Call Detail record CDRs data, and 33 distinct features were extracted initially, relevant features were selected, and eliminated unnecessary and irrelevant information, and different preprocessing methods, such as feature selection, and shaping the data were performed For the purpose of conducting this study.

As a result, the RF algorithm with options for Cross Validation (CV), which scored 90.1% accuracy, is determined to be the best classifier algorithm, the two algorithms come next with the best result, KNN and DT, which scored 89.6% and, 88.8%, respectively, Using cross-validation, the SVM algorithm performs inaccurately and exceeds the desired detection delay by more than an hour during training Time.

This outcome is the result of the RF algorithm's superior capacity to accurately handle vast amounts of data, form decision trees at random, and prevent overfitting by employing random subsets of characteristics to create smaller trees.

Keywords: CDR, Phishing, Outliers, correlation, cross-validation

Contents

List of tables

List of Figures

List of Acronyms

Cases

CHAPTER ONE

INTRODUCTION

1. Background

The development and expansion of Information Technology have resulted in greater internet connectivity and, increased usage of mobile phones. The literature shows [1] smartphones are becoming increasingly lavish and The Internet is being accessed globally by mobile phone users. Though over the past few years, there has been a significant increase in both the number of subscribers and revenue, it is extremely susceptible to deception. Due to the advancement of fraudsters' methods and techniques in tandem with the growth of the telecom business, the sector is severely threatened.

Telecommunication fraud is an activity or an intention to use any telecommunication infrastructure or services without willing to pay for the service. Globally there are different types of telecommunication fraud which brought a huge impact on telecommunication operators. These common telecom fraud causes the impact of, a company's image, brand, Revenue loss, Security and Quality of Service (QoS) reduction. Both telecommunication service provider and their customers can be affected by those fraudulent activities.

Due to its digital marketing popularity and lax security, deliberate abuse of telecommunication networks-based attackers are increasingly focusing on cellphones as their primary target. User privacy is becoming a significant issue, because users are storing their sensitive information on Android devices, such as through mobile banking systems, tele money apps, gifts, and credit card details in an Android device. Attackers aim to steal users' personal information by posing as a trustworthy organization, such as user passwords and financial information. Due to the following factors, mobile device users are more vulnerable to Smishing attacks [2]:

i. Due to the limited display of mobile devices and the fact that web addresses are not always fully visible, mobile users are unable to verify the veracity of texts they receive.
ii. Most mobile users don't know of security solutions that can be used to thwart phishing assaults, and they aren't aware of the hazards they may suffer as a result of adopting less secure practices.
iii. Mobile users are accustomed to providing their credentials whenever requested.
iv. Fraudsters are called them directly after sending smishing SMSs to the customer by saying I have sent a gift wrongly to you please send money back instead.

v. Expert surveys revealed that the majority of smartphones send cash back once they have entered their credentials into the user interfaces. Because of this, The perpetrator fabricates USSD user interfaces in order to steal the victim's financial and personal data.

The Short Message Service (SMS) enables the transmission and reception of brief messages between a mobile station and a wireless network as well as between a wireless network and an external device. Using cellular network signaling channels, it enables connectionless transmission of messages with a maximum character length of 160.

Today, SMS is an adaptable tool that can be automated and linked with other corporate systems. It is more than just a basic platform for rapid communication. Over many years, this communication environment has been the catalyst for and subject to numerous changes. The use of smartphones has grown quickly in Ethiopia, where text messaging has become the preferred method of communication. Text messaging is widely used to advertise and offer buy points of interest because it is the most affordable way to reach a large audience. Text messages are being used more frequently, which is quite advantageous for attackers. Experts estimate that more than 75% of SMS messages are read and responded to, which is crucial for focusing an attack.

Different SMS fraud exists since the beginning of the telecom service providers. Works of literature declare different telecom frauds and predict the number will skyrocket in the next decade. Smishing, in particular, can have a detrimental impact on mobile phone users, and service providers documented different types of Smishing, [3] [4] [5] [6]. SMS viruses, SMS spam, SMS spoofing, and Grey routes are a few of them.

Telecom services that are offered by Ethio Telecom. In addition to hybrid sim accounts, VSAT, mobile broadband, VPN, business portable and the web, M2M enterprise, fax, fixed wireless CDMA, telephone service, domain name service, mobile internet, EVDO, ADSL, roaming, and mobile phone services, the company also offers these services.

Ethio Tele registers its customers on Customer Relationship Management (CRM) systems based on the service request. CRM is a web-based system for organizing, automating, and coordinating sales, marketing, customer care, and technical support interactions with current and potential customers.

However, when service requests or applications are made, it will not have the opportunity to determine whether these subscribers are fraudulent or legitimate. False identity numbers, stolen logins, and stolen accounts can all be used in subscription fraud to obtain customers' full information and create numerous opening new accounts with malicious intent. Therefore,

the subscriber has the opportunity to begin their fraudulent activities as soon as they register and receive accounts to access the network. Some of the functions handled by this system are points of sales to register individual customers, Real-time reports to sum up payment reports, Service Changes like a change of SIM card, Service management of the status of the subscriber, Resource Business like selling customer resources and another one which is crucial for this study is General query all incoming and outgoing query CDR..

Figure 1. Customer care. Source Company image

From this illegal use of service SMS vulnerability are the leading telecom frauds. Moreover, some SMS frauds are the complexity of the technology, weakness in the operation system, free financial gain, irresponsible business models, and illegal acts.

Ethio Telecom now use conventional fraud management services (FMS) to find potential vulnerabilities. Fraudulent have a possibility to adopt these rigid rule-based FMS detection rules because of the domain rule-based detection system. Moreover, fraudsters steal customers' properties and damage the trust relationship with the company. Then after, they fasten losses of the company's reputation and increase revenue losses. The network expert, KPI expert, security experts systems, and domain troubleshooters are unable to handle the novel Behavior of scammers has changed.

.

To get control such problems there are Machine Learning (ML) approaches that can heal such illegal activities. Depending on the context of the analysis, SMS smishing detection can accept various input formats [15]. System logs generated by various network devices may be used as inputs for smishing detection. Alternatively, it can be call data records (CDRs), a type of communication log. The logs are presented generally as multidimensional data frames [14]. This method extracts an irregular user distribution near an access point or antenna that could be caused by an issue with the access point. This Smishing detection learns from users'

recorded data for their behavior change. Additionally, there is a general imbalance between the data classes (normal/harmful). This smishing healing approach gives the potential to detect smishing in rule-based systems.

Smishing refers to an attack where a perpetrator sends a user a text message that contains the perpetrator's phone number or email address, links to nefarious websites, programs, or user interfaces, or packages gifts and requests that the user enter their login information or mobile device PINs. Through these user interfaces, people divulge their private information, including usernames, passwords, credit card numbers, and even cash. Through this assault, the attackers obtain access to users' private information such as device PINs, contact information, mobile bank USSD codes, mobile money app IDs, etc. in addition to financial gain. SMS has a higher response rate from users than any other kind of theft, hence attackers prefer utilizing it for smishing [7] rather than email or other methods of theft.

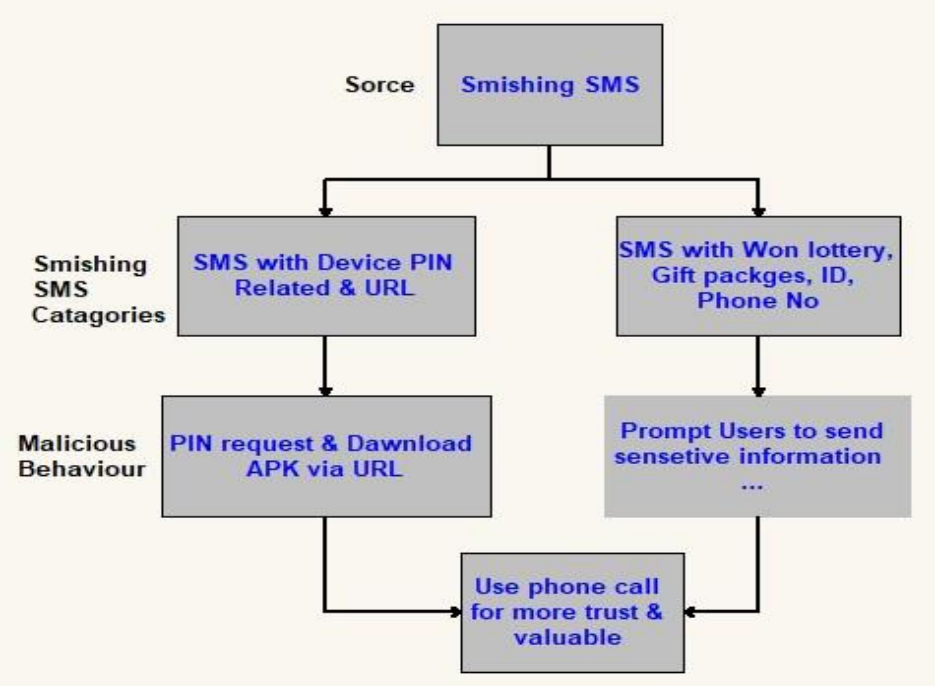


Figure 1-2 Malicious activities of a smishing SMS**.**

During smishing, attackers expertly design the user interface such that the victim cannot tell the slight changes between a legitimate service and a fake created by the attacker.

To detect SMS fraud a variety of detection techniques are proposed by the Department of ethio telecom fraud management team, the researcher, and other concerned parties [4] [8]. To identify between legal SMS and smishing SMS, a few ways have been put out by earlier researchers. We found that the blacklist technique was adopted by the majority of the tactics in use. [9] [2], the whitelist approach, and heuristic approaches [10] [11] to recognize

smishing attacks. However, the blacklist method is ineffective because smishing messages' domains change often. Particularly, unless and unless we discover that the victim is reporting to the provider's operator when attackers attempt to steal the user's credentials or after victims have been attached by the fraudsters, smishing SMS utilizing a stolen Sim and fake ID subscription Sim cannot be characterized as a smishing message. It is essential to check the historic behavior of the service number message communication, transactions and, calls to categorize it as smishing. The methods currently used by service providers to identify smishing messages have certain drawbacks. Therefore, it is imperative to develop a fresh and effective system that recognizes smishing messages. We have examined the smishing SMS's behavior, the average balances within each month, the average main balance account vs transfer rate within a part of days, frequencies of main account recharged in each of last month vs balance transfer rate with part of days, frequencies of main account recharged from different channels, numbers balance recharged by the user in last each month. A total number of balance recharged from different channels, and have also analyzed the behavior of the incoming calls. We have analyzed the recharge account of service numbers to check its malicious behavior.

1.2. Statement of the problem

Among many other types of social engineering attacks, Smishing is a cyber-security issue faced by mobile phone users and serious damage to individual users and remains a major security threat to organizations, in particular, smishing has a significant impact on individuals. According to a recent CFCA poll (2021), global fraud losses are pegged at $39.89 billion [12]. Fraud losses climbed by 28% in comparison to the year 2020. Smishing is about $2.03 billion lost due to SMS hacking, phishing, vishing, and other identity theft methods. The manually blacklisting/whitelisting method of smishing is Fatiguing & distraction, tedious, time consuming, highly subjective errors and impractical, limited visualization system and inadequate training and experience, in today fraud management systems where collecting information from complained individuals about smishing and analyze manually the historical behavior of the suspected smisher, However due to huge amount of fraud complains and lack of automated smishing detection system, there is a high rate of human faults/errors and dalliance to analyzing the fraudulent behaviors for smishing those many cases millions of remaining or unaware, individuals are affected even with a single fraudster and the behavior of smishing messages are changed fast, Attackers keep on changing fraud services frequently

and making the blacklisting technique is ineffective. So, it is imperative to develop a fresh and effective technology that detects smishing SMS. Therefore, to solve the problem, this thesis/research is aim to develop a smishing SMS detection System in Call Detail Records (CDR) data and Machine Learning Algorithms are compared. These can be help in developing countries where the number of mobile users who are innocent is huge, and telecom service providers cannot detect automatically.

The research questions that should be answered at the end of this work are:

Research Questions

1. What kinds of CDR features can be used to detect Smishing SMS?

   What Machine Learning techniques are used to identify Smishing SMS?

1.3. Objective

1.3.1. General objective

Our objective is to study and suggest methods for SMS Smishing detection using a Machine Learning ML approach. The proposed techniques are based on data analysis concepts. Specifically, We focus on developing algorithms that enable the extraction of data essential to the detection process from the vast amounts of data generated by call detailed records (CDRs). The objective of this work is based stated earlier; we focus on analyzing which ML algorithms perform better for the detection and demonstration of SMS Smishing.

The thesis has two main objectives:

SMS Smishing Detection: our aim is to detect SMS Smishing from Call Detail Records CDR data. It should be automatic and supervised. The suggested solution should be efficient (have a low mistake rate with few computational resources) and learn from prior data using expert rules. Additionally, the suggested solution must be implemented as a module of a monitoring system, process the data it generated, and identify Smishing patterns in the network measurements in order to identify malicious conduct.

Our goal is to develop a system that analyzes the cellular network's CDR data and locates the source of problems in the Data to localize the issue. The analysis should produce a result that includes all the data required for the detection procedures. It captures data and analyzes all incoming and outgoing calls, as well as all SMS input and output, within the network in order to reach a judgment based on all CDRs.

1.3.2. Specific objective

The specific objective of this thesis includes:

- choosing appropriate features for a Smishing SMS detection model
- To choose the appropriate machine learning tools, algorithms, and detecting methods for SMS smishing
- Creating models and conducting analyses with chosen algorithms after cleansing and converting the data into a format that is appropriate for the identification of SMS Smishing.
- Detecting smishing SMS based on CDR historical usage behavior
- Evaluate and compare the performance of the model
- to report the result and leave a recommendation for further research

1.4. Scope and limitation

The Smishing can be divided into a variety of categories and levels according to their breadth [16]. However, this study primarily exclusively addressed SMS Smishing. Telecommunication Call Detail Record (CDR) data, which is data recorded and created by Ethiopian telecom equipment and records the specifics of a communication transaction, is the raw data that is used. The research just keeps track of the source and destination phone numbers, call duration and time, connection and completion status, SMS inbound and outbound, service kinds, and total cost. The study only makes use of CDR data from five mouths.

1.5. Significance of the study

Nowadays, as telecommunications technologies and services develop quickly, so do the types and methods of telecom fraud, phishing, and smishing. Connecting newly developed technologies to an established network reveals new forms of fraud. Customers are diverted and annual revenue for telecom operators decreases as a result. SMS are susceptible to a variety of attacks among these unlawful activities. This paper's main contribution is a model for identifying SMS smishing based on ML techniques. The purpose of this thesis is aid the in the execution of.

- analyze user behavior
- locate and uncover hidden correlations
- Showing a direction and allowing telecom operators to point out and detect SMS Smishing before affecting subscribers and affecting revenues.
- Give concrete recommendations that experts used to manage service quality by detecting Smishing and advances about the universality of SMS Smishing.
- Provide insight about the concept of self-healing network, human expert troubleshooting tasks and its limitation
- Choose more sophisticated Smishing detection algorithms

1.6. Related work

In recent years, different researchers have made their studies on the detection and prevention of telecom fraudsters' concentrating more on Smishing because it's common in mobile attacks. Smishing fraud is one of the majority of the top listed fraud methods victimizing communication service providers and their subscribers. A number of researchers have been conducted in implementing machine Learning Algorithms in order to identify Smishing massages.

Research conducted by Ankit Kumar et al. [1] [17] [18] [19] [20] [21] proposed models to detect smishing messages. Similar strategies have been employed by various spam detection models [22] [23], and numerous researchers have discussed smishing tactics and detection methods to raise awareness among users and researchers [17] [24] [1]. Specifically, this section gives an overview of related literature presented in the context of smishing.

The research conducted by, foozy et al. [24] illustrates the taxonomy of mobile device phishing detection. The study outlined and commented on a number of phishing tactics, including Bluetooth, voice, SMS, and mobile application phishing. Additionally, the author tested a variety of phishing detection methods to compare and assess them. The author of [5] suggested a method dubbed "S-Detector" for the detection of smishing messages and employed Naive Bayes classification. The SMS monitor, SMS analyzer, SMS determinant, and database make up the detector. The URL is identified by the model in order to determine whether the.apk file is web-based rather than to classify SMS as smishing.

Another study by Ankit Kumar, et al. [11] proposed a rule-based method for the detection and identification of SMS smishing messages. The author developed nine rules to separate smishing messages from genuine messages using the classification algorithms Decision Tree, RIPPER, and PRISM. The system was tested and trained, and the mother's performance review reveals a success rate of more than 99%. A research work by Tarikua [3] characterizes the possible existence of SMS frauds and likes between the scheme and its root causes. The plan comprises four key phases to reduce SMS fraud: Theoretical analysis as well as empirical analysis, refinement, and evaluation Matrixes, which are used to infer for the taxonomy, and SMS fraud outs. The other is the taxonomy construction method, which made use of cause-and-effect and question-and-answer techniques. The four core nodes of this taxonomy technique technology, vulnerability, fraud, and mitigation are built utilizing the cause-and-effect and question-and-answer methods. These nodes are comparable to cause-and-effect diagrams where fraud represents the vulnerabilities and mitigations

represent the consequences of the technology, however the question-and-answer process involves developing unique questions for each node (Which), (were), (How), and (what).

Regarding the analysis of the researcher's experiment a database of 10,000 SMS fraud numbers from various channels is used by Tarikua [3]; approximately 98,491 of the total records are filed using mitigating methods. These mitigation strategies fail to discover trends in 2,081 records. This suggests that the new strategy can negate 98% of the signals before any effects are felt. By examining the application information obtained prior to the attack records, the study evaluated and greatly filtered SMS fraud in the telecommunications sector. This implies that 98% of the signals can be neutralized by the new method before any consequences are felt. The study assessed and significantly filtered SMS fraud in telecoms by looking at the application information prior to the assault.

Sonowal et al. [17] presents a model called SmiDCA and a machine learning strategy, show the design and implementation of a system for detecting smishing messages. 39 characteristics were extracted by the suggested model from various sources. The study experiment was conducted on different datasets and Machine Learning Algorithms. Experimental evaluation of SmiDCA shows good accuracy using a Random Forest classifier. This method of detection focuses only on different Smishing contents from a different source but the method misses to detect smishing calls after the text. On the other hand, Goel, D et al. [1] studies a smishing detection method for recognizing smishing communications, called'smishing classifier'. Using a Nave Bayesian Classifier, the suggested approach locates the SMS contents and SMS keywords. The study framework evaluates the sender's cellphone number and confirms the presence of URLs in SMS messages. Additionally, the model assesses the look of the login page and APK file downloads.

A research conducted by Chaudhari, A.S [25] security analysis of SMS and related technologies. This research basically focused on increasing awareness of SMS security as attackers illegally access sensitive data through messages and technological devices. The study mentioned the potential countermeasures to enhance the security of SMS-based authentication schemes against vulnerability and attacks. The basic analysis of this research is to secure the backbone traffic which the transmission of the data between various protocol layers and network entities should have secured and message filtering on channels. The thesis discovered the fundamental security issues related to authentication procedures.

On the other hand, Mishra, S. & Soni, D. [26] and Fitsum Tesfaye [27] presented the importance of CDR data analysis in the case of natural disasters. Botnets are an efficient

malware-launching platform where several bots instantly send out a fresh worm or virus, taking on a worldwide and menacing presence.

Research has shown that SMS messages can be used to spread malware via URLs sent within SMS messages, distribute SMS spam, perform denial-of-service (DoS) attacks to send premium rate SMS messages without user consent, and convey command and control (C&C) instructions. One well-known study on mobile botnets that spread malware through SMS messages' transmitted URLs was done by Mishra et al. [26] and is a well-known example of this type of research. Hua et al. [10] have done a well-known study on mobile botnets that use SMS as a vector of spread.

Min Kang, et al. [21] discussed a numerous phishing and smishing attack types. To verify the legitimacy of the URL given in the message, the author has suggested a URL validation test. The smishing box strategy has also been explored as a defense against programs that are downloaded during a smishing attack. Many academics work to identify fraudulent actions before they have an impact on businesses and their subscribers due to the effects of smishing fraud. To identify smishing communications, Mishra, et al. [26] employed a content-based technique. The most common words used in smishing SMS are determined using a machine learning system. Additionally, this model assesses the login page's look and the.apk file download to check for maliciousness in the URL.

Ravi, et al. [28] Describe the process by which the tool can be incorporated into a program so that a user can use it to determine if a suspicious SMS is real or fake. The suggested approach uses TF-IDF to vectorize the SMS's text using machine learning algorithms to interpret it. They have demonstrated a method for smishing fraud detection based on machine learning. The author discovered from the tools outcomes that even when random forest produces better assessment metrics. Goel, D., & Jain, A. K [1] gave a thorough analysis of smishing assaults. The author has talked about different mobile smishing attacks carried out by attackers and their defenses. Additionally, they went into detail on the taxonomy of smishing solutions, author strategies, various difficulties in smishing attack detection, and numerous smishing datasets.

Priyanka Maan et al. [29] present how to perform the spam inclusive Smishing messages identification based on a hybrid model. The paper has used web communications in the form of chat, emails, and some other message communication systems of the textual dataset as input and use fuzzy logic to transform this textual data into statistical information. After accepting this textual data as input information processing is done to minimize sizes and to

perform filtration, filtration includes the removal of non-keywords from the list and classify the positive and negative sense words. The proposed use of decision tree algorithms to perform the message classification and the result show that the work is effective in the recognition of normal messages.

Depending on the characteristics of cellular networks, SMS contents, and user behaviour, the existing technologies described each play a different role in detecting smishing attacks. Some of the rules in rule-based techniques addressed how smishing attacks progressed. As a result, various forms of smishing fraud detection methods and strategies have been noted. The study found that numerous variations of smishing attacks are not covered by rule-based techniques. Approaches analyze the contents of the text message whitelisting, blacklisting, Machine learning techniques used to SMS contents alone are unable to identify the varied smishing tendencies. In this study, we strongly feel that it is necessary to analyze the behavior feature numbers of incoming calls, outgoing calls, SMS In, SMS out, Actual Volume, main recharge account, frequency transaction and Actual Charge Amount, and other recharge channels.

1.7. Method

The main objective of this study is to build a smishing SMS detection model using machine learning techniques. The following actions are taken in order to fulfill the study's goal.

1. To identify and outline the problem domain, review the telecom kind's literature on smishing fraud detection in SMS, .apps links, recharges, balance transfers, and calls
2. Build the dataset for both subscribers to be used for testing and training the model by gathering CDR data from both honest and dishonest subscribers, selecting the proper attributes, performing preparation (data cleaning, integration, and aggregation, for instance), and building the dataset for both subscribers.
3. To define the issue and establish the study's objective, domain experts consult one another.
4. The suggested model for detecting smishing SMSs consists of six machine learning algorithms: Random Forest (RF), Logistic Regression (LR), Naive Bayes (NB), a support vector machine (SVM), a decision tree (DT), and K-nearest neighbor (K-NN).
5. Anaconda 3 Python programming tool version 3.7 64-bit is preferred for this specific research to create models using classification algorithms The Anaconda environment has the advantage of being designed with data science in mind. It

contains popular Python modules, Anaconda keeps module versions for reliability, and researchers can always add new modules to their home directory.

6. Use the performance measurement metrics (confusion matrix, received operating characteristics (ROC), F-measure, and accuracy) to test and assess the generated model. At the conclusion, recommendations are given and a more effective model is recommended after models have been compared.

Chapter two

2. TELECOMMUNICATION SERVICE BACKGROUND AND FRAUDS

As of this chapter, some communications services concepts are introduce relevant to the study, as we explain in the thesis objectives, our goal is to create a model for detecting smishing in mobile networks. We start by outlining the structure and features of the current cellular networks. The complexity of SMS services is then underlined in terms of performance and dependability. The state of the art of smishing detecting systems and some of their functions are then detailed. We aim to draw attention to the discrepancy between the performance of smishing detection systems and the required capacity. Whereas, the next section attempts to fill this gap.

2.1. TELECOMMUNICATIONS MOBILE SERVICE

The process of sending, transmitting, and receiving information over a long distance with the intention of conversing is known as mobile telecommunications service. This type of signal transmission requires the assistance of a mobile device, such as a cell phone, computer, or other wired or wireless devices.

Mobile services are the provision of telephone services through using a mobile phone and SIM cards, which may move around freely rather than stay fixed in one location within the service network coverage. It is supported by different stages of technologies such as 2G, 3G & 4G, which are incorporated with better quality & feature advancement of the technology on the mobile business.

Mobile technology Different mobile services, such as mobile voice, mobile internet, mobile SMS, etc., are supported by mobile technology. Of which the mobile voice service helps users to make and receive telephone calls to and from mobile and fixed-line telephones across the world.

To use this voice service there are different modes of payment. These are pre-paid Postpaid and hybrid modes of payment.

Prepaid customers are required to recharge using scratch-able cards which are available in various denominations to meet customers' needs.

Postpaid customers, Based on their usage, postpaid clients are invoiced at the end of the month.

Hybrid customers can use both prepaid and postpaid modes of payment at different times.

Currently, telecom providers offer a variety of services, including mobile services as well as fixed-network products (data retail, Internet retail, voice retail, and wholesale).

2.1.1. Fixed-data services

All dedicated/private line, packet, and circuit-switched access services (including frame relay, asynchronous transfer mode, IP, Integrated Services Digital Network, DSL, and satellite) are included in these retail sales. These services differ in the type of traffic or application they transport. There are many different types of transmissions, including non-voice data, images, video, fax, integrated services, and even voice.

2.1.2. Fixed-voice services

Fixed voice telephone service is a specific quality of voice telephone service offered through a fixed telephone network between fixed termination points of that network. This reflects the retail voice service revenue for all goods and services made available to end users as such, including voice-related local and long-distance services (calling costs, line rental/subscription fees, and connection fees are included in this category), enhanced voice services, data and fax transmission over the circuit-switched PSTN, and retail voice over IP revenue mobile telecom service.

2.1.3. Mobile telecom services

Mobile phone calls and use of mobile data Short Message Service (SMS), and, mobile data access will generate revenue for all mobile operators in that geographical market. Consumer charges are dropped. Revenue from mobile phone calling costs, mobile data access, SMS fees, line rental/subscription fees, and connection fees are included in this category.

2.2. CELULAR NETWORKS

Telecom companies provide a variety of services, from the most basic, like messages services, conversations voice, multi-media messaging service (MMS), and unstructured supplementary service data (USSD), to the more advanced services that are available on smartphones, like video chatting and streaming, as well as web-based services like browsing, email, social networking, and peer-to-peer (P2P) file transfer. More services based on the Internet of Things (IoT) and machine-to-machine (M2M) communication are anticipated to emerge on mobile networks for smart homes and cities [30] [31]. The variety of services increases the complexity of the network architecture.

2.2.1. Radio Access Network (RAN)

The RAN, which connects subscriber devices to the core network wirelessly, is a component of the network. One or more Radio Access Technologies (RATs) are used in its

implementation. GSM networks include the Radio Access Network (GERAN). The radio component of GSM/EDGE is known as GERAN, along with the network that connects base stations and base station controllers. With a bandwidth of 200 kHz, a data rate of up to 1.89 MBps, and a delay of at least 180 ms, this network standard uses the Enhanced Data Rates for the Evolution of GSM (EDGE) cellular technology, which enhances information transmission and serves as a compatible with previous versions addition to standard GSM [31] [32]. An access network made up of a network of interoperable evolved nodes B (eNodeBs) is known as the Enhanced Universal Terrestrial Radio Access Network (UTRAN). It is a wireless technology that is utilized by 3GPPTM systems between mobile terminals and base stations [32]. It might operate between 1 MHz and 20 MHz in frequency. The data speed is up to 300MBps, and the delay is under 2 ms.

2.2.2. **Core Network**

The central component of a telecommunications network, the core network provides clients linked by the access network with a wide range of services. Call routing over the public switched telephone network is its primary duty [33] as cellular networks advance, various core network types are developed.

This term typically refers to the incredibly efficient communication network that joins important nodes. Multiple sub-networks can share data with one another through the core network. Instead of "core network," the term "backbone" is typically used to describe enterprise networks that serve a single organization, however, the word "core network" is regularly used when discussing service providers.

2.3. **Monitoring Systems**

Cellular networks are extremely difficult to manage, as we covered in the last section. In this regard, the standards organizations and the communication market compelled telecom providers to provide high-quality service. These facts create the cellular networks' governing system problems.

Having complete visibility and control over the network and its operations is the core goal of cellular network monitoring. Network administrators, software technicians, IT professionals, IP technicians, security specialists, and telecommunications experts all perform this function. By automating tasks and disseminating important information, incorporating the monitoring systems aims to boost speed and efficiency.

2.3.1. Structure

The structure of monitoring systems varies depending on a number of variables, including the network's size, traffic volume, available services, operator preferences, fraud, and smishing. Centralized or decentralized monitoring systems are also possible. The next section focuses on the four crucial elements that are found in almost all monitoring systems. The mirrored components known as probes [8] can be added to the network monitoring tool or installed separately in the devices being monitored. The monitoring system's "brain" is a data processing entity. It implements fundamental features including data aggregation and filtering. Additionally, it has advanced features like analysis of trends, smishing detection, network assessment, capacity planning, and (QoE) prediction using BI and ML tools. Produced by this part are status updates, KPIs, and summary information.

The data processing institution's information triggers a real-time alarming mechanism. For instance, if a KPI falls below a set threshold, it alerts the administrators. It allows the network administrator to see performance and status reports and investigate the various network components.

2.3.2. Data Types

The researcher gathers thorough real-world data while maintaining the privacy and confidentiality of consumer records. This communication data is known as a Call Detail Record (CDR), which is a data record created by telecom equipment and records specifics of a communication transaction. Our record is limited to source and destination phone numbers, phone call times/duration, connection/completion status, SMS in/out, and total cost. Now we would like to briefly introduce the analysis and related information that we will be working.

Two sorts of logs are produced by mobile network monitoring systems: communication records and system logs [15] [14].

The records a monitoring system keeps of subscribers' behavior, such as call traces, are referred to as real-world communication logs. Call traces are unstructured data sets that contain all of the communications that network devices exchange during a subscriber-initiated mobile conversation. The call data records (CDRs) and the session data records (SDRs), which are organized as logs containing the details of the network equipment and the service information used during the mobile connection, are created from the combination of these traces. By integrating CDRs and SDRs into multifunctional counters, such as the average time to the response by RAT, service, and handset type, the efficiency of the network

is reported. The logs used to record system activities, operations, and communications, including the address resolution protocol (ARP) requests, are known as system logs. If an unusual occurrence occurs, these logs are utilized to produce alerts. System records and communication logs are combined into KPIs that show the historical development of performance metrics like CPU use and call drop rate across time series.

2.3.3. Limitation

As a significant business, Ethiopian telecom providers has its own shortcomings despite developing monitoring systems that are effective and autonomous. As the study mentioned in the previous section, operators are still below the requirements in many aspects. First, network monitoring, communication monitoring, and every system monitoring process are relies on the presence of humans. Customer complaints tasks, system failures, and fraud management monitoring tasks are performed manually. Although monitoring systems produce KPIs and alarms, professionals assess these later when diagnosing the network [31] [34]. Because of this, troubleshooting is an extremely expensive operation for operators. Second, cellular networks nevertheless occasionally (downtime) or continuously (in particular specific instances such as SDP and mobility) have low efficiency. The systems for tracking are not responsive and effective enough to navigate SDP and handovers effectively or to resolve the unavailability of services in real-time.

2.4. Telecommunication fraud

Any behavior intended to take unfair advantage of and obtain a competitive advantage over telecommunications businesses is considered telecommunication fraud [62]. Since the telecommunications sector is the oldest and biggest network in use in the world and generates 40% of global sales of consumer goods, fraudsters have devised tools and techniques to take advantage of it and extract profit. With increasing, mobile phone subscriber fraud attacks also usually involve quickly over time as companies squash them.

Fraud has a number of costs, including harm to a company's reputation and performance, in addition to unpaid invoices and significant revenue losses. Fraud may also result in the possible loss of both current and potential clients as well as negative press. Telecommunication fraud attacks particularly customers by calling and messages from a stolen phone and incorrect numbers subscriptions sold for a better commission from retailers. As the technology to run mobile networks become more available, operators and customers are becoming the target of fraud indirectly. The opportunity of using a fake or stolen identity, this helps fraudsters run illegal, very lucrative businesses with little capital commitment and

a low chance of being found. Additionally, the fraudster is more likely to conduct the same crime again. In the event that a subscription is terminated, the scammer will usually find a new subscription and carry on with the fraudulent actions.

2.5. Types of telecommunication frauds

Investment fraud comes in forms now's the day, in the telecommunication sector, there are a number of telecommunication fraud types that incur both direct and indirect losses for fraud. Due to significant customer discontent from bill disappointments and negative client experiences, it has had a negative influence on customers and resulted in customer turnover. Additionally, it hurts the operator's reputation because customers would gripe about being unaware of the abundant call rates charged to them.

According to a recent CFCA (2021), report The majority of the top ten fraud methods, including spoofing, Wangiri, SMS phishing (Smishing), account takeover, sim swapping, phishing, and robocalling, target subscribers of communication service providers, having a direct financial impact on customers and financial service providers.

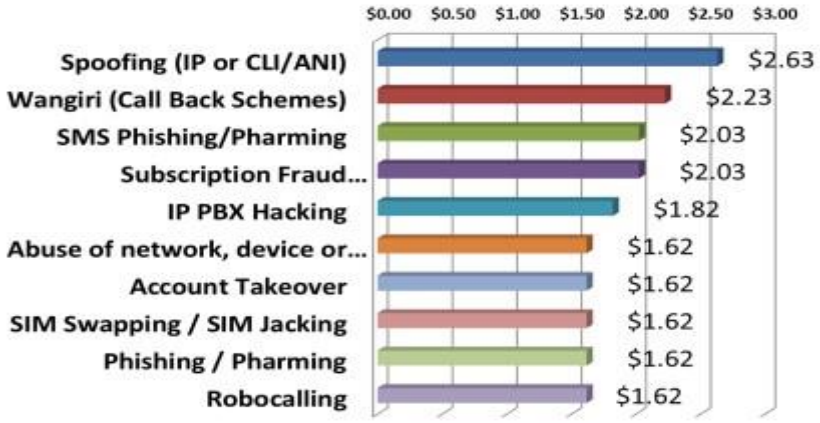


Figure 2-1. CFCA 2021 report [35]

There are also other many different fraud types that cause a huge amount of loss every year. In this subsection, the research explains the most important ones which are related to Smishing properties. Subscription fraud, cloning fraud, SMS premium rate fraud, roaming deceit, and wangiri fraud.

2.5.1. Subscription Deceit:

Due to the fact that they are registering for a subscription using stolen or made-up identities, fraudsters typically have no intention of paying. The most prevalent and harmful non-technical fraud on the GSM network is acknowledged to be subscription fraud. This type of fraud is the ground for another fraud type. Different methods can be used to manage subscriptions. By assuming their own identity, updating parts of their personal data, or by

assuming the identities of others. The personal use of the fraudster or someone he transfers the phone to falls under the category of subscription fraud. The second is done to make actual money; in this case, the fraudster poses as a small firm to acquire several handsets for Direct Call Selling reasons. The scammer, who has never thought of paying his account, now sells the airtime to those looking to place inexpensive long-distance calls, perhaps in exchange for cash [26] [36] [14].

2.5.2. **Cloning:**

Cloning, which is similar to SIM swapping, involves creating a duplicate SIM from the original. This technique is however highly advanced technically. When the app copies the real Sim card. Cloning is carried out to make calling someone else's subscription simpler. This method allows for the acquisition of the victim's IMSI (international mobile subscriber identifier) and encryption key. Cloning has the advantage of charging calls made from a cloned phone to the person with the original subscription. The fraudsters will be able to seize control and use the mobile phone to track, monitor, listen to calls, place calls, and send texts by cloning primarily the SIM. Cloning, however, is currently a reality within GSM as well. Cloning GSM phones is thought to be an extremely challenging process. Each subscriber is identified by their IMSI number. The IMSI is located on the SIM card. The SIM card houses the IMSI. The SIM card also contains a secret key that is used to authenticate the network subscriber [26] [14].

2.5.3. **SMS fraud.**

SMS fraud refers to the abuse of the mobile premium service (SMS premium rate) used for phone billing. This involves the abuse of premium rate service and can occur in different ways. This service is provided by many legitimate service providers due to its simplicity of use as a mobile payment mechanism. Fraudulent phone calls to expensive (premium) locations are known as Premium Rate Services (PRS) destination fraud. Users can, for instance, order a range of mobile content (such as ringtones, wallpapers, and donations), receive the item they ordered, and the cost will be added directly to their phone bill. The aggregator (middleman), who manages the technical service, pays the service fee to the service providers after the transaction is complete. To subscribe or receive advertised content, a user typically has to send an SMS to a given number. Figure 4 illustrates how an attacker exploits the premium-rate service by subscribing a victim to a vast number of premium service providers silently. In this SMS scam, the perpetrator establishes its own service provider and pays the aggregator according to the number of subscriptions. Depending on the country, some services require an acknowledgment from a user before [10] [36] [37] the

charge is processed, as part of the service provider's procedure. However, the attackers have exploited this mechanism by intercepting the acknowledgment messages from an infected mobile device without the user's consent.

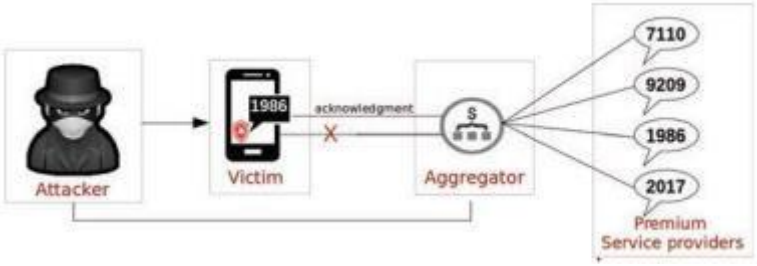


Figure 2-2.Premium rate SMS attack [37]

2.5.4. **Roaming Fraud:**

Roaming is the usage of a mobile service when you are unable to connect to the one you typically use. Operators let visiting subscribers use their networks and their subscribers use their partners' networks. This makes it possible for subscribers to use their mobile phones in foreign countries or areas not covered by their home operators' networks. Based on the work [1] [37] Roaming users' mobile phones are frequently stolen, typically while on vacation, because the time it takes the home provider to receive roamer call data can range from one to many days. The capacity to utilize telecom services, such as voice or data services, outside of the home network without the intention of paying for them. The extended time-frames necessary for the home network are used in these situations by fraudsters. When SIM cards are obtained and sent to a foreign network, roaming fraud might begin as subscription fraud or internal fraud in the home network.

2.5.5. **Wangiri fraud.**

A callback scam dates have its origin in the Japanese word 'one ring and cut' telecom operators are facing this for over a decade now and this is only growing exponentially year after year around the globe. In wangiri, a con artist uses an international or odd number to leave a missed call on the phones of multiple victims in several countries. To the user, the caller number ID is modified in such a way that it looks like a genuine call when the victims call back. It turns out to be a premium rate service number owned by a fraudster for which a victim is charged heavily for the calls. The fraudster intends to keep the victim on hold to increase the building amount. The premium rate provider pays the fraudster a certain share of the call revenue for each minute of call received by the premium number [38] [39].

There is also an SMS version of this when scammers send a message instructing victims to phone or text a particular number back.

Telecommunication incurs both direct and indirect losses for wangiri fraud. Significant consumer unhappiness from bill shocks and negative client experiences has negatively influenced a negative influence on customers and caused customer churn. It also has a negative impact on the operator's brand image as subscribers would complain about the high call rates charged to them without them being aware of it.

According to CFCA 2021 [40] fraud loss survey report, wangiri is one of the top fraud methods used by fraudsters to carry out fraudulent activities. It is also estimated that telecommunication is close to $2.23 as a result of Wangiri or Callback fraud schemes.

2.5.6. **SIM Swapping.**

SIM swapping fraud happens when con artists use the phone number you provide to access your account using two-factor authentication and verification, where the second requirement or step is a text message (SMS) or phone call to a mobile device. A SIM swap scam is also known as a port-out scam, SIM splitting, Smishing, and sim jacking. In order to scale their SIM-swapping attacks, fraudsters are increasingly using social media profiles to attract staff of mobile phone providers. Attackers have the chance to entice insiders with the promise of financial advantage by pretending to be a company hiring for available positions through these accounts [1] [41].

2.5.7. **SMS Smishing.**

The term "smishing" combines the short messaging service (SMS) and the texting-related technologies. Phishing is the act of stealing money or personal information through false pretenses communications, primarily emails, financial data, and online account credentials. Traditionally, SMS Phishing refers to a form of phishing that uses the social engineering technique as a method of information retrieval to acquire users' sensitive information. For instance, a fraudster sends the victim an SMS message asking for sensitive information including credentials via a Web link or a telephone number [42] The term "SMS Phishing" refers to a technique of spreading malware that uses SMS messages to disseminate and convince users to take certain actions in the context of Mobile financial malware, several actions, as depicted in Figure 5.

Regards to Ghorbani et al. [18] describe the methods of distributing malware via SMS messages which are spoofing and malicious apps. This means spoofing refers to manipulating the sender's information by changing the originating mobile number with different international codes for the purpose of impersonating another person, company, or product.

This method makes use of unrestricted SMS spoofing services that are available for free. These services were primarily developed for customers who do not own a mobile phone but must send an SMS from a number they had previously given to the recipient. However, the attackers are using this service as a vehicle for the distribution of malware.

The other scenario which is SMS Smishing via malicious apps consists of several processes: the attackers first upload malware to their hosting sites to be linked with the SMS. They control their attack instructions via a C&C server i.e. send false SMSs, eavesdrop on SMSs, and steal data. Once the victim received the SMS and visited the malicious URL, the malware is installed on the victim's device without the victim's knowledge [2] [22]. After being installed, the virus behaves in a manner that is characteristic of phishing attacks by asking for device administrator rights and then running in the background to carry out a number of evil deeds.

Typically, these actions include the following:

- All incoming and outgoing SMS messages should be captured.
- send an SMS text message to each contact in the victim's phone book in response to receiving command and control (C&C) instructions in this manner
- using pre-defined keywords like "Pay," "Check," "Bank," "Balance," and "Validation" to intercept SMS messages and steal sensitive information (such as financial data).

All collected data is sent to a distant C&C server. As a result, the attacker can use the victim's personal information to open a new credit card or transfer money from the victim's account.

2.5.7.1. Smishing fraud properties.

The primary characteristics of smishing fraud are included in the list below.

- Fraudsters make a lot of outgoing phone calls.
- Fraudsters Send Smishing texts with some common characteristics of messages like conveying a sense of urgency, containing a link, and asking for personal information.
- Local calls are the primary destination of fraudulent customers' calls.
- After sending smishing SMS, call victims repeatedly while posing as someone else to take advantage of their confidence.
- There might be virtually no incoming calls.
- Fraudulent users are using SMS to send expensive cell package gifts, claim tax refunds, check their online bank accounts, and claim prizes.
- Receives a high amount of money from different online banking and as soon as transfers to others

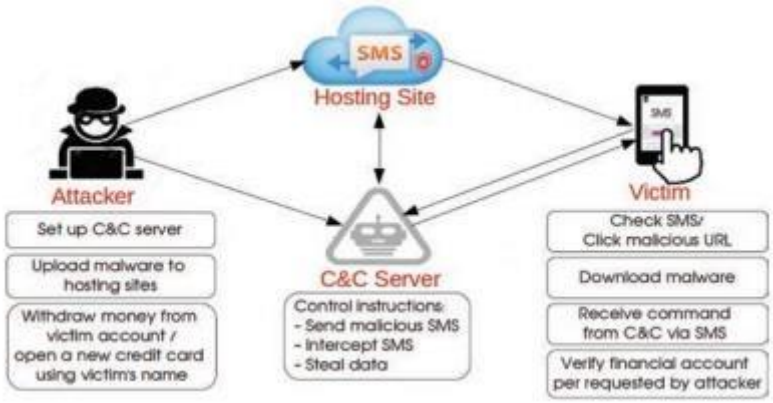


Figure 2-3. SMS smishing fraud [2]

CHAPTER THREE

3. MACHINE LEARNING

Machine learning is an application of AI that enables systems to take lessons from their past performance without needing to be explicitly programmed. The goal of machine learning is to enable computers to access data and learn independently [43].Machine learning techniques have made it possible for computers to operate independently without explicit programming. Apps with machine learning (ML) capabilities can freely learn from new data and evolve. Machine learning uses algorithms to find patterns and learn in an iterative process, extracting valuable knowledge from massive amounts of data ML algorithms employ computation methods to acquire knowledge instead of relying on any preexisting equation that may be used as a model, one should calculate directly from data. The basic process of ML is to train and test the model to generate a new set of rules based on inference from source data [22] [44]. During the 'learning' processes, the performance of ML algorithms adaptively improves with a rise in the quantity of accessible samples. To build a model, machine learning algorithms are applied to a training dataset. The trained ML algorithm uses the built-in model to anticipate outcomes as new input data is fed into the system. Through the development of rapid, efficient algorithms and models that utilize data for the analysis of real-time data. It can generate more accurate outcomes and data analysis [3] [27].

By human standards, processing much data takes time and is challenging, but high-quality data is the best source for teaching a machine learning system. A large dataset's ability to train a machine-learning algorithm will increase with the amount of clean, usable, and machine-readable data present [43].

As stated, machine learning algorithms have the ability to improve themselves through training. Currently, three popular techniques are used to train ML algorithms. Unsupervised learning, reinforcement learning, and supervised learning are the three categories of machine learning. This research focuses on machine learning algorithms for classification. The next sections describe these ML categories. These algorithms are supervised, unsupervised, and reinforcement techniques that are used for the classification task. At last, we describe supervised ML techniques with the proposed algorithms because the data used in this work is labeled data.

3.1. Unsupervised learning

Machine learning techniques known as "unsupervised learning" examine and train models using unlabeled datasets. These algorithms identify hidden patterns or groups of data without the assistance of a human. It gathers information and processes the whole input, in order to make decisions based on the entire input. When using unsupervised learning the system does not know the right "answers". However, the methods' capacity to identify similarities and contrasts in data makes them the ideal choice for cross-selling tactics, picture identification, and exploratory data analysis. The goal of the unsupervised [3] [14] [43]. The system must classify unsorted data based on resemblances, patterns, and differences without any prior training on data.

3.2. Semi-supervised learning

Semi-supervised learning is a large category of machine learning methods that can be used to analyze both labeled and unlabeled data. It is a combination of supervised and unsupervised learning methods. To treat a data point differently depending on whether or not it has a label is the main goal of semi-supervision. While the algorithm for unlabeled data minimizes the variance in predictions between other similar trained data, the technique for labeled data uses supervision to update the model weight.

3.3. Reinforcement learning

In order to optimize a numerical reward signal, reinforcement learning involves learning what to perform and how to link situations to actions. The learner is not given instructions on what to do; instead, he or she must experiment to determine which activities would result in the greatest rewards. Reinforcement learning characteristics with trial-and-error search and delayed reward.

Supervised learning, the type of learning that is the focus of the majority of recent studies in the field of machine learning, is distinct from reinforcement learning. Supervised learning is the process of learning from a training set of labeled examples that are provided by an experienced outsider. Additionally, unsupervised learning, which machine learning experts refer to as learning, differs from reinforcement learning in that it often focuses on uncovering underlying structures in collections of unlabeled data. The opposite approach is used by reinforcement learning, which begins with a fully developed, interactive, goal-seeking agent.

3.4. **Supervised learning**

Supervised learning is a machine learning technique to acquire information about the input-output correlations of a system. It is based on a set of paired input-output training examples. An input/output training sample is also known as labeled training data or supervised data since the output is thought of as the label of the input data or the supervision [13].

By using supervised learning, an artificial system may be built that can foresee its output in response to new inputs and can identify the relationship between input and output. If the output only accepts a finite set of discrete values that correspond to the input's class labels, the learned mapping defines how the input data is classed. Assuming continuous data, regression of the input produces results from the output. Learning-model parameters are widely used to describe the information about the input-output relationship. When these parameters cannot be obtained directly from training samples, a learning system must perform an estimating method to obtain these parameters. Supervised learning is different from research called Unsupervised Learning, the training data for Supervised Learning need to be supervised or labeled information, while the training data for unsupervised learning are unsupervised as they are not labeled.

A machine can learn how people act or object behaviors for specific tasks using supervised learning. The machine can then execute comparable operations on these tasks using the learnt information.

In the supervised learning paradigm, there are various methods for designing learning systems. This can be categorized as regression, naïve Bayes, and classification. Additionally, there are various algorithms used in supervised learning system techniques. Nearest Neighbor algorithms KNN, SVM, Decision Trees DT, Neural Networks NN, and, Naive Bayes NB [14].

3.4.1. **Regression**

A supervised learning method called regression is utilized to comprehend the relationship between reliable and independent variables. It is also a kind of supervised learning that gains knowledge from labeled data sets to forecast continuous results for various inputs in an algorithm. The model forecasts classes of new variables to which they belong, and it is thought to be commonly utilized in cases where the result must be a finite value [43].

There are two types of regressions: the first is linear regression used to identify the relation between two variables, and when the dependent variable is categorical or contains binary outcomes, such as "yes" or "no," logistics regression is utilized.

3.4.2. **Naïve Bayes**

A branch of supervised learning used for huge datasets is the Naive Bayes method. The strategy's foundation is the algorithm's independent operation of each program. This indicates that having one trait does not affect having the other.

Generally, it is used in text classification and recommendation systems [45].

3.4.3. **Classification**

A supervised machine learning technique called classification is used to predict group membership for individual data instances [13]. Classification is the technique to machine learning that is most widely used, despite the fact that there are numerous others. Classification is a well-known challenge in machine learning, especially in the context of information discovery and long-term planning.

Classification is categorized as one of the supremely studied problems by researchers in the machine learning and data mining fields. The commonly used classification techniques include Decision Trees (ID3 and C4.5), Bayesian Networks, K-Nearest Neighbors, Artificial Neural Networks - ANN, and Support Vector Machines.

3.4.3.1.**Decision Tree Induction**

The most used algorithms in classification are decision tree algorithms [13]. A decision tree provides a modelling tool that is simple to comprehend and streamlines the categorization process [11]. The decision tree is a transparent mechanism it facilitates users to follow a tree structure easily in order to see how the decision is made [13] [46]. In this section, the study specifically describes decision tree methods and has discussed their strengths, limitations, and applications. The main goal of the decision tree is to produce a model that calculates the value of a required variable based on numerous input variables [12] [13]. Usually, most decision tree algorithms are built in two stages [13] [42].

Stage 1. Tree growth; in which training set based on local optimal criteria is splitting recursively until most of the records belonging to the partition have the same class label [13] [42].

Stage 2. Tree pruning; in which the size of the tree is reduced making it easier to understand [10]. In this section, we focus on ID3 and C4.5 decision tree algorithms.

The name "ID3" (Iterative Dichotomiser 3) refers to the algorithm's iterative (repeated) dichotomization (dividing) of attributes into two or more groups at each phase. This decision tree algorithm was first introduced in 1986 [44] [13]. Due to its efficiency and simplicity, it

is one of the techniques that is frequently employed in the fields of data analysis and machine learning [13] [27]. This model can easily understand and in the final decision whole training example is considered, but, weaknesses include no backtracking According to et al. [13], C4.5 is a famous algorithm for decision tree production. It is an expansion of the ID3 algorithm and it minimizes the drawbacks caused by ID3. In the pruning stage, C4.5 tries to eliminate the un-comfort branches by swapping them with leaf nodes by going back through the tree once it has been generated [13] [47]. The capabilities of C4.5 include pre- and post-pruning, dealing with training data with missing feature values, and dealing with both discrete and continuous features. [10] [48].

ID3 has the following Metrics. In order to determine the optimal feature, the ID3 method uses Information Gain, or simply Gain, at each stage of creating a decision tree.

Information Gain assesses how well a certain characteristic categorizes or separates the target classes and determines the reduction in entropy. The best feature is determined by its Information Gain score. Entropy is a measure of disorder in general, and the entropy of a dataset is a measure of instability in its target feature.

D3 Entropy equals 0 if all the values are homogenous (similar) and 1 if there are an equal number of values in each class when there are only two classes in a target column for binary classification.

Our dataset is designated as S, and entropy is determined as follows:

Entropy(S) equals - pi * log2 (pi); i ranges from 1 to n.

Where,

In our example, n = 2, or YES and NO, is the total number of classes in the target column. Pi is the probability of class "i" or the proportion of "the total number of rows in the dataset" to "the number of rows with class i in the target column".

ID3 Steps

I. Determine the amount of knowledge obtained from each feature.
II. Since not every row belongs to the same class, split dataset S into smaller groups by utilizing the characteristic for which information gain is maximum.
III. Use the attributes that provide the greatest information to create a decision tree node.
IV. If all of the rows belong to the same class, create a leaf node with the class as its label for the current node.
V. Continue until either the decision tree has all leaf nodes or we run out of features, whichever happens first.

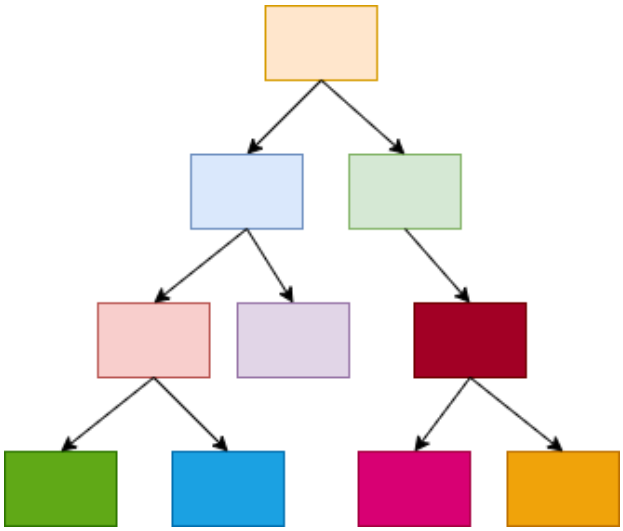


Figure 3-1 Decision Tree [48]

3.4.3.2. **Random Forest (RF)**

Random Forest is a collection of decision trees. According to their name, trees are constructed by randomly choosing m attributes for each tree node to get around the overfitting issue with DT, the [44]. Choose N number of attributes at random from a total of M number of attributes/features, build the first tree, then repeat this process K times to generate the desired K number of forests. There will be an equal number of randomly chosen cases in each tree. Every single tree functions in the same way as any other decision tree, with the selection of the root node and decision nodes requiring calculation. The performance of every single tree in the forest and its correlation with other trees determine the outcome of a random forest machine learning system. The algorithm's final result is established by which tree received the most votes, although each tree in the forest made its own choice.

One of the key characteristics of the Random Forest Algorithm is its ability to handle data sets with both continuous variables, as in regression, and categorical variables, as in classification. It performs better results for classification problems [43].

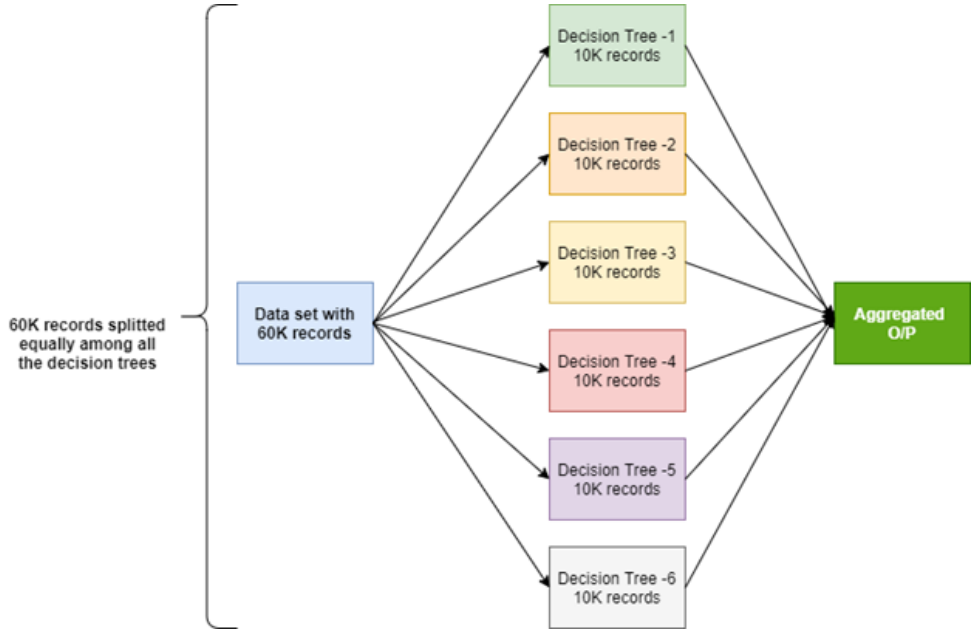


Figure 3-2 the aggregation of Random forest [65]

3.4.3.3. **K-Nearest-Neighbors (k-NN)**

K-nearest neighbors (KNN) is a kind of supervised learning technique used for both regression and classification. By computing the distance between the test data and all of the training points, KNN attempts to predict the appropriate class for the test data. Then select the K number of points that is close to the test data. The KNN algorithm calculates the probability of the test data belonging to the classes of 'K' training data and the class that holds the highest probability will be selected. Currently, the K-NN algorithm is the most widely used when doing classification due to its simplicity in the way the algorithm works and in the way of implementing it. As long as all of the values are continuous and the NA values are removed or changed into values. This goes for most of the algorithms using numeric calculations to do the predictions, there are also some that use text for pattern recognition [2]. There are two main obstacles with nearest neighbor-based classifiers are highlighted in [14] which include; space requirement and classification time.

The advantages of KNN include simplicity, transparency, Robust to noisy training data, easy to understand and implementation, and disadvantages include. Computation complexity, memory limitation, poor runtime performance for large training sets, and irrelevant attributes can cause problems.

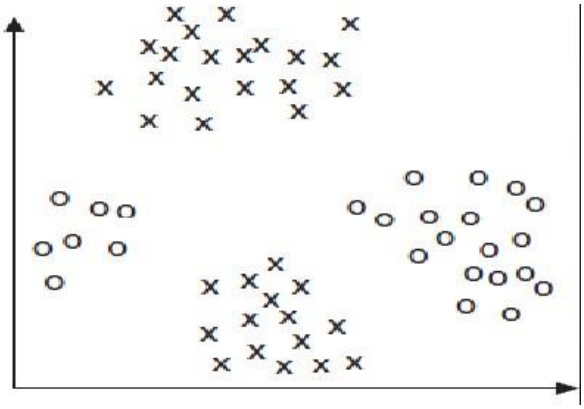


Figure 3-3. K-Nearest-Neighbors. ***Source****: SemanticsScolar.com*

3.4.3.4. **Support Vector Machine (SVM)**

One of the most well-known and practical methods for addressing issues relating to the classification of data is the use of SVM. Support Vector Machine (SVM) is a supervised machine learning algorithm that may be used for both regression and classification. SVMs are a new potential non-linear, non-parametric classification method that is ideal for binary and multi-class classification tasks [22]. SVM is a very widely used algorithm is when doing classification, this is mainly due to it is good accuracy and that it can be used to do multiclass classification while automatically avoiding overfitting the training data [14]. The SVM classifier reduces the computing complexity by classifying data from a small fraction of the entire training set. By employing the kernel trick, the reduction is attained, and overfitting of the data is prevented by classifying with a maximum margin [47].

Many research papers can show accuracy in the numbers close to 100 % which would be the wanted outcome but is almost impossible to achieve at least in a real-life scenario [43]. SVM can be used in many ways and there are several parameters that are given to the model. One can use SVM using a linear kernel, what this means explained in a very simple matter is that the hyperplanes are linear. In essence, SVM identifies a hyper-plane that establishes a distinction between the various types of data. This hyper-plane is only a line in two-dimensional space. Each dataset item is plotted in an N-dimensional space using SVM, where N is the total number of features and attributes in the dataset. The best hyperplane should then be found to divide the data. Only binary classification, or choosing between two classes, is possible using SVM. The SVM classifier has three main ideas: Maximum margin separator: draw the line or hyperplane that maximizes the distance between the separator and the training data, thus introducing a margin slab, a soft margin separator (draw the best separator line when data with various labels are mixed together while accounting for the samples within the

margin slab), and a kernel trick: for more complex models in which the data separation boundary is not linear, allow for higher-order polynomials or even not polynomial functions [49].

In this study, the training phase is used to draw the line at which data can be considered normal. The cases are compared to that zone during testing, and if they fall inside it, they are labeled as normal, otherwise fraudulent. The basic goal is to as much as possible reduce the number of points in the margin.

Consider a general two-class classification problem of assigning a class label $y \in \{-1, +1\}$ to an input feature vector $\mathbf{x} \in R^N$. We are given input-output training data pairs $\{(\mathbf{x}_1, y_1) \ldots (\mathbf{x}_n, y_n)\}$.

S(x) = w1.xj1 +w2.xj2 +w3.xj3 +w4.xj4 +.... + ++ wn.xjn + b (3.5)

The classification function in an SVM classifier can be expressed as follows [52, 73]:

S(x) = xTjwj + (b)

Where S(x) is a linear discriminant function, X is a feature vector used for classification, W is the hyperplane's weight, and B is the bias governing the hyperplane's position.

Maximum Margin Separator

The objective of SVM calcification is to find the most optimal line / hyperplane which separates the points correctly thereby each point correctly representing the class it belongs to. This problem can be solved using two different algorithms. One of them maximum margin separator of SVM.

When using SVMs, the goal of optimization is to increase the margin so that the points may be correctly classified. The margin is the angle at which this line or hyperplane is perpendicular to the closest points on each side that correspond to various classes.

It is important to understand the concept of hyperplane to understand the concept of SVM before understanding the Maximal Margin Classifier in SVM. It is basically a boundary that separates the dataset into different classes. In Figure 3.4.3.4, the description of maximum margin separator. The remaining variables are not crucial for creating the model because the answer solely depends on the support vectors.

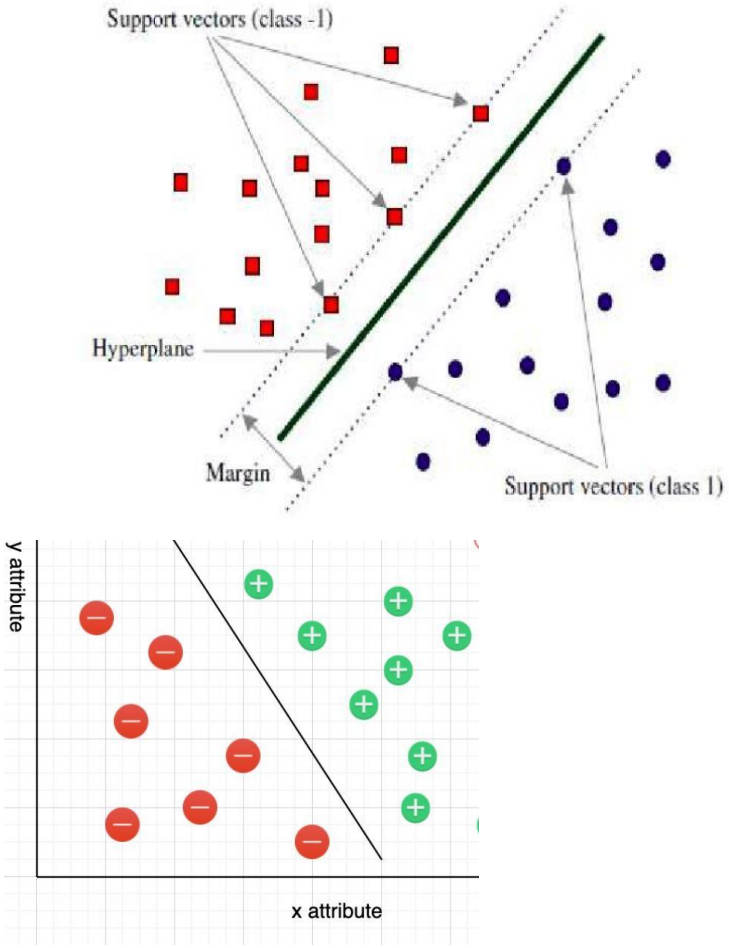


Figure 3-4. SVM Classification [14] [43] [37]

The equation of a line is

$$ax + by + c = 0$$

Where a, b are coefficients.

We can generalize the hyperplane as below:

$$w_1 x_1 + w_2 x_2 + w_0 = 0$$

Where

w_1 And w_2 are weights.

x_1 And x_2 are attributes.

The data points belong to the positive side are strictly $ax + by + c > 0$ and point belong to another side of hyperplane us$ax + by + c < 0$.

The math behind the distance between a point(x) and hyperplane**.**

A hyperplane is defined as:

$$w^T x + b = 0$$

Let's consider a point 'X' and 'd' is the distance of the point from the hyperplane and perpendicular to the plane.

x^h is the point on the hyperplane and the distance between X and x^h is below.

$$x - x^h = d$$

As the perpendicular line is parallel to w, we can rewrite it as below.

$d = \alpha w$ Where α is a constant

As x^h lies on the plane $w^T x + b = 0$

Therefore

$$w^T(x - d) + b = 0$$

$$w^T(x - \alpha w) + b = 0$$

$$\alpha = \frac{w^T x + b}{w^T w}$$ ____> Equation-1

$$\|d\| = \sqrt{d^T d}$$

As $d = \alpha w$

$$\|d\| = \sqrt{\alpha^2 w^T w}$$

$$\|d\| = \alpha\sqrt{w^T w}$$

Replacing α as Equation-1

$$\|d\| = \left(\frac{w^T x + b}{w^T w}\right)\sqrt{w^T w}$$

$$\|d\| = \left(\frac{w^T x + b}{\sqrt{w^T w}}\right)$$

$$\|d\| = \left(\frac{w^T x + b}{\|w\|}\right)$$

The same distance can also be found using the distance rule.

Based on the below rule to find the distance from any point (x_0, y_0) to a

line $Ax + By + C = 0$

$d = \frac{Ax_0+By_0+C}{\sqrt{A^2+B^2}}$

Following the above rule, the distance of the hyperplane will be

$\|d\| = \left(\frac{w^T x+b}{\|w\|}\right)$

Now let's maximize the margin such that each data point can be classified correctly.

We know that data points on either side of the hyperplane should follow the below criteria.

$w^T x + b \geq +1$ Where $y_i = +1$

$w^T x + b \leq -1$ Where $y_i = -1$

Both can be combined into:

$y_i(w^T x + b) \geq 1$ ——> Equation-2

To maximize margin $\|d\|$

$$\max_{w,b} \frac{1}{\|w\|} \min_x w^T x + b$$

Based on Equation-2,

$\min_x w^T x + b = 1$

Substituting the above value, we get

$\max_{w,b} \frac{1}{\|w\|} * 1$

$= \min_{w,b} \|w\|$

We can rewrite this as below.

$= \min_{w,b} \sqrt{w^T w}$

This is a constraint optimization problem and this can be solved Lagrangian multiplier method. Upon solving this equation we get below.

$L = \frac{1}{2}\|w^2\| - \sum \alpha_i[y_i(wx_i + b) - 1]$

$: w - \sum \alpha_i y_i x_i$

$w = \sum \alpha_i y_i x_i$

Once we find **'w'** then we can find the distance (**d**) which is nothing but the margin

CHAPTER FOUR

PROPOSED APPROACH

4. Introduction

In this section, the study presents the experimental process of the smishing detection Model. Figure 4.0.1 depicts the major phases of the evaluation process:

Stage 1: Data collection and Dataset creation, which is collecting data and creating a dataset based on the identified categories in CDR data Stage 2: preprocessing, which is to analyzing, data clearing, data integration, and aggregating the collected dataset in stage 1. Stage 3: Evaluation and Prediction, which is to train, and build a model based on relevant attributes, and understand the behavior of each smishing category. Details of the tasks done under these modules are described in the coming sections.

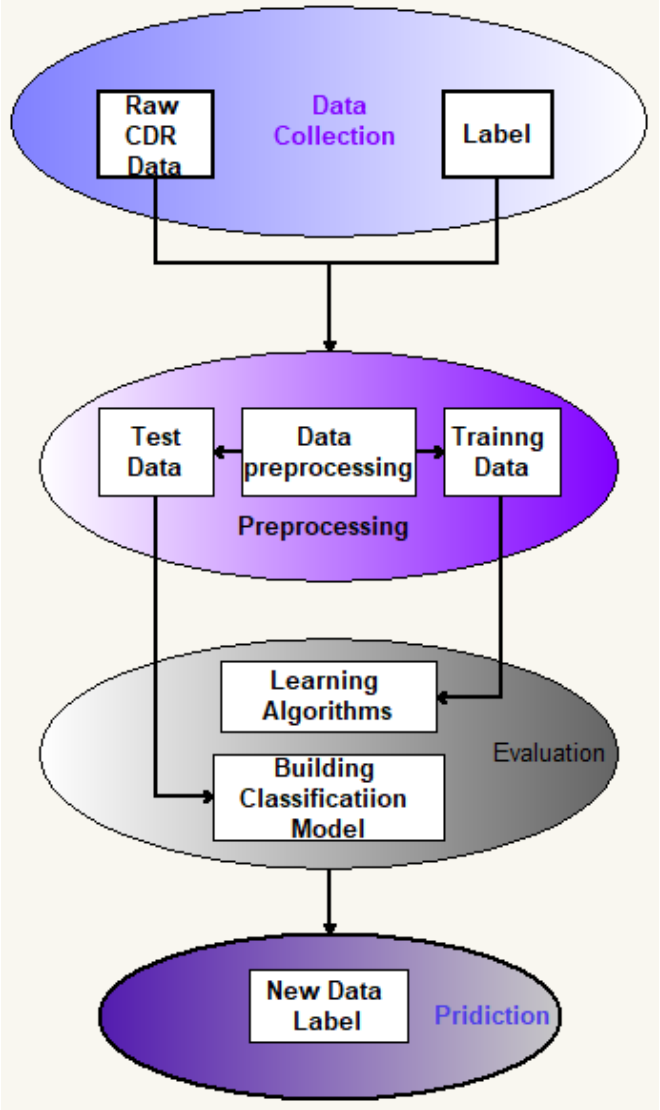


Figure 4-1 System Model [22] [14]

4.1. **Data collection**

The study collected a large number of Smishing samples representing more than three smishing families. The accumulated dataset samples from ethio telecom customers' CDR data, samples provided by contact center advisers received from customers, as well as samples provided by multi-channel services blogs, and anti-fraud sections finally these samples are analyzed and confirmed by the fraud specialist section.

The collected dataset includes 52400 unique samples spanning a five-month period from March 01- 2022 to September 30-2022 and 157200 normal data in the same period.

The research examined the dataset with a different storage location because the CDR data size is enormous to assure the proper labeling of samples. Oracle's 19th generation version database was installed on Windows Server 2012 R2 specifications V.6.3.9600 with 8GB RAM, a quad-core processor, 2.5 TB of storage, and 2.5 TB of mounted storage.

Each day, the CDR data is gathered, stored, and loaded into the database using an automatic data loader script in the following format: figure 4-2.

Figure 4-2. Dumped CDR data

Because the CDR data is so large, more storage space is needed. Other tables are developed for typical behavioral data, smishing behavior data based on the types of service SMS table, data table, and voice table with specified attributes. Tables are created for uploading the imputation dumped CDR data as it is the record in the network. Daily gathered text format data are batch-filed into a table, and this table is then translated into another table with additional batch files using the service's chosen attributes. The original dumped text format CDR data are then deleted after loading these data into their relevant table.

4.2. **Understanding CDR Data**

The analysis proceeds in this part using CDR data from the Ethiopian telecom monitoring system. Giving the database the imputed CDR data is the first step in the research. It then goes into detail on the tasks the study performs in collaboration with subject-matter experts to comprehend and assess the data using their chosen attribute values relevant to the study's

issue domain. Additionally, we go over checking the data's usefulness, completeness, redundancy, missing values, and acceptability of attribute values in light of the ML objectives. Call Detail Records (CDR) are an all-encompassing and flexible data source that telephone service providers initially used for billing. Each record includes the mobile phone user's transaction time and location. Encrypted user ID, location area code, cell ID, timestamp, billing, Recharge channels, Mobile Station International Subscriber Directory Number (MSIDN), Service Type, and event type as shown in Figure 4.3 are among the components of CDR data.

A call detail record additionally contains data fields that characterize a specific instance of a telecommunication transaction to the information previously provided. In actuality, call detail records are far more extensive in modern usage, and they include data such as:

- ✓ the subscriber's phone number (the "calling party," or "A-party")
- ✓ The recipient phone number (called party, B-party)
- ✓ the call's start time (date and time)
- ✓ the call duration
- ✓ the phone line used for billing that will be billed for the call
- ✓ the name of the phone company or piece of equipment that created the record
- ✓ the record's individual sequence number
- ✓ the extra digits that were added to the calling number to route or bill the call
- ✓ the outcome of the call, such as whether or not the call was connected, or its disposition
- ✓ the path the call took to reach the exchange
- ✓ Call type (voice, SMS, etc.)
- ✓ voice call types (setup, continuation, operation, termination, idle, busy, and out-of-service calls),

Service_type	Billing_Number	Calling_Number	Outgoing_call	Called_Number	Start_Time	End_Time	Actual_Volume	Measure_Unit	Actu..
On-net SMS	912020303.0...	25191202...	25192...	251910635...	3-Apr-2022...	3-Apr-2022 1...	127	Item	0
On-net Voice	920232418.0...	25192023...	25192...	251910635...	3-Apr-2022...	3-Apr-2022 1...	37	Second	0
On-net Voice	946462019.0...	25194646...	25194...	251910635...	3-Apr-2022...	3-Apr-2022 1...	100	Second	1
On-net Voice	946462019.0...	25194646...	25197...	251910635...	3-Apr-2022...	3-Apr-2022 1...	22	Second	0
On-net Voice	946462019.0...	25194646...	25192...	251910635...	3-Apr-2022...	3-Apr-2022 1...	13	Second	0
On-net Voice	946462019.0...	25194646...	25196...	251910635...	3-Apr-2022...	3-Apr-2022 1...	191	Second	1
On-net Voice	942588258.0...	25194258...	25199...	251910635...	3-Apr-2022...	3-Apr-2022 1...	39	Second	0
National ro...	946462019.0...	25194646...	25199...	251910635...	3-Apr-2022...	3-Apr-2022 0...	65	Second	1

Figure 4-3 Screenshot of CDR data

In order to obtain the longitude and latitude of the cell tower, LAC and Cell ID must be linked to the cell tower database to jointly establish the location (coordinates) of the cell tower. The event ID keeps track of the transaction's kind, which often includes calls in and out, messages, and web browsing.

The CDR also retains additional data that can be used to infer the socio-demographic characteristics of the individuals in addition to the spatial and temporal data already described. For instance, researchers can deduce the sort of phone the user is using, along with the brand, manufacturer, model, and system. As a stand-in for the user's disposable income, this can be employed. Additionally, by learning the person's registration country and city, it is possible to deduce the user's nationality or hometown. There are several uses for inferring the user's attributes.

4.2.1 Why Are CDRs Important?

A CDR log records every billable communication sent through your phone system. This allows phone companies to generate your phone bills and lets you keep definitive records of how and when your phone system was used. In order to help with call reporting and invoicing, businesses generally employ them.

Billing departments utilize CDRs to record phone system usage, resolve conflicts, and keep tabs on how funds are spent. CDRs can be used by IT teams to check for phone service interruptions.

CDRs can be used to determine calling patterns and learn more about how employees use their phones. This enables managers to see patterns and trends and improve management and hiring decisions.

4.3. Data Selection

To create a dataset that is based on the identified category in calcification. Creating a target data set is what this stage is all about. All attributes Stored from CDR data are not useful for the study processing the entire set of data is not advisable for economical and practical reasons. We must first determine the parameters that could help us achieve the study's goal before we can choose qualities. An algorithm's evaluated output is dependent on the carefully chosen input data.

Under this subsection of data selection, we followed the same procedure of data collection and labeling. Five months of mobile network CDR data were selected. Following the generation of CDR data, the study concentrates on the data sample used for the study. After

assessing the usefulness of the data, the study chooses pertinent qualities that could identify subscriber smishing behavior. Irrelevant data from collected CDR data are removed.

Figure 4-4 screenshot of only SMS data sample

Service_type	Billing_Number	Calling_Number	Outgoing_call	Called_Number	label	msisdn	age	daily_spent_1m	daily_spent_5m	Avg_main_bal1m	Avg_main_bal5m	las_re ch_	last_rech_date_da	last_rech_amt_ma
On-net Voice	960517036.0000...	251960517...	0.000000000000...	25197338...	0	21408I...	272.0000	3055.050000...	3065.150000000...	220.13	260.13	2.0000	.0000	1539
On-net Voice	962360145.0000...	251962360...	2.519495564360...	25197338...	1	76462I...	712.0000	12122.00000...	12124.750000...	3691.26	3691.26	20.0...	.0000	5787
On-net Voice	962360145.0000...	251962360...	2.519224512300...	25197338...	1	17943I...	535.0000	1398.000000...	1398.000000000...	900.13	900.13	3.0000	.0000	1539
On-net Voice	962360145.0000...	251962360...	2.519313774290...	25197338...	1	55773I...	241.0000	21.22800000...	21.228000000...	159.42	159.42	41.0...	.0000	947
On-net Voice	928849360.0000...	251928849...	2.519313774290...	25197338...	1	03813I...	947.0000	150.6193333...	150.619333300...	1098.90	1098.90	4.0000	.0000	2309
On-net Voice	928524540.0000...	251928524...	2.519787937910...	25197338...	1	35819I...	568.0000	2257.362667...	2261.460000000...	368.13	380.13	2.0000	.0000	1539
On-net Voice	928524540.0000...	251928524...	0.000000000000...	25197338...	1	96759I...	545.0000	2876.641667...	2883.970000000...	335.75	402.90	13.0...	.0000	5787
On-net Voice	928524540.0000...	251928524...	0.000000000000...	25197338...	1	59772I...	1191.0000	90.69500000...	90.695000000...	2287.50	2287.50	1.0000	.0000	1539
On-net Voice	929253206.0000...	251929253...	0.000000000000...	25197338...	1	32893I...	1511.0000	12.89600000...	12.896000000...	790.44	790.44	8.0000	.0000	1539
On-net Voice	928849360.0000...	251928849...	0.000000000000...	25197338...	0	82417I...	82.0000	65.16666667...	65.166666670...	326.20	326.20	17.0...	.0000	7526
On-net Voice	928849360.0000...	251928849...	0.000000000000...	25197338...	1	11435I...	154.0000	227.0410000...	227.041000000...	240.41	240.41	2.0000	.0000	1547
On-net Voice	960517036.0000...	251960517...	0.000000000000...	25197338...	1	66580I...	887.0000	55.90933333...	55.909333330...	208.80	208.80	2.0000	.0000	1539
On-net Voice	929253206.0000...	251929253...	0.000000000000...	25197338...	1	63139I...	707.0000	8919.000000...	10317.350000...	399.25	2453.78	3.0000	.0000	770
On-net Voice	900878444.0000...	251900878...	0.000000000000...	25197338...	0	24075I...	1037.0000	12.00000000...	12.000000000...	1216.80	1216.80	.0000	.0000	0
On-net Voice	960020471.0000...	251960020...	2.519641243230...	25197338...	0	82053I...	1583.0000	1000.000000...	1000.000000000...	1000.80	1087.88	.0000	.0000	0
On-net Voice	986370210.0000...	251986370...	0.000000000000...	25197338...	1	37204I...	929.0000	10.68800000...	10.688000000...	40.00	40.00	.0000	.0000	0
On-net Voice	960448493.0000...	251960448...	2.519397712110...	25197338...	1	44217I...	832.0000	14.40000000...	14.400000000...	1660.96	1660.96	1.0000	.0000	2309
On-net Voice	960448493.0000...	251960448...	2.519742353760...	25197338...	1	19611I...	450.0000	48.93500000...	48.935000000...	726.30	726.30	1.0000	.0000	1539
On-net Voice	960448493.0000...	251960448...	2.519108456270...	25197338...	1	67813I...	100.0000	769.6140000...	777.460000000...	1050.57	1167.30	6.0000	.0000	770
On-net Voice	901144560.0000...	251901144...	2.519742353760...	25197338...	0	75522I...	378.0000	514.6933333...	515.200000000...	56.26	58.20	2.0000	.0000	773
On-net Voice	952832560.0000...	251952832...	0.000000000000...	25197338...	1	61590I...	463.0000	1540.000000...	1541.000000000...	969.12	969.12	4.0000	.0000	770

Figure 4-5 screenshot of only voice data sampe

Service_type	Billing_Number	Calling_Number	Outgoing_call	Called_Number	label	msisdn	age	daily_spent_1m	daily_spent_5m	Avg_main_bal1m	Avg_main_bal5m	las_re ch_	last_rech_date_da	last_rech_amt_ma
On-net SMS	928524540.0000...	251928524...	0.000000000000...	25197338...	1	09832I...	768.0000	12905.00000...	17804.150000...	900.35	2549.11	4.0000	55.0000	3178
On-net SMS	929253206.0000...	251929253...	0.000000000000...	25197338...	1	56331I...	536.0000	29.35733333...	29.357333330...	612.96	612.96	11.0...	.0000	773
On-net SMS	8686.0000000000	0	2.519641243230...	25194373...	1	90392I...	291.0000	33.82000000...	33.820000000...	1106.40	1106.40	3.0000	.0000	770
On-net SMS	8686.0000000000	0	2.519208345010...	25194373...	1	14958I...	1756.0000	78.12000000...	78.120000000...	295.60	295.60	3.0000	.0000	1539
On-net SMS	989318042.0000...	251989318...	2.519742353760...	25194373...	1	49700I...	493.0000	8674.670000...	10711.490000...	4622.56	5008.54	4.0000	.0000	770
On-net SMS	962311609.0000...	251962311...	2.519208345010...	25194373...	1	98423I...	1283.0000	13776.05300...	13813.510000...	2069.15	3569.39	2.0000	.0000	773
On-net SMS	962311609.0000...	251962311...	2.519131584970...	25194373...	1	86227I...	693.0000	2600.000000...	2600.000000...	1243.61	1243.61	8.0000	.0000	1539
On-net SMS	8686.0000000000	0	2.519313774290...	25194373...	1	86231I...	274.0000	105.7943333...	105.794333300...	730.84	730.84	1.0000	.0000	2309
On-net SMS	8686.0000000000	0	2.519397712110...	25194373...	0	88380I...	1102.0000	23.13300000...	77.110000000...	14351.27	19634.66	33.0...	.0000	5787
On-net SMS	962872477.0000...	251962872...	2.519388958240...	25194373...	1	89125I...	175.0000	37.70000000...	37.700000000...	2086.84	2086.84	6.0000	.0000	1539
On-net SMS	8686.0000000000	0	0.000000000000...	25194373...	1	36283I...	624.0000	9837.000000...	14591.890000...	4413.43	11014.33	11.0...	.0000	3178
On-net SMS	8686.0000000000	0	2.519742353760...	25194373...	1	58963I...	246.0000	7892.000000...	10312.260000...	2316.20	5220.25	2.0000	.0000	770
On-net SMS	8686.0000000000	0	2.519641243230...	25194373...	1	21241I...	1162.0000	56.66666667...	56.666666670...	398.14	398.14	1.0000	.0000	1539
On-net SMS	947543834.0000...	251947543...	2.519742353760...	25194322...	1	86850I...	71.0000	19.12033333...	19.120333330...	1889.17	1889.17	20.0...	.0000	1547
On-net SMS	910452622.0000...	251910452...	2.519721292410...	25194322...	0	24419I...	126.0000	27.90000000...	27.900000000...	1117.80	1117.80	2.0000	.0000	1539
On-net SMS	905465548.0000...	251905465...	2.519208345010...	25194322...	1	17348I...	2282.0000	6809.000000...	7572.950000000...	1642.88	1933.88	7.0000	.0000	2309
On-net SMS	922887030.0000...	251922887...	2.519853029540...	25194572...	1	60826I...	1215.0000	37.03666667...	37.036666670...	1345.70	1345.70	4.0000	.0000	2309
On-net SMS	922887030.0000...	251922887...	0.000000000000...	25194572...	1	24240I...	355.0000	12502.00000...	12907.870000...	5890.77	6130.77	1.0000	.0000	1539
On-net SMS	922887030.0000...	251922887...	0.000000000000...	25194572...	0	11094I...	207.0000	.0000000000...	.000000000000...	1644.66	1644.66	.0000	.0000	0
On-net SMS	915308633.0000...	251915308...	2.519801962840...	25194572...	1	89807I...	1305.0000	8704.000000...	10617.730000...	8323.53	17344.64	7.0000	.0000	1547
On-net SMS	965133404.0000...	251965133...	2.519742353760...	25193690...	1	46118I...	231.0000	17811.50000...	17943.000000...	5342.57	7919.87	1.0000	.0000	1539

Figure 4-6 Screenshot of only long distance voice data sample

Service_type	Billing_Number	Calling_Number	Outgoing_call	Called_Number	label	msisdn	age	daily_spent_1m	daily_spent_5m	Avg_main_bal1m	Avg_main_bal5m	las_rech_	last_rech_date_da	last_rech_amt_ma
Long dista...	962698583.0000...	251962698...	2.519224512300...	25194373...	1	75630I...	817.0000	3680.000000...	4435.5000000...	299.69	879.95	1.0000	.0000	1539
Long dista...	934645881.0000...	251934645...	2.519495564360...	25194373...	1	42449I...	433.0000	10.96300000...	10.963000000...	1967.83	1967.83	3.0000	.0000	773
Long dista...	934645881.0000...	251934645...	2.519495564360...	25194373...	1	01210I...	725.0000	9740.000000...	16337.040000...	299.96	1457.22	15.0...	63.0000	5787
Long dista...	962698583.0000...	251962698...	2.519721292410...	25194373...	1	28146I...	133.0000	8.400000000...	8.4000000000...	231.60	231.60	3.0000	.0000	773
Long dista...	962835550.0000...	251962835...	2.519397712110...	25194373...	1	60754I...	875.0000	1619.048000...	1620.4400000...	336.18	387.90	11.0...	.0000	1924
Long dista...	902452777.0000...	251902452...	2.519224512300...	25194373...	1	41026I...	138.0000	67.67933333...	67.679333330...	242.00	242.00	6.0000	.0000	2309
Long dista...	941697427.0000...	251941697...	2.519313774290...	25194373...	0	27460I...	766.0000	-.250000000...	-.2500000000...	.00	.00	.0000	.0000	0
Long dista...	947543834.0000...	251947543...	0.000000000000...	25194322...	1	34684I...	233.0000	11394.00000...	12219.420000...	6675.38	9197.12	3.0000	69.0000	1924
Long dista...	905425469.0000...	251905425...	2.518220000000...	25194322...	1	30030I...	827.0000	5461.000000...	6014.7700000...	300.26	717.56	24.0...	46.0000	3178
Long dista...	947543834.0000...	251947543...	2.518220000000...	25194322...	1	56589I...	111.0000	1820.000000...	2221.4000000...	1068.68	1520.18	24.0...	.0000	770
Long dista...	927529823.0000...	251927529...	2.518220000000...	25194322...	1	91716I...	365.0000	44836.00000...	52266.890000...	6394.29	8163.82	1.0000	.0000	2309
Long dista...	948798903.0000...	251948798...	0.000000000000...	25194322...	0	64987I...	274.0000	800.0000000...	800.00000000...	.00	.00	.0000	.0000	0
Long dista...	941475965.0000...	251941475...	0.000000000000...	25194322...	1	24480I...	1094.0000	3207.000000...	3207.0000000...	900.13	900.13	1.0000	.0000	1539
Long dista...	948798903.0000...	251948798...	0.000000000000...	25194322...	1	43748I...	1312.0000	12699.00000...	14512.920000...	4687.74	5126.86	1.0000	.0000	1539
Long dista...	948798903.0000...	251948798...	2.518220000000...	25194322...	1	69235I...	1268.0000	5399.852000...	5494.1300000...	3314.97	5736.51	7.0000	.0000	2309
Long dista...	941475965.0000...	251941475...	2.518220000000...	25194322...	1	07684I...	625.0000	2005.833333...	2007.0000000...	75.00	90.00	11.0...	.0000	2309
Long dista...	943597834.0000...	251943597...	2.518220000000...	25194322...	1	33970I...	710.0000	23.52133333...	23.521333330...	1152.76	1152.76	1.0000	.0000	1539
Long dista...	905465548.0000...	251905465...	2.518220000000...	25194322...	1	43430I...	164.0000	33750.00000...	39567.480000...	3301.65	5061.54	1.0000	.0000	1539
Long dista...	905465548.0000...	251905465...	2.518220000000...	25194322...	1	48581I...	397.0000	38.78000000...	38.780000000...	560.56	560.56	1.0000	.0000	4048
Long dista...	947543834.0000...	251947543...	2.518220000000...	25194322...	1	28734I...	1221.0000	95.99333333...	95.993333330...	374.17	374.17	1.0000	.0000	1539
Long dista...	941475965.0000...	251941475...	2.518220000000...	25194322...	1	73196I...	198.0000	18301.00000...	28936.470000...	8634.10	11994.34	1.0000	.0000	4048

Figure 4-7 Screenshot of only National Roaming Service sample

Service_type	Billing_Number	Calling_Number	Outgoing_call	Called_Number	label	msisdn	age	daily_spent_1m	daily_spent_5m	Avg_main_bal1m	Avg_main_bal5m	las_rech_	last_rech_date_da	last_rech_amt_ma
National ro...	925008410.0000...	251925008...	2.519742353760...	25193690...	1	71629I...	1080.0000	31.000000000...	31.000000000...	122.40	122.40	7.0000	.0000	1539
National ro...	912126057.0000...	251912126...	2.519853029540...	25193690...	1	71767I...	914.0000	148.1666667...	148.16666670...	-98.60	-98.60	7.0000	.0000	1924
National ro...	910670274.0000...	251910670...	2.519284995750...	25193690...	1	77712I...	1251.0000	36.30733333...	36.307333330...	1559.14	1559.14	1.0000	.0000	2309
National ro...	965133404.0000...	251965133...	2.519721292410...	25193690...	1	63273I...	356.0000	77.61000000...	77.610000000...	557.60	557.60	1.0000	.0000	11874
National ro...	965133404.0000...	251965133...	2.519721292410...	25193690...	1	08108I...	1151.0000	19.80000000...	19.800000000...	1653.75	1653.75	4.0000	.0000	770
National ro...	965133404.0000...	251965133...	2.519853029540...	25193690...	1	79293I...	344.0000	121.7850000...	121.78500000...	278.53	278.53	7.0000	.0000	4048
National ro...	965133404.0000...	251965133...	2.519283178070...	25193690...	1	03027I...	-37.0000	9.160000000...	9.1600000000...	203.04	203.04	1.0000	.0000	770
National ro...	901488602.0000...	251901488...	2.519283178070...	25193690...	1	47359I...	692.0000	1096.540000...	1097.8000000...	145.00	150.00	2.0000	.0000	770
National ro...	931123496.0000...	251931123...	2.519284995750...	25193690...	1	44284I...	319.0000	20.74600000...	20.746000000...	21.34	21.34	10.0...	.0000	773
National ro...	925008410.0000...	251925008...	2.519853029540...	25193690...	1	87592I...	211.0000	970.7733333...	972.80000000...	3655.20	3764.00	2.0000	.0000	1539
National ro...	965133404.0000...	251965133...	2.519283178070...	25193690...	1	70656I...	539.0000	10425.00000...	14410.680000...	1817.78	5755.38	2.0000	.0000	1539
National ro...	912846169.0000...	251912846...	2.519853029540...	25193690...	1	85447I...	90.0000	5642.573333...	5664.3600000...	2526.46	2804.66	2.0000	.0000	770
National ro...	968595288.0000...	251968595...	0.000000000000...	25193690...	1	42360I...	172.0000	13826.78000...	13861.670000...	5949.58	7224.28	1.0000	.0000	2309
National ro...	983067895.0000...	251983067...	0.000000000000...	25193690...	1	36827I...	1314.0000	59.02933333...	59.029333330...	1376.32	1376.32	2.0000	.0000	773
National ro...	912846169.0000...	251912846...	2.519397712110...	25193690...	0	06194I...	595.0000	115.8776667...	115.87766670...	1652.56	1652.56	6.0000	.0000	8000
National ro...	910670274.0000...	251910670...	0.000000000000...	25193690...	1	54408I...	51.0000	13.63666667...	13.636666670...	127.56	127.56	3.0000	.0000	773
National ro...	909839717.0000...	251909839...	2.519801962840...	25193690...	0	40480I...	539.0000	23.40000000...	23.400000000...	965.70	965.70	2.0000	.0000	773
National ro...	912126057.0000...	251912126...	2.519397712110...	25193690...	1	79122I...	1204.0000	7783.466000...	7805.3300000...	1862.26	2058.82	1.0000	.0000	1539
National ro...	6353.0000000000	0	2.519965035650...	25193690...	0	14718I...	1000.0000	-1.73333333...	-1.733333333...	.00	.00	.0000	.0000	0
National ro...	983899397.0000...	251983899...	2.519313774290...	25193690...	1	10187I...	183.0000	31.84800000...	31.848000000...	1836.36	1836.36	8.0000	.0000	770
National ro...	983899397.0000...	251983899...	2.519110748960...	25193690...	1	39331I...	1308.0000	647.5263333...	650.83000000...	.00	.00	.0000	.0000	0

4.3.1. **Attribute Selection**

Finding a set of representative attributes from which to create a classification model for a specific job is the main challenge in machine learning. In the data mining process, the approach of attribute selection is utilized to reduce the amount of data. Data reduction reduces the size of the data to enable more efficient analysis.

There could be a lot of attributes in the data set. However, some of those characteristics might be superfluous or unnecessary. The aim of attribute selection is to find a minimum set of qualities and eliminate those that are unnecessary.

To solve problems efficiently attribute selection is important for many reasons: run time requirements, performance generalization, constraints, and interpreting problems [14]. The process of choosing a subset of features that are pertinent to the analysis is known as feature selection. Depending on the model input (numerical or categorical data) and the analysis

output (classification or regression), there are many strategies available. Table 1 lists some instances of the methods that are employed.

For this study, from imported CDR data 33 attributes are listed as shown in Table 4.1 with their description 13 selected. The chosen attributes reveal the usage patterns of subscribers. The remaining characteristics that weren't pertinent to the study's goal were eliminated.

Table 4-1 Selected Attributes

NO	Attributes	Reason
1	SERVICE _TYPE	To associate the service type
2	CALLING _NUMBER	for identifying the call's initiator's subscriber
3	CALLED_NUMBER	Identify the destination number
4	START_TIME & END_TIME	Duration of call start time/end time and last recharge date, last recharge amount
5	ACTUAL_VOLUM_RECHARGED	The actual amount of balance recharged
6	ACTUAL_CHARGE_AMOUNT	The actual amount spent on service
7	VOICE	To identify which is On-net Voice, long-distance, and National roaming Voice
8	ON-NET SMS	Describes how much SMS was send and received
9	RECHARGE_CHANNEL	To identify the recharge channels
10	TRANSFERRED_BALANCE	To identify the transfer balance rate
11	RECHARGE_BALANCE_SELF	Numbers of balance recharge by the user
12	TOTAL BALANCE	balance recharged by the user in the last 30 days
13	AVERAGE MAIN BALANCE	Average balance vs transfer rate in five month

The researcher assesses the information content of the attributes by assisting the domain expert in order to choose the best attributes from this first acquired dataset. This study depended on smishing SMS attack usage patterns with respect to normal usage behaviors.

To create a dataset with the help of domain experts and feature extraction which is based below identified category [14].

- *Service Number/ Indicator of a call*: the Indicator of Call Behaviors, especially the Selected Investigated Sample of outgoing, local or international Call Patterns

  SMS In and SMS out, voice, long distance, national roaming, call type
- *Time Dimension*: Describe time as the flow of past, present, and future events.
  Start Time, End Time, Calling number, Last Recharge Date, Balance Transfer Rate, SMS, and time of sent with respect to the called number, and calling Number.
- *Usage category*: Number of SMS items sent, Number of SMS received, Number of calls made (call frequency), number of incoming calls (incoming call frequency), Main account Recharged (frequency), frequency of main recharge channels, Number of balance recharged by the user vs balance transfer rate with five days.
  Called number, Calling Number, actual Volume, Recharge Channel, SMS, Service Type

4.3.2. **Sampling**

Before the study has even started, sample selection, a critical component of research design, can decide whether research questions will be answered. Before crossing to the data preprocessing phase, sampling is the primary task to be done. The purpose of sampling is to comprehend sampling from statistical data in order to foretell smishing behavior. For both practical and financial reasons, it is not advisable to process all of the CDR data because it is large data. In these cases, sampling is a necessity rather than another option. Sampling [44] [14] is defined as the selection of a subset of samples from a data set on which the analysis will be performed. Sampling allows for the application of sophisticated models while reducing processing time. There are numerous methods that can be used to conduct sampling, including stratifying sampling, cluster sampling, and random selection [31] [14]. Selection of sample size is not easy and straightforward forward it depends on the techniques that will apply, the computational resource, and the accuracy of the output. The amount of proportionalities between the categorized parties could vary depending on the problem domain area. The majority of academics suggested a ratio of 20% bogus numbers to 80% genuine numbers [14] [39]. We decided to use it in a similar sample selection manner.

The proposed work is evaluated by taking from a total of 61 million subscribers' stored data in a database and out of this data about 205413 legitimate numbers have been chosen using systematic sampling techniques. Each active subscriber has an equal probability to be selected in a database. The sample selection process is done in simple steps of random sampling each subject is independent of the selection of every other unit. Additionally, with the help of domain experts who identified fraudulent subscribers that are detected and suspended numbers are used for further experiment and analysis purposes. The research is going to use a total data preprocessing and feature selection of 205,413 sample subscriber numbers, and 52,400 high-usage fraudulent subscribers were identified. Table 4.2 demonstrates the subscriber sample size of valid and phony service numbers.

Table 4-2 Required Sample Size Record

	Smishing	**Normal**	**Total**
Subscriber	52,400	153,013	205413
Record	829,450	2,998,291	3,080,741

4.4. **Data Preprocessing**

Data preparation is the process of converting raw data into a suitable format that can be interpreted. The raw data frequently lacks key trends, is inconsistent, noisy, and full of inaccuracies. Before using machine learning algorithms, the quality of the data should be verified. To improve classification performance and properly interpret the outcome, well-executed preprocessing processes are essential. Clearing null values, deleting noisy data, combining tables, aggregating attributes, and integrating tables are examples of preprocessing tasks.

Under this study, there are a few data preprocessing techniques that help to improve data quality. These methods include data cleaning, which eliminates noise and corrects inconsistencies in data, and data integration, which combines data from several sources into a single coherent set of data. Data aggregation, feature deletion, or grouping are all methods for reducing the amount of instance data. Data transformation is the final strategy, which is also used to normalize data that has been scaled to fit inside a narrower range. They are collaborating to attain better results.

4.4.1. **Data Cleaning**

Data cleaning is an important early step for fixing values, smishing, and possible inconsistent data. The data cleaning process is time taking and requires high attention not to avoid the creation of irrelevant data at the end. Since the collected data are stored in different tables or places, all the data cleaning activities are applied to all Make sources. Make sure the same columns found in each table must have equal size, data type, and the same format. Like SMS communication, called Numbers, calling Numbers, Call Start Time, and Call End Time are found in more than two places or tables [13].

The research cleansed the data by eliminating the stored data from each column that had missing values or partial data. The CDR data contain duplicate records, null values, and missing values. The goal of this study is based on mobile network monitoring systems: only those SMSs, Called Numbers, Calling Numbers, Service Numbers, Service kinds, Recharge Amounts, and Recharge Channels are chosen for the study. However, records that differ from such values are eliminated from the CDR data that has been collected. The gathered CDR includes 33 columns in total, plus null-valued and unnecessary attributes were removed. Records that were incomplete, outliers, or missing values were removed [14]. The total recorded has many attributes from this attribute some of them are not important for this study. The entire column is then erased because using those undesirable attributes would be pointless.

Accordingly, CALL FLOW, CDR EVENT TYPE, TOTAL FEE (MAIN ACCOUNT), TOTAL FEE (BONUS MONETARY ACCOUNT), TOTAL FEE, FREE ITEM, FREE SECOND, FREE VOLUME, REFUND INDICATOR, LOCATION ID, TOTAL TAX AMOUNT, UN-SUBSCRIPTION KEYWORD, OFFER NAME, PARTNER NAME, PAYMENT ACCOUNT CODE, TOTAL TAX AMOUNT, PAY BY SELF, ROAM TYPE, BILL CYCLE, REALUSINGIMSI, REALUSINGMSISDN, RECHARGE TYPE, ORIGINAL BALANCE, CURRENCY was not important for this research and discarded. Due to the rarity of such entries and the fact that their removal has no impact on the dataset as a whole, it was decided to exclude them.

Additionally, an outlier is a variable with an unusual value that could otherwise negatively affect the specification of the model, lead to biased parameter estimates, and produce inaccurate findings. The variable that appears to be skewing the data's distribution may be at its maximum or minimum [4] [14]. Extreme values that deviate from the majority of other data points in a dataset are called outliers. For reliable findings, it's crucial to carefully spot potential outliers in the dataset and deal with them in the right way. The presence of many

data points that are dispersed far from the mean in one direction is another sign of a variable's skewed distribution, which is another indicator of outliers. In situations where data have a skewed distribution or a high number of outliers, it is crucial to choose the right statistical tests or measures. Interquartile Range (IQR) is the most popular data distribution measure for spotting outliers [14] [50].

Finding outliers in data is commonly acknowledged as using the interquartile range The entire dataset is divided into four equal halves, or quartiles, when the interquartile range, or IQR, is used. The IQR is computed based on the gaps between the quartiles. IQR, or interquartile range a statistical gauge of data dispersion is the interquartile range (IQR). It is equivalent to the distance between the third (Q3) and first quartiles (Q1), or the 80th and 20th percentiles. As a result, the middle 50% of the data, sometimes referred to as the middle 50% or mid spread, makes up the IQR. Q1 from Q3 can be subtracted to create the IQR.

$$IQR = Q3–Q1 \quad (4.1)$$

Equations 4.2 and 4.3 indicate the upper and lower bounds, which were derived using an outlier's factors that are essentially set to "1.5".

$$Upper\_Limit = Q1 + (1.5 \_ IQR) \quad (4.2)$$

$$Lower\_Limit = Q1 - (1.5 \_ IQR) \quad (4.3)$$

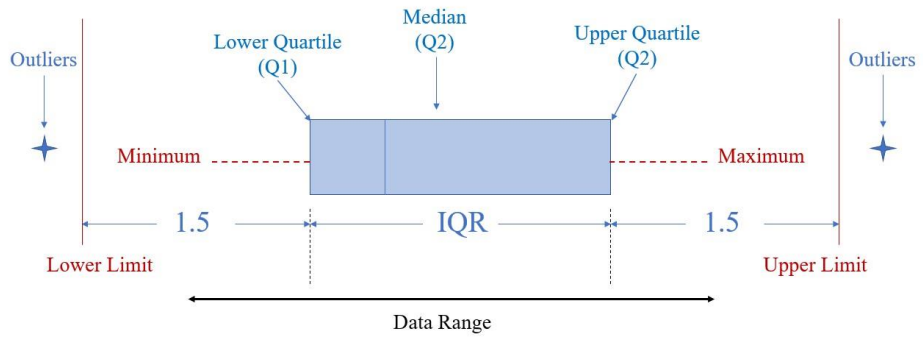


Figure 4-8 Pictorial representation of IQR [39]

The outcome of the earlier computation shows when a number of the other values deviate "too far" from the fundamental value. The term "outliers" is used to describe these "too far away" points since they "lie outside" the range.

In this study, outliers were eliminated using the IQR methodology, and a total of 12625 data were eliminated.

4.4.2. **Data Integration**

Data integration is an integral part of data operations in which data can be obtained from several sources. It is an important technique because it provides a uniform view of dispersed data while ensuring data accuracy and integrates data from various sources to make it available to consumers in a single uniform view that reflects their state. In this research, there are seven tables are created to manipulate the recorded CDR data. These are voice, data, recharge, and SMS for both fraudulent and legitimate users. To create labels for each month, the attribute tables from individual CDR files must be combined and saved into that various tables. It took a long time for the research to integrate the data. The amount of data was quite difficult to handle. A distinct identification is needed for each aggregation time range in order to locate and gather the aggregated information from each table. Combining the data from all of the aforementioned tables in a single aggregation hour. The values of each service number record from those tables must correspond to the time, date, and column values as stated in the appendix.

4.4.2.1. **Data Aggregation**

The process of gathering information and expressing it in a summarized form is known as data aggregation. The atomic data rows that are normally gathered from numerous sources when data is aggregated are frequently replaced by amounts or summary statistics. Groups of observed aggregates are replaced by summary statistics based on those observations. It is a type of information and data mining process where information is sought out, gathered, and presented in a report-based, condensed style to accomplish particular corporate goals or procedures. In this study, aggregation is going to apply depending on the Service Quality Management (SQM) of SMS transaction, Service Number, SMS in, SMS out, time dimension, frequency, Call out, Call in recharge amount, and operator name to detect smishing. This accumulated CDR data record needs to be aggregated together in order to give a full picture of user behaviors. An aggregation is the cumulative result of each individual user within a given time span. Table 4.3. displays the daily agglomerated data across time depending on the chosen CDR data attributes.

Individual data pieces with personally identifiable information produced by aggregation can be integrated and replaced with a summary that represents the group as a whole. SMS, voice calls, recharges, and Internet data are all combined at the subscriber level by aggregation to create a single instance. In addition, as shown in the appendix, a class field that indicates the subscriber type is introduced for training purposes.

Table 4-3 Aggregated and derived feature description.

Attribute	**Description**	
TOT_CALLS	*Total number of calls Subscribers made*	
RECEIVED_SMS	*Total number of SMS received*	
SENT_ SMS	*Number of SMS sent*	
INC_CALLS	*Number of incoming calls*	
DIST_CALL	*Number of Unique called numbers*	
DIST_SMS	*Number of Unique SMS sent*	
AVD_MA_BAL150 *LAST_RECH_DATA_MA*	*Average main balance account in 150days* *Last recharge date main account*	
FR_MA_RECH150	*Frequency of main account Recharged in last 150*	
CNT_BAL150	*Number of Balance transferred by user in last 150 and frequency*	
AMNT_BAL150	*Total Number of Balance transferred by user in last 150 and its frequency*	
CNT_MA_RECH_CHLLS150	Main account recharge channels and its frequency	
SMISHING_STATUS	Identify the subscriber type	

After the data has been extracted, the following step is to prepare it in a file format that machine learning algorithms can use. Additionally, the points below attempt to reflect consumer behavior.

- The number of voice calls made by the user in a specific amount of time
- The number of SMS sent and received during the specified period
- The volume of transactions made during a given period of time
- How long did the consumer use the service?
- How frequently uses all the services

4.4.3. **Validation techniques**

Validation is the process of gathering and analyzing data from the product design stage and throughout the process to demonstrate through scientific proof that the process can generate high-quality products consistently [28].

The statistical significance is a crucial factor to take into account when determining the relevance of a data analysis technique. To reach a broad conclusion, the analysis should be used on a number of substantial data sets. To avoid overfitting problems, the data set should be realistically representative of the data. The outcome should also be precise and consistent. Results that are ambiguous and deceptive cannot be the basis for decisions. Among many validation techniques, cross-validation is the one used to compute and evaluate unseen data.

After data aggregating and integration are completed, this section performs the training and builds a classification model using selected models. In this study, to train a model the researchers used one training techniques. These models are K-fold cross-validation.

Cross Validation

The cross-validation option is used to resample and evaluate machine learning models on a limited data sample [14] [37]. Create k-partitions from the dataset. Train a model with all the components except for the one that is kept out as the test set, then carry out this process k times to make k distinct models and give each fold a chance to survive as the test set. This process is repeated iteratively by changing the test fold starting from the first to the Kth fold. Finally, the cumulative average error of each training and testing result is provided [1] [14] [37]. A single instance has been used in 10-fold cross-validation for both training and testing. This study determines the performance measures across all iterations. The averages of the performance measurements over all iterations make up the overall performance measures. The estimation bias of performance metrics is decreased by this validation technique. By dividing the total number of correctly categorized samples across all 10 rounds by the total number of samples in the initial dataset, the performance estimate is derived. As shown in Table 4.4. In a similar way, this study uses the technique. The technique has eight steps: one, Choose a few folds-k. Two: Divide the dataset into k identical folds. Three: select k-1 folds as the training set and the rest as test folds. Four: train a model on the training set. Five: Validate the test set, then record the outcome. Six: Repeat steps 3-6 k times then return to step 6.

Table 4-4 Process of 10-fold cross-validation

NO	Train Classifier	Test Classifier
6-1	1,2,……9,10	1
2	1,2…….9,10	2
3	1,2,……9,10	3
4	1,2……..9,10	4
5	1,2……9,10	5
6	1,2……9,10	6
7	1,2……9,10	7
8	1,2……9,10	8
9	1,2……9,10	9
10	1,2…….9,10	10

Separate test data

Evaluation of data mining methods involves dividing data into training and testing sets. The majority of the data is typically used for training, and a smaller piece is used for testing when dividing a data set into a training set and a testing set. The former is used to construct the classifier, and the latter to assess its effectiveness. Data splitting is typically carried out to avoid overfitting.

The portion of the data used to train the model is known as the training set. The model should observe the training set and take notes, then adjust any of its parameters as necessary.

The testing set is an additional collection of data that is compared to the prior data sets and reviewed in the final model. The testing set is used to evaluate the final mode and algorithm. In the test set provided: On the basis of how accurately it assigns a class to a group of cases loaded from a file, the classifier is judged. A total of 20,5413 data points are divided into two datasets for training and testing in this experimental approach. As illustrated in figure 4.9, the training dataset has 143,789 instances, while the testing dataset contains 61,624 instances.

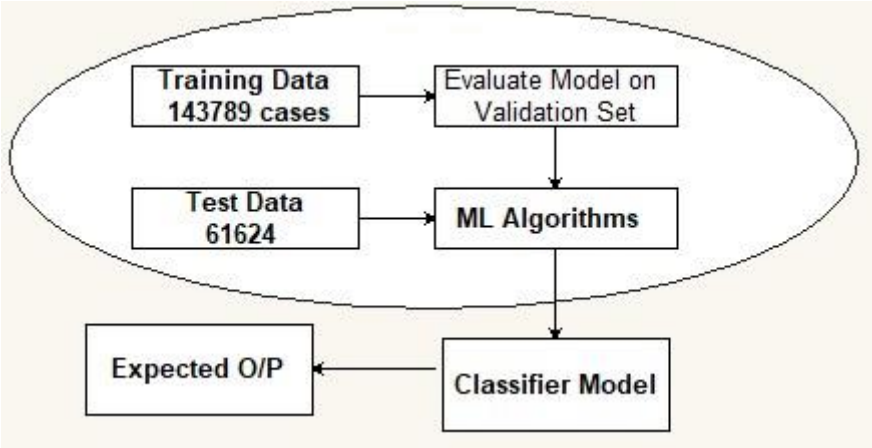


Figure 4-9 .Data Train and test Method

In this study, once all the data preprocessing, feature selection, aggregation, and integration is completed, all experiment are evaluated using cross-validation. Before crossing to the experiment process, labeled attributes are inserted into the dataset then algorithms experimented with the given parameter.

4.5. **Constructing Model**

In data mining, modeling is a key effort and developing models requires validated data. Once the data is prepared to be used for model construction, the best-performing model will be selected after computing performance metrics and model construction. Once the model is used in production, model upkeep is essential. This step might use various methods to complete the data mining assignment. Choosing a modeling technique, setting up an experiment, creating a model, and assessing the model are a few of the tasks.

Techniques for supervised categorization are learned in this subject. Appropriate categorization algorithms are chosen for model creation based on the research's objective. The chosen algorithms for this study's environment are RF, LR, DT, NB, SVM, and KNN. This work aimed to construct models utilizing chosen methods for the classification of Smishing SMS. A confusion matrix, performance metrics such as Precision, Recall F-measure, and accuracy, and a comparison of the models' performances were used to assess the models. Once trained and tested the algorithms, and evaluate the outcomes of the algorithm based on their performance measures parameters. The next subsections explain parameters for comparing and analyzing the algorithms.

4.5.1. **Confusion Matrix**

The classification models' performance with respect to a certain set of test data is evaluated using a matrix called the confusion matrix. It can only be decided once the actual test data values are known. There are two classes of classifiers in this situation. The matrix is a 2*2 table with two different axes for the projected values, actual values, and the total number of forecasts. Table 4.5 below displays the confusion matrix for the two class designations.

Table 4-5 Confusion Matrix

	Prediction class	
	Reality: No	Reality: Yes
Predicted: No	True Negative TP	False Positive FP
Predicted: Yes	False Negative FN	True Positive TP

The following cases appear in the table above:

True Negative: TN - Model predicted No, and the real or actual value was also No, proving that the projected class value and the actual class value are both valid.

True Positive: TP - The model predicted yes, and the fact that the actual number was also accurate suggests that both the predicted and actual values of the class are false.

False Negative: FN - A Type II error is when the model predicted no but the actual value was yes.

False Positive: FP - Despite the model's projection of "Yes," the actual outcome was "No." It's also known as a type-I error**.**

We can calculate the model's accuracy, precision, recall, and F-measure after we have a firm understanding of these four parameters, utilizing the matrix. Here are these calculations:

4.5.2. **Accuracy**

It is a key element in figuring out how effectively classification problems are resolved. It shows how frequently the model predicts the outcome properly. The ratio of the number of accurate predictions made by the classifier to the total number of predictions made by the classifiers can be used to compute it. The equation is shown below:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{TP+FP+FN+TN} \quad (4.2)$$

4.5.3. **Precision**

The indicator can be viewed as the number of accurate model outputs or the percentage of correctly anticipated positive classes that actually occurred. You could determine it by applying the following formula:

$$\text{Precision} = \frac{TP}{TP+FP} \quad (4.3)$$

4.5.4. **Recall:**

It is sometimes referred to as the percentage of all affirmative classes that our model accurately predicted. There must be the greatest recall possible. Even if a sample is genuinely positive, a classifier with a recall of 1.0 but lower accuracy would output positive for every sample. A classifier's recall increases with decreased false negatives.

$$\text{Recall} = \frac{TP}{TP+FN} \quad (4.4)$$

4.5.5. **F-measure:**

Comparing two models that have low precision but high recall, or vice versa, is difficult. F-score is thus applicable in this context. We may evaluate recall and precision simultaneously using this score. The F-score is maximum when recall and precision are equal. The following formula can be used to calculate:

$$\text{F-measure} = 2X \frac{Precision\ X\ Recal}{Precision+Recal} \quad (4.5)$$

Other important terms used in the Confusion Matrix:

Null Error rate: This statistic illustrates how frequently our model if it were to be consistently reliable, would forecast the majority class accurately. According to the accuracy paradox, the best classifier has a higher mistake rate than the null error rate.

ROC Curve: A graph known as the ROC displays how well a classifier performs for each possible threshold. On the graph, the true positive rate (Y-axis) and the false positive rate (X-axis) are plotted.

4.6. **Exploratory data analytics**

4.6.1. **Getting to know the data set**

Simple graphs, charts, and tables can reveal significant relationships that could point to fertile areas for additional research. We explore the Smishing dataset using exploratory techniques. For the Exploratory Data Analysis (EDA), we employ Python programming language and the 19th Oracle generation database. The data set contains 52400 unique samples and 157200 normal data, and with an indication status for Smishing.

Table 4-6 Subscriber information (call, demographic, and, payment data)

Feature	Unique Subscriber ID	Categorical
Service type	Total No of SMS, Voice, and data	integer
GenTotCalls	Totals number Calls	integer
SMSCalls	Total Number of SMS	integer
MSISDN	Mobile Number of the user	integer
Nan	Numbers of ages on cellular networks in days	integer
Daily_spent in a month	Total daily spent from the main act, avg over the last 1mount	integer
Daily_spent in 5 months	Total daily spent from the main act, avg over the last 5 months	integer
Label	Flag Indicated balance label	integer
Avg_main_bal_1m	Average main account balance over last 1 month	integer
Avg_main_bal_5m	Average main account balance over the last five months	integer
Last_rech_date_ma	Number of days till the last recharge of the main account	integer
Last_rech_amt_ma	Amount of last recharge account	integer
Time_ma_rech_1mount	Number of times the main account got recharged in the last 30 days	integer

Fr_ma_rech_1mount	Frequency of the main account recharged in the last 30 days	integer
Tot_ma_rech_1month	The total amount of recharge in the main account over the last 30 days	integer
Med_Amt_ma_rech_1month	Median of amount recharges done in main account over the last 30 days at user label	integer
Med_Amt_ma_rechprebal_1month	Median of main account balance just before recharge in last 30 days at user label	integer
Time_ma_rech_5mount	Number of times the main account got recharged in the last 5 months	integer
Fr_ma_rech_5mount	Frequency of the main account recharged in the last 5 months	integer
Tot_ma_rech_5month	The total amount of recharge in the main account over the last 5 months	integer
Med_Amt_ma_rech_5month	Median of amount recharges done in main account over the last 5 months at user label	integer
Med_Amt_ma_rechprebal_5month	Median of main account balance just before	integer

	recharge in last 5 months at user label	
Cnt_bal_rec_1mounth	Tot Number of balance received by the user in last 30 days	integer
Tol_bal_rec_1month	Total amount of balance received by the user in last 30 days	integer
Maxmnt_bal_rec_1month	Maximum amount of balance received by the user in last 30 days	integer
Med_bal_1month	Median of amount of balance received by the user in last 5 months	integer
Cnt_bal_5mounth	Number of balance received by the user in last 30 days	integer
Tol_bal_5month	Total amount of balance received by the user in last 5 months	integer
Maxmnt_bal_5month	Maximum amount of balance received by the user in last 5 months	integer
Med_bal_5month	Median of amount of balance received by the user in last 5 months	integer
Avd_tra_bal_1month	Average transfer time in last 30 days	integer
Avg_tra_bal_5month	Average transfer time in last 450 days	integer
Bal_rech_chls_1m	Balance recharge from different channels in last 30 days	integer

Tol_bal_rech_chls_5m	Balance recharge from different channels in last 5 months	

4.6.1. Exploring Variables

Investigating the variables, examining the links between numerical variables, analyzing categorical variable distributions, and examining the connections between sets of variables are some of the main goals of exploratory data analysis.

a. Subscribers with main balance levels vs transferring balance within a week

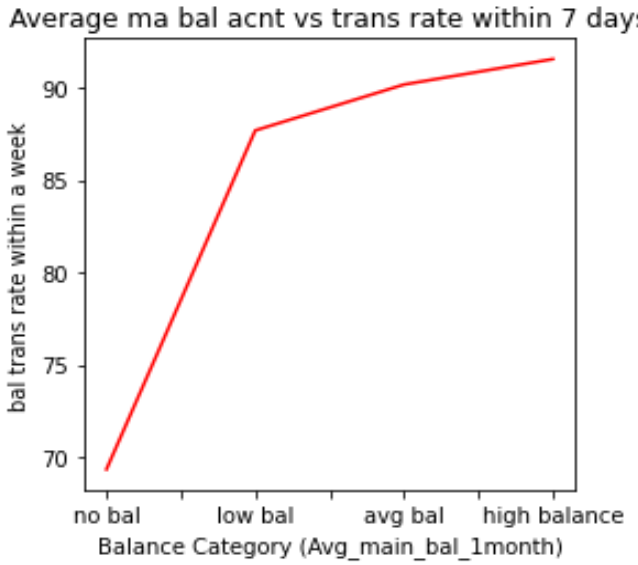


Figure 4-10 Subscribers with main balance levels vs transferring balance in 7days

The above bar plot indicates how different main balance levels of clients have shifted balances within five days. The high balance level individuals have a 100% chance of balance transfer, receive but no recharge within 5 days. Regarding those with average and low balances, 11% of individuals do not transfer within 5 days. Regarding users with low balances, it has been found that 40% of them do not transfer their balance within the allotted five days. People with a totally transferred balance or 100% balance transferred people are creating major losses to the subscribers by smishing them within five days.

b. Customers with different frequency levels main account recharge vs transfer rate

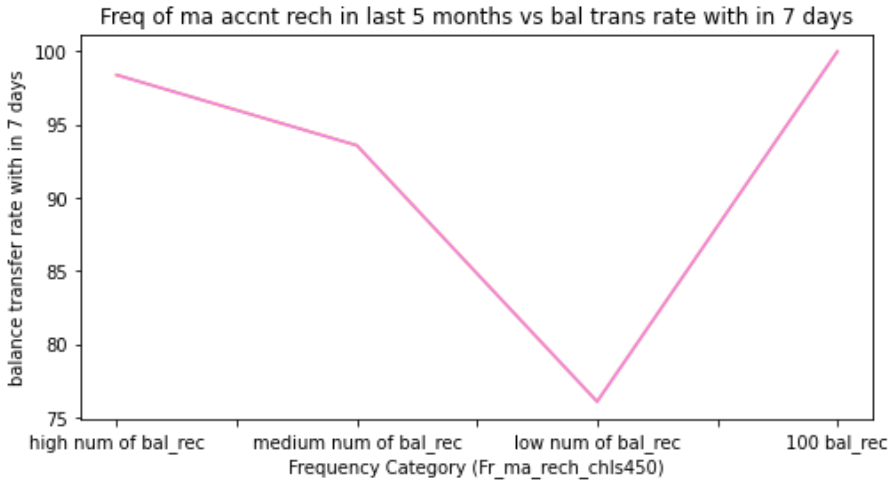


Figure 4-11 Customers with different frequency levels recharge vs transfer rate

The above bar plot shows us how customers with different frequency levels (main account recharge) are transferred their balance within five days. Coming to the average and low & medium frequency people it is observed that around 6% of people are not transferred within 5 days. Regarding low frequency users, it is noted that 30% of people have their balances transferred within the allotted five days. The 30% of people who don't have their main account recharged for 30 days are costing the consumer a lot of money.

c. Customers with different received balance level vs balance transfer rate within 7 days

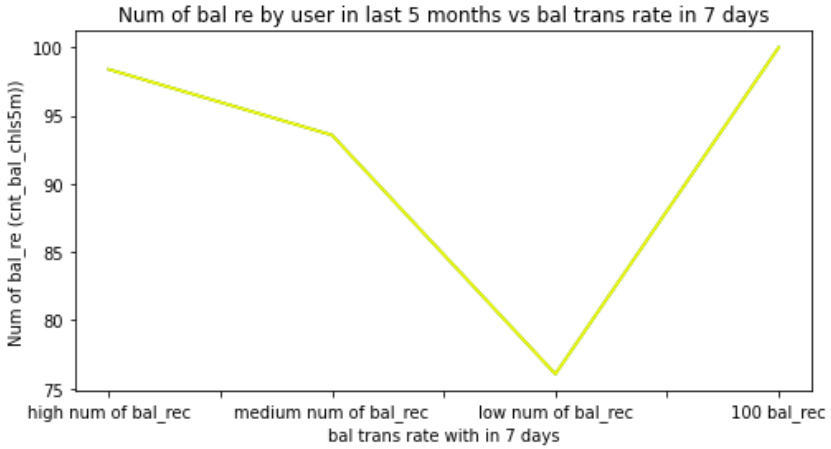


Figure 4-12 Customers with different received balance level vs balance transfer rate within 7 days

The above bar plot illustrates how different balance levels of clients received and transferred money within a week.

People who received a balance level 100% rate within 7 days can be found when the remaining levels are taken into account. Regarding the significant number of customers who have received balances, 22% do not obtain them within 7 days. Only 3% of those in the category with few balances do not get transferred within 5 days. This is followed by individuals with a total transfer of balanced and 100% of received balance is mostly smishers.

d. Total amount of recharged by the user in last one month vs transfer rate within seven days

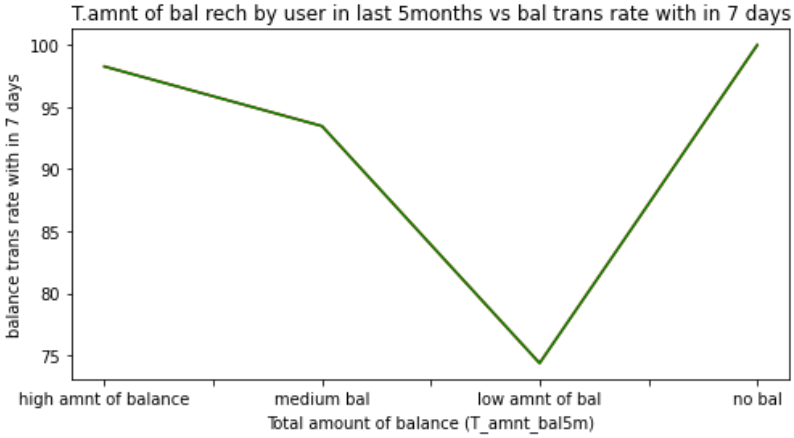


Figure 4-13 total amount of recharge by the user in last 30 vs transfer rate within seven days

The graph displays how customers with different balance levels recharged and transfer their balance within seven days do not considering customers who are not recharged.

By leaving clients there who can't recharge their account within seven days while taking into account the remaining levels. Regarding those with low balances, it is noted that approximately 27% of customers have their balances transferred within 7 days. Only 3% of customers who recharge their accounts with a significant amount do not transfer the balance within 7 days. The next group, those with a medium amount of debt, has roughly 10% defaulters.

e. Customers with different incoming call frequencies vs called rate

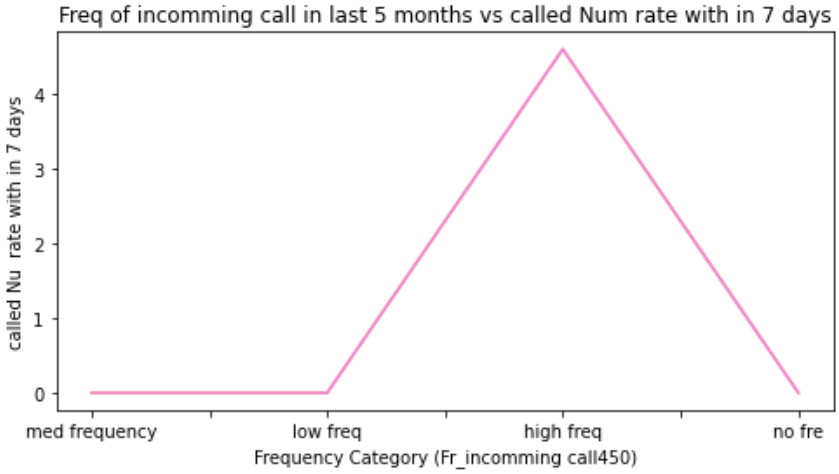


Figure 4-14 customers with difference incoming call frequency vs called rate

The above bar plot indicates how incoming calls levels of clients have shifted balances within five days. The high balance level individuals have a 100% chance of balance transfer, receive but no recharge within 5 days. Regarding those with average and low balances, 11% of individuals do not transfer within 5 days. Regarding users with low balances, it has been found that 40% of them do not transfer their balance within the allotted five days. Peoples with totally transferred balance or 100% balance transferred people are creating a major loses to the subscribers by smishing them within five days' time.

f. **Distribution of SMS attribute with status overlay**

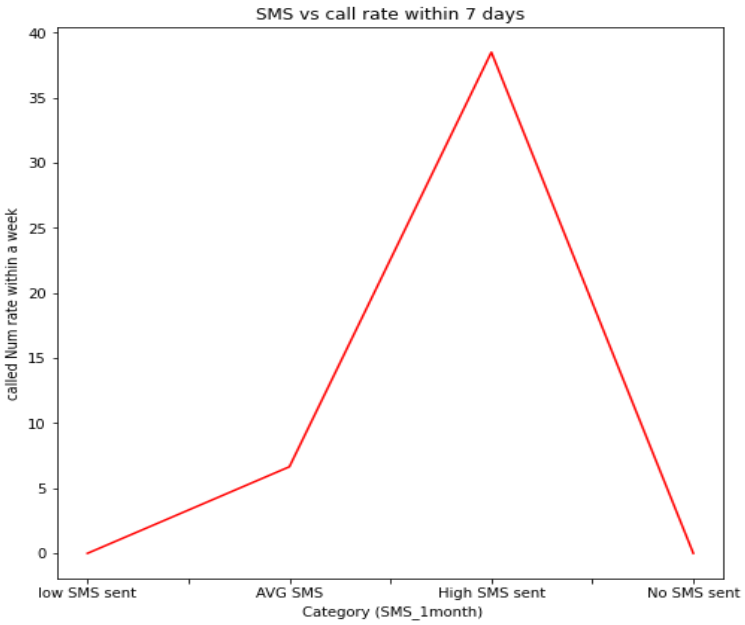


Figure 4-15 Distri**bution of SMS attribute with status overlay**

As a result of Exploratory Data Analysis, we have found that some variables like SMS, incoming_call, outgoing_call, bala_rec, bal_transfer, recharge amount, frequency_lebel, and

average main balance with status overlay show important tendency for fraudulent use. Some of the other variables exhibit a very modest tendency, while others exhibit none at all.

4.6.2. Dealing with Correlated Variables

Analytics should be careful to avoid providing correlated variables to one's data mining and statistical models in order to check the correlation with the dependent variable labels. At best, using linked variables will highlight one aspect of the data too much.

The CDR behavior data set contains the following variables: SMSs in/out, *calls (outgoing/incoming)*, daily amount spent, average main account, last recharge of main account, number of times main account got recharged, amount of last recharge main account, frequency of main account recharge, total amount of recharge in main account, median, average transfer amount, average receive amount, number balance transfer by the user, Total balance transfer and age on the network in the last five month. The data description indicates that the recharge variable may be a function of receive, transfer, calls and SMS, with the result that the variables would be correlated.

There does seem to be relationship between recharge, calling and called number. One may have expected that the number of calling number and the called number would tends to increase and similarly for SMS, recharge main account receive high amount, transfer would tend to show something relate, resulting in appositive correlation these fields. The figure show the correlation these fields.

1.000000	-0.004104	0.168228	0.166104	0.057969	0.075417	0.003349	0.002122
-0.004104	1.000000	0.000298	-0.000339	-0.001052	-0.000662	0.001939	-0.001575
0.168228	0.000298	1.000000	0.977740	0.441897	0.458868	0.000949	-0.001827
0.166104	-0.000339	0.977740	1.000000	0.434595	0.471689	0.001415	-0.001986
0.057969	-0.001052	0.441897	0.434595	1.000000	0.955316	-0.000675	0.003660
0.075417	-0.000662	0.458868	0.471689	0.955316	1.000000	-0.001200	0.003241
0.003349	0.001939	0.000949	0.001415	-0.000675	-0.001200	1.000000	0.001921
0.002122	-0.001575	-0.001827	-0.001986	0.003660	0.003241	0.001921	1.000000
0.131756	0.003455	0.276428	0.264974	0.127854	0.121617	-0.000122	-0.000488
0.237323	-0.002754	0.450983	0.426147	0.233414	0.230385	0.004406	0.001532
0.001017	-0.001001	-0.000245	-0.000092	-0.000905	-0.000293	-0.001549	0.000394
0.202651	0.000410	0.636603	0.603960	0.273041	0.260168	0.002433	-0.000600
0.141475	0.003507	0.295722	0.283559	0.130698	0.121047	-0.001158	0.000612
-0.004879	0.003358	-0.001162	-0.000738	-0.001481	-0.001244	0.004237	0.002944
0.236439	-0.002555	0.586789	0.592251	0.312265	0.345485	0.004420	0.001151
0.084281	0.004015	-0.078284	-0.079527	-0.033470	-0.036494	-0.000021	0.000623
0.205663	0.000637	0.762908	0.768658	0.342695	0.360995	0.002591	-0.000945
0.120770	0.004190	0.258427	0.251289	0.111399	0.103644	-0.000700	-0.000355

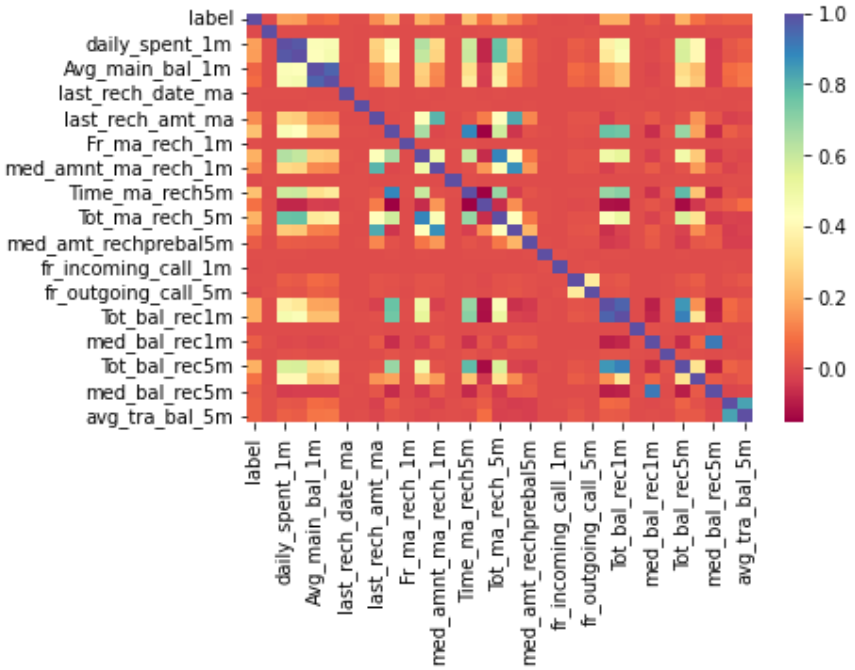


Figure 4-16 Heat map and Pearson matrix for correlation of features

A statistical metric called a correlation expresses how strongly two variables are related to one another. The two primary types of correlations are positive and negative correlations. A positive correlation exists when two variables move in the same direction; when one rises, the other rises as well.

The 33 features are formed into the heat map matrix, also known as the correlation matrix, as seen in figure 4.3 below. The correlation's magnitude ranges from -1 to 1. A score nearer to one or a negative value near zero suggests a positive association between features, whereas a value less than zero or negative suggests an independent relationship between features. The diagonal values are related to value 1, and vice versa. It is hence an exact correlation. The 33 features are displayed in a heat map matrix with their correlation values in Figure 4.10. Both positive and negative correlations provide an understanding of the data's structure. In contrast to red, which shows a negative link between two questions, blue denotes a positive relationship. The relationship's strength is depicted by the size and intensity of the colored dots.

4.6.3. **Data normalization**

The process of converting the distribution's shape to a normal distribution is known as normalization.

Table 4-7 normalization

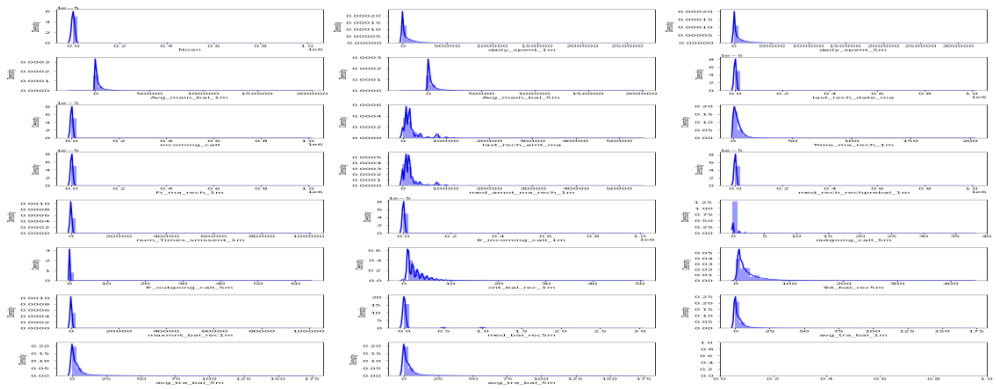


Figure 4-17 Normalization

4.6.4. **Outlier Transformation**

Outliers are data points in a distribution that deviate from the norm. The dataset underwent a number of modifications in order to be ready for analysis. There are no null values in the data set, hence null value imputation is not necessary for the data set. The data set contains outliers for a number of variables. The study discovered a method to perform an outlier's imputation technique for the data of the features utilizing IQR with Q3-Q1 by observing these characteristics. IQR is defined as the middle 50% of data, or Q3–Q1. A quartile divides an ordered dataset into four equal-sized groups, with Q1 being the first quartile and Q3 the third. To find Q1 and Q3 in Python, we can utilize the percentile function in the NumPy module. Outliers are values greater than Q3 + 1.5 * IQR or smaller than Q1 - 1.5 * IQR, according to the interquartile range technique. First, we determine Q1 and Q3 using the percentile function. The data is the first parameter, and the second is the percentiles to be calculated.

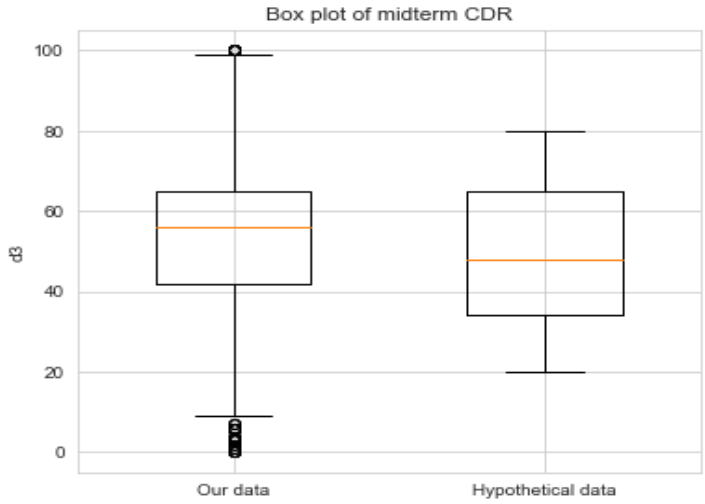


Figure 4-18 Outlier transformation

The above box plot infers as in the first box there are dots which are the maximum and minimum values from other data points. So, these values are outliers that need to be removed which leads to the biased performance of the prediction, and based on that we removed these values using the IQR method

4.6.5. **Variance inflation factor (VIF)**

How inflated the variance is effectively measured by the variance inflation factor (VIF). In this context, the standard error is referred to as variance. Thus, when multicollinearity in correlation exists, the variances of the computed coefficients are inflated. It is a statistic used to assess the degree of collinearity between several variables. The formula to determine the Variance Inflation Factor (VIF) of any property is as follows:

$$\text{VIF}(x) = \frac{1}{1-R1^2} \qquad (4.1)$$

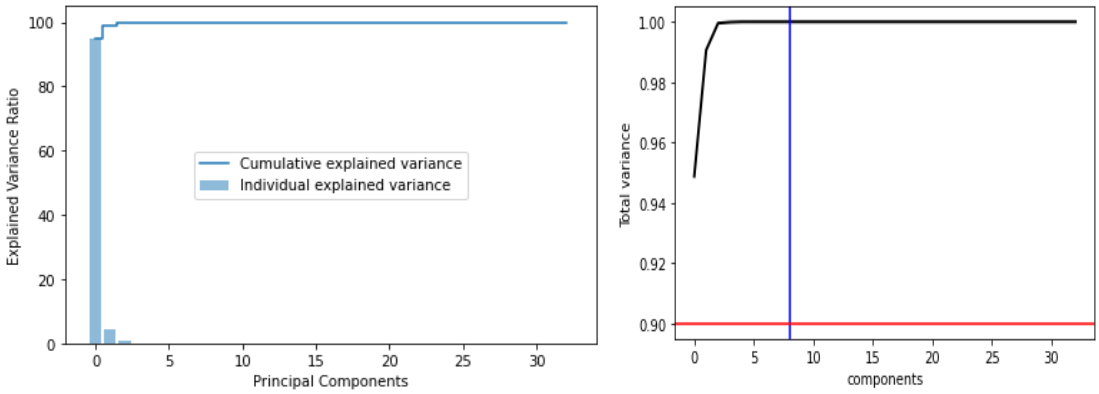
Table 4-8 VIF values

No	Feature	VIF values	Feature	VIF
1	label	8.302055e+00	fr_incoming_call_1m	7.915284e+00
2	Age	2.950934e+03	outgoing_call_5m	10.547497e+01
3	daily_spent_1m	2.259525e+01	fr_outgoing_call_5m	5.192125e+00
4	daily_spent_5m	3.272503e+01	cnt_bal_rec_1m	6.648623e+00
5	Avg_main_bal_1m	7.185817e+00	Tot_bal_rec1m	1.432314e+02
6	Avg_main_bal_5m	7.737966e+00	maxmnt_bal_rec1m	1.492621e+01
7	last_rech_date_ma	1.723406e+06	maxmnt_bal_rec5m	4.870312e+00
8	incoming_call	5.416527e+06	med_bal_rec5m	4.282132e+00
9	Fr_ma_rech_1m	5.268532e+06	avg_tra_bal_1m	6.915284e+00
10	Tot_ma_rech_1m	1.030806e+01	fr_incoming_call_1m	9.547497e+01
11	med_rech_rechprebal_1m	4.346014e+02	outgoing_call_5m	4.192125e+00

12	Time_ma_rech5m	7.717201e+00	fr_outgoing_call_5m	6.648623e+00
13	Tot_ma_rech_5m	1.202088e+01	cnt_bal_rec_1m	1.432314e+02
14	med_amt_ma_rech_5m	4.334420e+00	Tot_bal_rec1m	1.492621e+01
15	num_Times_smssent_1m	2.301126e+02	maxmnt_bal_rec1m	5.870312e+00
16	fr_incoming_call_1m	1.192661e+07	maxmnt_bal_rec5m	5.282132e+00

According to the aforementioned findings, the data set has numerous features that exhibit significant multicollinearity. This suggests that Principal Component Analysis is necessary. The model prediction and model outcomes will be impacted by noise or correlation between the independent variables if Principal Component Analysis is not performed. In order to lessen the multicollinearity effect among the independent variables, PCA is required because more than half of the features have a vif >4 value.

Figure 4-19 principal component analysis



From the above result we can observe that, 33 components (or 91% of the data) are covered by primary component analyses. So the total number of principal component analytic was taken as 33.

CHAPTER FIVE

5. RESULT AND DISCUSSION

The experiment's findings will be discussed in this chapter. The findings will be displayed using various plot types, bar charts, tables, and some text to describe the various presentations. This chapter also describes the various experiments' methods of operation.

The comparison of the experiments' performance analysis of six machine learning algorithms mostly takes place in the discussion chapter, smishing SMS is presented. The outcomes of the various experiments will be examined and fully explained. Key figures and findings will be mentioned along with the graphs, charts, and other visual presentations. This chapter's goal is to present the results of the tests conducted. It will provide some cues that will be expanded upon and discussed later. It will be utilized to talk about which algorithm performed the best overall in the experiments.

There is a total of six experiments as discussed earlier in chapter 3. The six experiments will be performed and the results of each and every experiment will be analyzed. For each algorithm, the experiments were performed with the validation cross-validation. The best model for smishing SMS detection was offered after the results of validation procedures on each algorithm were documented with regard to their processing times and performance measurement factors. The many parameters are examined during the tests, and the one that performs the best is used as the parameter in the model.

This model is then again a parameter in the prediction.

5.1. Results and comparison

The primary objective of this work is to choose a better classification method for creating a model that works the best for handling prediction and spotting smishing SMS. The research chooses six categorization models, and table 5.1 below lists each algorithm's greatest performance accuracy.

The research's goal is to compare the top classifier models and choose the one with the best classification accuracy as explained below. On the basis of thorough experimentation, various models have been run. Thus, the most effective classification algorithm suitable for this problem area has been chosen. The researcher chose smishing features and modified several parameter settings in consultation with the domain specialists. The experts explained that

more focus should be placed on smishing users reported by customers who were attacked, the high total number of outgoing calls, the high total number of sent SMSs in relation to an outgoing call, the high balance recharges from different channels, as well as the high balance transfers in relation to outgoing calls and SMSs and no incoming calls after the start of fraud. Sending threatening texts and making many phone calls throughout the day are two of the main habits of smishing fraudsters. Another sign of their activity is when a subscriber or customer becomes inactive or leaves the network. Following is a detailed description of the outcomes of the overall classification, mobile telecommunication smishing fraudsters, and non-fraudsters for both classification models. The experiment was run with the detection techniques that were chosen. In the experiment using supervised ML algorithms, data aggregation and 10 cross-fold validation approaches are used to obtain the final instance data. Re-call, precision, accuracy, F-measure, and ROC curves are examples of performance measurements that are used for comparisons. The classifiers algorithm classification performance results and the validation methodologies are shown in Table 5.1. When the models were compared, the RF ML algorithm's cross-validation test possibilities using Python showed the highest accuracy with 90.11% percentage values. The accuracy of 87.6% from the total experimental score utilizing cross-validation is a very low result in another classifier algorithm's SVM model performance when compared to all other classifier algorithm models. According on cross-validation 89%, KNN is comparatively the second-highest classifier model.

Table 5-1 performance metrics of all the classifiers

Validation Technique	Algorithms	Build Times in s	Recall	Precision	F-measure	ROC	Accuracy	support
10-Fold/ Python	LR	660	0.99	0.87	0.94	0.822	0.88	61624
	RF	1740	0.97	0.92	0.95	0.84	0.90	61624
	K-NN	4597.6	0.96	0.92	0.94	0.80	0.89	61624
	DT	1980	0.97	0.92	0.94	0.82	0.89	61624
	NB	1860	0.78	0.95	0.85	0.80	0.80	61624
	SVM	36000>	0.99	0.88	0.93	0.79	0.876	61624

As was already mentioned, the main goal of data ML algorithms is to compare various models and choose the one with the best classification accuracy. Various models are run based on the experimentation that is chosen. The optimal classification algorithm for this problem area has been chosen in accordance with the findings.

In particular, all algorithms SVM, KNN, and others used a substantially varied amount of time in the cross-validation comparison of classification in terms of time (model building and evaluation). SVM takes longer to execute than the other two algorithms on the scripts, taking nearly an hour to complete. However, the results showed that RF had a second faster run time of 1740s with the cross-validation test while Logistic Regression took a much better time in both cases of building and evaluation time than the other algorithms.

Models for classification are used to predict the target class of the data sample in classification problems. The classification model calculates the probability that each event falls into a particular class or classes. To be able to rely on using classifications models in production to solve real-world problems, it is crucial to assess their performance. Performance measurements are employed to assess how well machine learning categorization models perform in a specific circumstance. Performance indicators like accuracy, precision, recall, and F1-score can be used to assess a model's strengths and weaknesses while making predictions in unexpected contexts. Confusion matrices, which represent counts from expected and real values, demonstrate how well an algorithm performs on a set of test data for which the true values are known.

In general, the study work titled Comparative Study of KNN, SVM, Naive Bayer (NB), RF, LR and Decision Tree Algorithm for CDR smishing detection prediction was used for the quantitative analysis. Following model testing using the testing data and as previously discussed in the performance metrics of all the classifiers in Table 5.1, the resulting confusion matrices for the LR, RF, KNN, DT, NB, and SVM algorithms have been given in Table 5-2, respectively. Out of these observations, it concluded that the RF and KNN algorithms render the best performance and should be the best model for student performance prediction to their accuracy. After the models were tested, it was discovered that the SVM model actually performed the best in terms of prediction. Additionally, it is safe to say that the RF and KNN method performs the best when compared to the other four algorithms after calculating other factors including building time, recall, accuracy, F-measure, ROC, and Accuracy.

POINTS OF SIGNIFICANCE In comparison to the other four algorithms, RF, KNN, and DT performed the best. Machine learning (ML) examples that highlight numerous key trade-

offs include RF, DT, and KNN. While KNN avoids making a priori assumptions about the shape of the class boundary as the amount of training data increases, it is a nonparametric method that can more closely adapt to nonlinear borders. KNN is more variable than other models. When looking for patterns with a large number of input variables, KNN rapidly degrades [22]. The final/leaf node of a decision tree will be averaged for regression in a random forest (RF), on the other hand. Fundamentally, RFs perform better when they are less overfitting. A logically based ML technique is DT [51]. Classification trees with top-down recursion give results similar to flowchart structures. Test and compare the property values on the internal node of the tree, find the corresponding branch, and then come to a decision in the leaf node of the DT till the cut-off value. Simple decision rules created from the properties of the data are taught to DT. In general, the accuracy of a forest of trees classifier depends on the strength of each individual tree in the forest and the association between them. Using noise-resistant random decision forests, overfitting of the DTs is prevented.

Information that is frequently utilized in publications that classify fraud using machine learning approaches is shown in Table 5.2. As was already mentioned, the confusion matrix serves as the foundation for this forecast. The prediction's outcome is simply numbered, and the percentages must be calculated manually. This is accomplished by adhering to a standard that outlines how to compute the various rates. The rue Positive rate, or TP, indicates the percentage of correctly classified smishings. False Negative Rate, or FN, refers to how many smishings were categorized as typical traffic. All of the predictions that were successfully identified as regular traffic have a TN, or True Negative rate. All of the regular traffic is labeled as smishing due to the FP rate, or false positive rate. This provides a clear image of how accurate the method is and may also provide other information, such as the potential for calculation when these values are available.

How many samples the algorithm correctly and erroneously identified is shown in Table 5.2 below. The columns are arranged so that the samples that were successfully categorized appear in the appropriate columns. The four algorithms SVM, LR, KNN, DT, and RF incorrectly classify 12215, 7479, 6761, 6384, and 6091 from their highest scoring using cross-validation performance results, respectively. In the column described below in the table, the NB cross-validation technique achieves relatively a smaller incorrectly classified 0.69% or 14196 from the total 205413 instances.

Table 5-2. Summary of confusion matrix

		Correctly classified		Incorrectly classified	
Validation technique	Algorithm	TP	TN	FP	FN
10 Fold	LR	53922	223	7364	115
	K-NN	51777	2988	4599	2162
	DT	52465	2778	1575	4809
	NB	42178	5286	2301	11859
	SVM	51136	5142	2356	9859
	RF	52638	2895	4692	1399

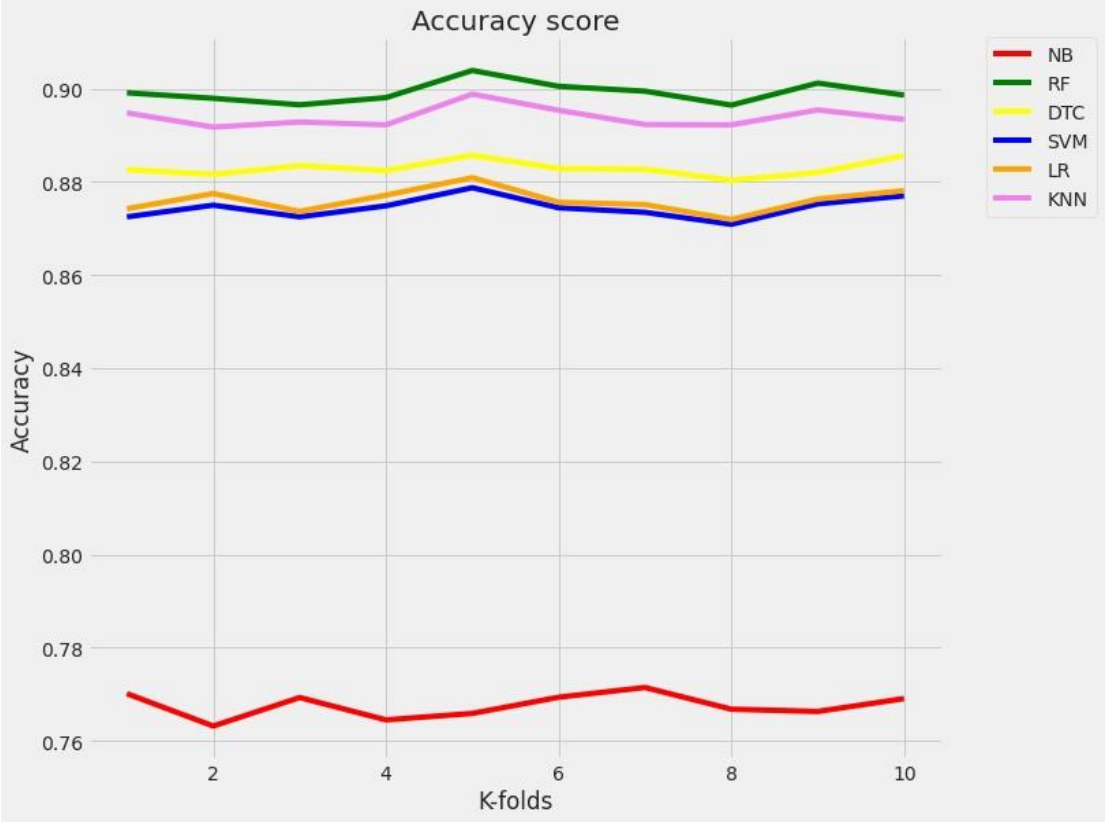
Similar to how it was described previously, Figure 5.2's algorithms must be compared using cross-validation options with the identical performance measurement metrics recall, precision, and F-measure values.

Form the above Table 5-2 Summary of confusion matrix result we have some of false positive rate or Type I errors values and false negative rate or Type II errors value are varied. Which means false positive rate or Type I errors is occur when we see things that are not there. Type II errors or false negative errors occur when we don't see things that are there. As we see from the Table 5-2 above result our Random forest RF classifier batter than the rest of other models. The value RF scores TP = 52638, TF = 2895, FP = 4692(Type 1 error), FP = 1399(type 2 errors). From this result we can conclude that only 9% of the data are under type 1 and type 2 errors. RF model which best in our case is unable to classify effectively only 9% records whether the customer is smisher or not.

The measurements yield various results for the three performance measuring metrics, as shown above in table 5.1. For the three metrics values of precision (0.92), recall (0.97), and F-measure (0.95), respectively, RF received the best score, while DT came in second. For the three metrics values of precision (0.92), recall (0.97), and F-measure (0.94), respectively, DT received the lowest score. Precision (0.87), recall (0.99), and F-measure (0.94) scores for LR were moderate. But for the three assessed metrics of accuracy (0.88), recall (0.99), and F-

measure (0.93) and, respectively, precision (0.95), recall (0.78), and F-measure (0.85), SVM and NB score the lowest values. In contrast, KNN outperforms SVM and NB algorithms in terms of precision (0.92), recall (0.96), and F-measure (0.84) metrics. The best classifier has the highest precision and recall values.

Figure 5-1. Algorithm Comparison

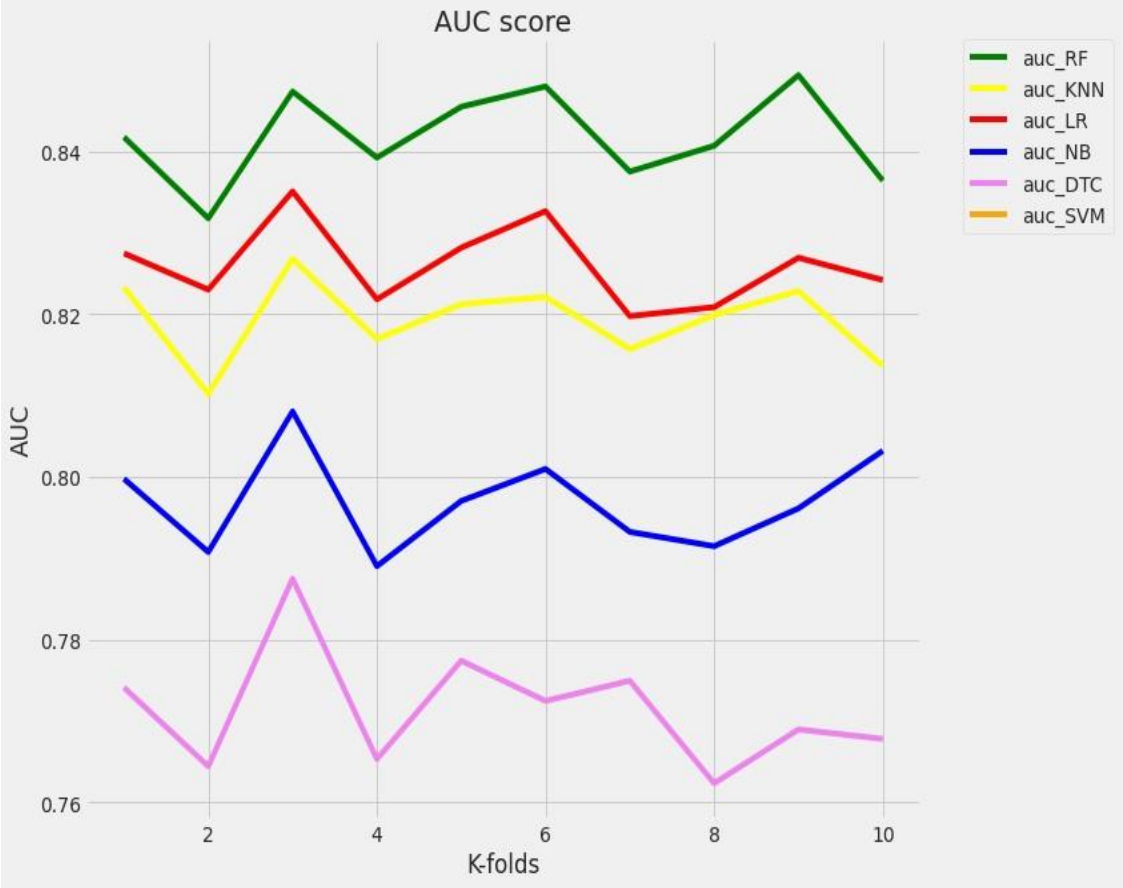


ML algorithm's classification performance can be compared using a graphical representation. ROC curve is one of the graphical representations technique that easily shows the performances of the models and mainly shows the relationship between the True Positive Rate (TPR) and the False Positive Rate (FPR) graphically. Another option for contrasting the effectiveness of the suggested algorithms is the ROC curve. It is a common technique for analyzing classifier performance using a variety of false positive and true positive error rate tradeoffs. According to performance measures the lowest ROC value for each method on the LR, RF, KNN, DT, NB, and SVM with cross-validation was 0.82, 0.84, 0.80, 0.82, and 79, respectively, according to Table 5.1. As can be seen from the results, RF once more receives the highest rating on each ROC value, followed by similarly high ratings for LR, DT, and RF.

However, the highest ROC value for each method is presented as shown in Figure 5.3 for further comparison in order to choose the best classifier based on the Area under the Curve (AUC). SVM is the least among those with the poorest ROC results. ROC Due to their greater coverage of the curve in Figure 5.3, RF is the strongest classifier, followed by LR and DT. The vertical true positive rate axis, which was about at [0.2], was where the RF curve was constructed. This shows that the algorithm was reliable because it assesses the percentage of real positives that are actually appropriately identified as positive values.

In contrast, based on its coverage area, SVM is the least classifier among the five chosen (highest) methods.

Figure 5-2. ROC evaluation against the best classification method



This thesis compares overall performance to determine which algorithms are more effective at identifying SMS phishing attempts based on classification performance using machine learning techniques. This is what was done and discussed in the chapters before, and the results of the cross-validation option comparisons above show the highest accuracy results for each algorithm, which are reported in Table 5.1. The experiments are planned in line with a solution to the problem statement.

The problem statement serves as the foundation for the entire thesis and serves as the impetus for this endeavor. The problem statement provides a clear problem or question that this thesis attempts to address. The aim of this thesis is to attempt to address the question of which classification method has the best overall performance, as stated in the problem description. The problem statement as listed in the introduction of this thesis is:

1. *What kind of CDR data features can be used in order to Detect smishing SMS?*

   This thesis aims to detect smishing SMS fraud based on historical subscriber usage trends data features were identified.

   - Calling number - Identify the phone numbers that place the call.
   - Called number which it shows the destination number
   - Incoming calling number which receives the calls
   - Call type which identifies the call destination
   - Average main balance account in five mounts vs transfer rate within five days
   - Frequency of the main account recharged in the last five months vs balance transfer rate
   - Number of balances recharged by the user in the last 30 days for five months from different channels vs transfer rate
   - The total amount of balance recharged by the user in the last 30 days for five months in self vs transfer rate
   - Start time which indicates the beginning of the calling
   - Duration which details how long the call lasted.
   - SMS which indicating the amount of SMS the subscriber sent/received
   - A Service type that specifies the type of service.

2. *What kinds of ML algorithms can be used in order to detect smishing SMS fraud?*

   There are six alternative implementations of the algorithms in this thesis. These are recommendations for how they might be put into practice to produce results that ought to be optimized.

   Table 5.3 lists each classifier's results with the highest accuracy. Therefore, it can be said that for newly discovered instances of smishing SMS, Random forest RF using cross-validation outperforms better than the other five algorithms in terms of accuracy, precision, recall, and ROC.

Table 5-3. Summary of the highest accuracy

Algorithms	Validation	Accuracy
RF	10-fold CV	0.901
DT	10-fold CV	0.896
KNN	10-fold CV	0.888
LR	10-fold CV	0.878
NB	10-fold CV	0.800
SVM	10-fold CV	0.876

CHAPTER SIX

CONCLUSION AND FUTURE WORK

6. CONCLUSION

The advancement of telecom technology very rapidly leads to certain characteristics of fraudsters. Due to fraudsters' advancement with new technologies and behavioral change, telecom operators and telecom users have suffered from fraudsters' fraudulent activities. A list of telecom frauds categorizes under the behavior of their fraudulent activities. Smishing SMS fraud is one of the common phishing fraud categories where criminals contact potential victims by SMS to trick them into providing personal information or bank account information, which greatly impacted customers' dissatisfaction up to huge revenue loss. Smishing employs a number of techniques to deceive people into submitting personal information. Users are far more likely to trust text messages, hence smishing is frequently profitable for attackers phishing for private information, banking credentials, and other credentials. Even with many controls in the area, smishing SMS fraud is still widespread and affects every telecom operator and customer.

The main target of this research is which algorithms perform better in detecting smishing SMS fraud using ML algorithms. Evaluate, and compare each algorithm based on its classification performance. In order to achieve the objective of this study the researcher used real CDR data from Ethio telecom. Preprocessing tasks were used to eliminate needless missing data values and outlier data in order to obtain a clear instance of the data. Feature selection is important for the study since it allows for the removal of redundant and irrelevant attributes from a data set, which reduces the dimensionality of the data. Depending on the kind of filter and wrapper used, different feature selection techniques are grouped according to the feature assessment measure. Smishing SMS habits were identified to help choose features/attributes with the assistance of domain expert guidance in addition to a reading of related papers. 33 qualities were chosen after deleting non-relevant ones. A subscriber-level aggregate of attributes was used to distinguish between genuine service numbers and fraudulent ones based on the behavior of smishing attacks in order to have complete information on the smisher and achieve the goal of the study.

The studies were carried out with supervised ML algorithms RF, LR, DT, KNN, NB, and SVM employing 10-fold cross-validation approaches and correlation algorithms. The models' experimental findings are noted, assessed, and contrasted.

All classifier methods were evaluated for performance using the cross-validation methodology, correlations, and the metrics used included accuracy, precision, recall, F-measures, ROC, and time. These measurements' results give us a quantitative knowledge of how well they work for detecting smishing SMS. The outcomes of these tests provide us with quantifiable information on how well they work to identify smishing SMS.

As a result, the RF method (90.1%) is a super classifier that perfectly fits the prediction solution of Smishing SMS detection, according to the performance measure evaluation matrices.

Six independent models were built using those ML algorithms, on the 10-fold cross validation training technique, and all six models of the RF algorithm achieves better performance than the other models. RF classifier tends to outperform most without issues of overfitting. It is capable to learn alternate expressions and minimize error pruning accounts for this outcome. By reducing overfitting, pruning decreases the complexity of the final classifier and improves predicted accuracy. The RF classifier does not require feature scaling, and it is more resistant to training dataset noise and the choice of training samples. On the Other hand algorithms with the greatest cross-validation scores were DT and KNN, scoring 89.6% and 88.8%, respectively. The amount of time taken by the RF algorithm with CV for classification is relatively much less than the classification time of the other algorithms in both building and evaluation time.

With regard to smishing SMS identification, this research will be crucial in containing and preventing the current fraudulent danger in telecom operators, especially in developed countries. Additionally, this research gives telecom operators the ability to distinguish between fraudulent and legitimate subscribers, allowing them to reduce revenue losses brought on by fraudsters, increase profitability and customer satisfaction, strengthen customer relationships, and build a stronger brand.

The researcher encountered a variety of difficulties. The difficulty with memory size and obtaining the dataset presented the first barrier. Accessing the sizable wet files was the second challenging task. P preparation of the data and selecting those essential qualities for this investigation was the third challenge. Finally, the difficulty was in comprehending the subject matter. The researcher has made an effort to collaborate with the subject matter specialists. As a result, the researcher spent a lot of time gathering, obtaining, and preparing the CDR data.

6.1. FUTURE WORK

The primary goal of this research is purely academic. This study has demonstrated the usefulness of several machine learning (ML) classification approaches, including DT, Random Forest, KNN, LR, NB, and SVM. These techniques automatically unearth buried knowledge that is intriguing and acknowledged by the subject matter expert.

Using the findings of the study as a foundation, the researcher makes the following suggestions:

1. This study used solely CDR data from prepaid mobile devices for five months; however, additional research including other forms of telecom data is needed.
2. The performance and accuracy of the method may be enhanced by expanding the dataset size.
3. Similar methodologies and procedures can be used to conduct research with extra, unforeseen characteristics.
4. With these comparable methods and algorithms, research into other fraud types can be carried out.
5. The historical behavior of pre-paid service numbers used for smishing SMS is the sole focus of this study. By clicking on URLs that download malware to their phones and looking into the many sorts of smishing texts sent, the research could be conducted.

REFERENCES

[1] D. &. J. A. K. Goel, "Smishing-classifier: a novel framework for detection of smishing attack in mobile environment," *third international conference* , pp. 502-512, 2017.

[2] Kadir, A. F. A., Stakhanova, N., & Ghorbani, A. A. , "Understanding Android financial malware attacks: Taxxonomy, characterization and challengs," *Joural of cyber security and mobility* , pp. 1-52, 2018.

[3] T. worku, Short Message Service Fraud Mitigation Taxonomy: The Case of ethio telecom, Addis Ababa: addis abab University , 2018.

[4] A. A, "The short message service: Standards, infrastructure and innovation," *telematics and informatics,* pp. 559-568, 2014.

[5] K. G. D. .. C. Kahelwala, "Real-time fraud detection in telecommunication network using call pattern analysis," 2017.

[6] "A2P MESSAGING FRAUD FRAMEWORK," in *Ecosystem Forum* , UK, 2015.

[7] M. S. G. B. A. &. L. J. Mdini, "ARCD: a solution for root cause diagnosis in mobile networks.," *14th International Conference on Network and Service Management (CNSM),* pp. 280-284, 2018.

[8] D. Tekeste, "Comparative Analysis of Machine Learning Algorithms for Subscription fraud Detection: The case of ethio telecom," addis ababa university , addis ababa, 2019.

[9] "Understanding android financial malware attacks: Taxonomy, characterization, and challenges," *Journal Of Cyber Security And Mobility*, pp. 1-15, 2018.

[10] "GSMA fraud Munal," GSM association , 2019.

[11] S. &. S. D. Mishra, "Smishing Detector: A security model to detect smishing through SMS content analysis and URL behavior analysis.," *Future Generation Computer Systems,* pp. 803-815, 2020.

[12] S. &. S. D. Mishra, "A content-based approach for detecting smishing in mobile environment.," *In Proceedings of International Conference on Sustainable Computing in Science, Technology and Management (SUSCOM),* 2019.

[13] A. K. &. G. B. B. Jain, " Feature based approach for detection of smishing messages in the mobile environment.,," *Journal of Information Technology Research (JITR),* pp. 12(2), 17-35., 2019.

[14] A. K. &. G. B. B. Jain, "Rule-based framework for detection of smishing messages in mobile environment.," *Procedia Computer Science,* pp. 125, 617-623, 2018.

[15] C. &. S. J. P. Mulliner, "Rise of the iBots: Owning a telco network.," *In 2010 5th International Conference on Malicious and Unwanted Software,* pp. 71-80, 2010.

[16] G. &. K. K. S. Sonowal, "SmiDCA: an anti-smishing model with machine learning approach.," *The Computer Journal,* pp. 61(8), 1143-1157, 2018.

[17] S. G. G. G. &. L. M. Gastellier-Prevost, "Decisive heuristics to differentiate legitimate from phishing sites.," *In 2011 conference on network and information systems security ,* pp. 1-9, 2011.

[18] J. W. M. S. Y. S. S. &. P. J. H. Joo, "S-Detector: an enhanced security model for detecting Smishing attack for mobile computing.," *Telecommunication Systems,* pp. 66, 29-38, 2017.

[19] A. K. &. G. B. B. Jain, " Rule-based framework for detection of smishing messages in mobile environment. Procedia Computer Science,," pp. 125, 617-623, 2018.

[20] A. D. L. J. K. W. M. B. L. &. P. J. H. Kang, " Security considerations for smart phone smishing attacks. In Advances in Computer Science and its Applications: Springer Berlin Heidelberg.," *CSA,* pp. 467-473, 2014.

[21] T. haddish, "constructing predictive model for subscription fraud detection using data mining techniques: the case of ethio-telecom," 2018.

[22] L. &. M. R. A. Auria, "Support vector machines (SVM) as a technique for solvency analysis.," 2008.

[23] C. F. M. A. R. &. A. M. F. Foozy, " Phishing detection taxonomy for mobile device.," *International Journal of Computer Science Issues (IJCSI)*, pp. 10(1), 338-344, 2013.

[24] A. S. Chaudhari, " Security analysis of SMS and related technologies. Research Master Thesis, Dept. of Mathematics and Computer Science, Eindhoven University of Technology.," 2015.

[25] S. &. S. D. Mishra, "Can mobile phone data improve emergency response to natural disasters?.," *IPLoS medicine*, pp. 8(8), e1001085), 2015.

[26] f. tesfaye, "Near-Real Time SIM-box Fraud Detection Using Machine Learning in the case of ethio telecom," 2020.

[27] B. E. R. A. &. M. M. Boukari, "Machine learning detection for smishing frauds.," *In 2021 IEEE 18th Annual Consumer Communications & Networking Conference (CCNC)* , pp. (pp. 1-2). IEEE., 2021.

[28] P. &. S. M. Maan, "Fuzzy Improved Decision Tree Approach for Outlier Detection in SMS.," *International Journal of Computer Applications,* , p. 119(16), 2015.

[29] P. Crocker, "Converged-mobile-messaging analysis and forecasts. Giga Omni Media, New York.," 2013.

[30] Mdini, M., Simon, G., Blanc, A., & Lecoeuvre, J. , "ARCD: a solution for root cause diagnosis in mobile networks.," *In 2018 14th International Conference on Network and Service Management (CNSM)* , pp. (pp. 280-284). IEEE., (2018, November).

[31] Boukari, B. E., Ravi, A., & Msahli, M., " Machine learning detection for smishing frauds.," *In 2021 IEEE 18th Annual Consumer Communications & Networking Conference (CCNC)* , pp. (pp. 1-2). IEEE., (2021, January)..

[32] "What is a Core Network? - Definition from Techopedia".

[33] Niu, K., Jiao, H., Deng, N., & Gao, Z. , "A real-time fraud detection algorithm based on intelligent scoring for the telecom industry.," *In 2016 International Conference on Networking and Network Applications (NaNA),* pp. (pp. 303-306). IEEE., (2016, July). .

[34] CFCA Survey Group, ""2019 Fraud Loss Survey," www.CFCA.org, New Jersey," 2018.

[35] Sultan, K., Ali, H., & Zhang, Z. , " Call detail records driven anomaly detection and traffic prediction in mobile cellular networks.," pp. IEEE Access, 6, 41728-41737., 2018.

[36] "https://seon.io/resources/telecommunications-fraud-detection-and-prevention/".

[37] L. M. Pompilio, "Analysis of 10 CFR Part 810 General Authorization Data on Assistance to Foreign Atomic Energy Activities (Doctoral dissertation).," (2017)..

[38] "https://www.expert.ai/blog/machine-learning-definition/".

[39] |. CFCA, " CFCA - Communications Fraud Control Association," www.cfca.org, New Jersey," *Fraud Loss Survey ,* 2019.

[40] Yadav, K., Kumaraguru, P., Goyal, A., Gupta, A., & Naik, V. , " SMSAssassin: Crowdsourcing driven mobile-based system for SMS spam filtering.s," *In Proceedings of the 12th Workshop on Mobile Computing Systems and Application,* pp. (pp. 1-6), (2011, March)..

[41] Wu, L., Du, X., & Wu, J., "MobiFish: A lightweight anti-phishing scheme for mobile phones.," *In 2014 23rd international conference on computer communication and networks (icccn),* pp. (pp. 1-8). IEEE., (2014, August).

[42] Kadir, A. F. A., Stakhanova, N., & Ghorbani, A. A. , " Understanding android financial malware attacks: Taxonomy, characterization, and challenges.," *Journal of Cyber Security and Mobility,,* pp. 7(3), 1-52, (2018).

[43] E. Ulker-Demirel, " Development of Digital Communication Technologies and the New Media.," *In Handbook of Research on Narrative Advertising,* pp. (pp. 164-175). IGI Global., (2019).

[44] A. A. &. A. A. Soofi, "Classification techniques in machine learning: applications and issues. J," *Basic Appl. Sci,,* pp. 13, 459-465, 2017.

[45] Ouali, Y., Hudelot, C., & Tami, M., " An overview of deep semi-supervised learning. arXiv preprint arXiv:2006.," p. 05278., (2020).

[46] Yadav, K., Kumaraguru, P., Goyal, A., Gupta, A., & Naik, V. , "SMSAssassin: Crowdsourcing driven mobile-based system for SMS spam filtering.," *In Proceedings of the 12th Workshop on Mobile Computing Systems and Applications,* pp. (pp. 1-6)., (2011, March).

[47] L. &. M. R. A. Auria, "Support vector machines (SVM) as a technique for solvency analysis," (2008)..

[48] Naboulsi, D., Stanica, R., & Fiore, M., "Classifying call profiles in large-scale mobile traffic datasets.," *In IEEE INFOCOM 2014-IEEE conference on computer communications ,* pp. (pp. 1806-1814). IEEE, 2014.

[49] H. Joe Steinhauer , Tove Helldin, Gunnar Mathiason, Alexander Karlsson, , "“Topic modeling for smishing detection in telecommunication networks”," January 2019..

[50] Bala, R., & Kumar, D. , "Classification using ANN: A review. Int. J. Comput. Intell. Res,," pp. 13(7), 1811-1820, (2017). .

[51] Fei Deng, Jibing Huang, Xiaoling Yuan, Chao Cheng & Lanjing Zhang, "Performance and efficiency of machine learning algorithms for analyzing rectangular biomedical data," *laboratory investigation,* p. pages430–441 (2021), 11 February 2021.

[54] "Universal Mobile Telecommunications System (UMTS); Telecommunication management;".

[55] D. Tekeste, " Comparative Analysis of Machine Learning Algorithms for Subscription fraud Detection: The case of ethio telecom".

[56] Houshmand, "Houshmand “SMS Spam Detection using Machine Learning Approach”".

[57] A. Chaudhari, " “Security analysis of SMS and related technologies”," 2015.

[58] M. S. Priyanka Maan, "“Fuzzy Improved Decision Tree Approach for Outlier Detection in SMS," *” International Journal of Computer Applications),* vol. Volume 119 – No., p. (0975 – 8887), 16, June 2015.

[59] T. H. G. M. A. K. H. Joe Steinhauer, " “Topic modeling for smishing detection in telecommunication networks”," January 2019. .

[60] "Subscriber and equipment trace; Trace control and configuration management," *” 3GPP, Tech. ,* Vols. TS 32.422, V7.4.0, 2017.

[61] F. L. a. N. A. J. Liu, " “Monitoring and analyzing big traffic data of a large-scale cellular network with Hadoop,," *” IEEE Network,,* 2014.

[62] G. K. a. A. Chhabra, " “Improved j48 classification algorithm for the prediction of diabetes," *” International Journal of Computer Applications,* Vols. vol. 98, , no. 22, 2014.

[63] W. Kellerer, A. Basta, P. Babarczi, A. Blenk, M. He, M. Klügel, and M. Alba, " “How to measure network flexibility? A proposal for evaluating softwarized networks," *” IEEE Communications Magazine",* 2018.

[64] H. J. N. D. a. Z. G. K. Niu, " “A real-time fraud detection algorithm based on intelligent scoring for the telecom industry,” Proceedings – 2016 International Conference on Networking and Network Applications,," Vols. vol. 1, , pp. pp. 303–306, 2, NaNA 2016,.

[65] S. S. Aksenova, " “Machine learning with weka weka explorer tutorial for weka version 3.4. 3," *” sabanciuniv. edu,,* 2004..

[66] "https://seon.io/resources/telecommunications-fraud-detection-and-prevention/".

[67] "Understanding Android Financial Malware Attacks," *Journal of Cyber Security and Mobility,* January 2018.

[68] "https://www.expert.ai/blog/machine-learning-definition/".

[69] Y. W. Qiong Liu, "‘SUPERVISED LEARNING’ Classification Techniques in Machine Learning: Applications and Issues".

[70] D. D. K. Rajni Bala1, "‘’ Classification Using ANN:’’," *International Journal of Computational Intelligence Research. ISSN 0973-1873,* vol. Volume 13, pp. pp. 1811-1820, Number 7 (2017).

[71] Prof. Ziad AlQadi, " Digital Image processing using Artificial Neural Networks".

[72] H. H. Volden., "Anomaly detection using machine learning techniques A comparison of classification algorithms.," *Master’s Thesis Spring,* 2016.

[73] A. A. a. A. F. Natalia Stakhanova, " Understanding Android Financial Malware Attacks: Taxonomy, Characterization, and Challenges," *Article in Journal of Cyber Security and Mobility* , January 2018.

[74] a. A. A. Aized Amin Soofi, " Classification Techniques in Machine Learning: Applications and Issues,," *Journal of Basic & Applied Sciences,,* p. 2017, 13, 459-465.

[75] "https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2678744/".

7. Appendix

DATASET

Service_type	Billing_Number	Calling_Number	Outgoing_call	Called_Number	label	msisdn	age	daily_spent_1m	daily_spent_5m	Avg_main_bal1m	Avg_main_bal5m	las_rech_	last_rech_date_da	last_rech_amt_ma
On-net Voice	960517036.0000...	251960517...	0.000000000000...	25197338...	0	21408I...	272.0000	3055.050000...	3065.1500000...	220.13	260.13	2.0000	.0000	1539
On-net Voice	962360145.0000...	251962360...	2.519495564360...	25197338...	1	76462I...	712.0000	12122.00000...	12124.750000...	3691.26	3691.26	20.0...	.0000	5787
On-net Voice	962360145.0000...	251962360...	2.519224512300...	25197338...	1	17943I...	535.0000	1398.000000...	1398.0000000...	900.13	900.13	3.0000	.0000	1539
On-net Voice	962360145.0000...	251962360...	2.519313774290...	25197338...	1	55773I...	241.0000	21.22800000...	21.228000000...	159.42	159.42	41.0...	.0000	947
On-net Voice	928849360.0000...	251928849...	2.519313774290...	25197338...	1	03813I...	947.0000	150.6193333...	150.61933330...	1098.90	1098.90	4.0000	.0000	2309
On-net Voice	928524540.0000...	251928524...	2.519787937910...	25197338...	1	35819I...	568.0000	2257.362667...	2261.4600000...	368.13	380.13	2.0000	.0000	1539
On-net Voice	928524540.0000...	251928524...	0.000000000000...	25197338...	1	96759I...	545.0000	2876.641667...	2883.9700000...	335.75	402.90	13.0...	.0000	5787
On-net SMS	928524540.0000...	251928524...	0.000000000000...	25197338...	1	09832I...	768.0000	12905.00000...	17804.150000...	900.35	2549.11	4.0000	55.0000	3178
On-net Voice	928524540.0000...	251928524...	0.000000000000...	25197338...	1	59772I...	1191.0000	90.69500000...	90.695000000...	2287.50	2287.50	1.0000	.0000	1539
On-net SMS	929253206.0000...	251929253...	0.000000000000...	25197338...	1	56331I...	536.0000	29.35733333...	29.357333330...	612.96	612.96	11.0...	.0000	773
On-net Voice	929253206.0000...	251929253...	0.000000000000...	25197338...	1	32893I...	1511.0000	12.89600000...	12.896000000...	790.44	790.44	8.0000	.0000	1539
On-net Voice	928849360.0000...	251928849...	0.000000000000...	25197338...	0	82417I...	82.0000	65.16666667...	65.166666670...	326.20	326.20	17.0...	.0000	7526
On-net Voice	928849360.0000...	251928849...	0.000000000000...	25197338...	1	11435I...	154.0000	227.0410000...	227.04100000...	240.41	240.41	2.0000	.0000	1547
On-net Voice	960517036.0000...	251960517...	0.000000000000...	25197338...	1	66580I...	887.0000	55.90933333...	55.909333330...	208.80	208.80	2.0000	.0000	1539
On-net Voice	929253206.0000...	251929253...	0.000000000000...	25197338...	1	63139I...	707.0000	8919.000000...	10317.350000...	399.25	2453.78	3.0000	.0000	770
On-net Voice	900878444.0000...	251900878...	0.000000000000...	25197338...	0	24075I...	1037.0000	12.00000000...	12.000000000...	1216.80	1216.80	.0000	.0000	0
On-net Voice	960020471.0000...	251960020...	2.519641243230...	25197338...	0	82053I...	1583.0000	1000.000000...	1000.0000000...	1000.80	1087.88	.0000	.0000	0
On-net Voice	986370210.0000...	251986370...	0.000000000000...	25197338...	1	37204I...	929.0000	10.68800000...	10.688000000...	40.00	40.00	.0000	.0000	0
On-net Voice	960448493.0000...	251960448...	2.519397712110...	25197338...	1	44217I...	832.0000	14.40000000...	14.400000000...	1660.96	1660.96	1.0000	.0000	2309
On-net Voice	960448493.0000...	251960448...	2.519742353760...	25197338...	1	19611I...	450.0000	48.93500000...	48.935000000...	726.30	726.30	1.0000	.0000	1539
On-net Voice	960448493.0000...	251960448...	2.519108456270...	25197338...	1	67813I...	100.0000	769.6140000...	777.46000000...	1050.57	1167.30	6.0000	.0000	770

Aggregating data

```
GET
  FILE='C:\Users\samson\Desktop\merg\mod.csv'.
DATASET NAME DataSet1 WINDOW=FRONT.
SORT CASES BY label Service_type.
AGGREGATE
  /OUTFILE=* MODE=ADDVARIABLES
  /PRESORTED
  /BREAK=label Service_type
  /Billing_Number_mean=MEAN(Billing_Number)
  /Calling_Number_mean=MEAN(Calling_Number)
  /Outgoing_call_mean=MEAN(Outgoing_call)
  /Called_Number_mean=MEAN(Called_Number)
  /msisdn_first=FIRST(msisdn)
  /aon_mean=MEAN(aon)
  /daily_decr30_mean=MEAN(daily_decr30)
  /daily_decr150_mean=MEAN(daily_decr150)
  /Avg_main_bal30_mean=MEAN(Avg_main_bal30)
  /Avg_main_bal150_mean=MEAN(Avg_main_bal150)
  /last_rech_amt_ma_mean=MEAN(last_rech_amt_ma)
```

/cnt_ma_rech_chls30_mean=MEAN(cnt_ma_rech_chls30)

/Fr_ma_rech_chls30_mean=MEAN(Fr_ma_rech_chls30)

/sumamnt_ma_rech_chls30_mean=MEAN(sumamnt_ma_rech_chls30)

/cnt_ma_rech150_mean=MEAN(cnt_ma_rech150)

/Fr_ma_rech_chls150_mean=MEAN(Fr_ma_rech_chls150)

/sumamnt_ma_rech_chls150_mean=MEAN(sumamnt_ma_rech_chls150)

/cnt_da_rech_chls30_mean=MEAN(cnt_da_rech_chls30)

/fr_da_rech_cnls30_mean=MEAN(fr_da_rech_cnls30)

/cnt_da_rech_chls150_mean=MEAN(cnt_da_rech_chls150)

/fr_da_rech_chls150_mean=MEAN(fr_da_rech_chls150)

/cnt_bal_chls30_mean=MEAN(cnt_bal_chls30)

/T_amnt_bal30_mean=MEAN(T_amnt_bal30)

/cnt_bal_chls150_mean=MEAN(cnt_bal_chls150)

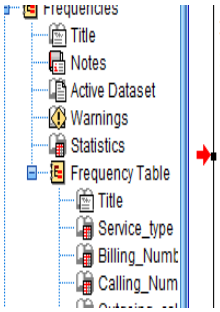
/T_amnt_bal150_mean=MEAN(T_amnt_bal150)

/medianamnt_bal150_mean=MEAN(medianamnt_bal150)

/transfer30_mean=MEAN(transfer30)

/transfer150_mean=MEAN(transfer150).

Data cleaning




```
DATASET NAME DataSet1 WINDOW=FRONT.
FREQUENCIES VARIABLES=Service_type Billing_Number Calling_Number Outgoing_call Called_Number label
    aon daily_decr30 daily_decr90 Avg_main_bal30 Avg_main_bal90 last_rech_date_ma last_rech_date_da
    last_rech_amt_ma cnt_ma_rech_chls30 Fr_ma_rech_chls30 sumamnt_ma_rech_chls30
    medianamnt_ma_rech_chls30 medianmarechprebal_chls30 cnt_ma_rech90 Fr_ma_rech_chls90
    sumamnt_ma_rech_chls90 medianamnt_ma_rech_chls90 medianmarechprebal90 cnt_da_rech_chls30
    fr_da_rech_cnls30 cnt_da_rech_chls90 fr_da_rech_chls90 cnt_bal_chls30 T_amnt_bal30 maxamnt_bal30
    medianamnt_bal30 cnt_bal_chls90 T_amnt_bal90 maxamnt_bal90 medianamnt_bal90 transfer30 transfer90
    pcircle pdate
```

Data Integration

```
ADD FILES /FILE=*
  /RENAME (aon Avg_main_bal30 Avg_main_bal90 Billing_Number Called_Number Calling_Number
    cnt_bal_chls30 cnt_bal_chls90 cnt_da_rech_chls30 cnt_da_rech_chls90 cnt_ma_rech90
    cnt_ma_rech_chls30 daily_decr30 daily_decr90 fr_da_rech_chls90 fr_da_rech_cnls30 Fr_ma_rech_chls30
    Fr_ma_rech_chls90 last_rech_amt_ma last_rech_date_da last_rech_date_ma maxamnt_bal30 maxamnt_bal90
    medianamnt_bal30 medianamnt_bal90 medianamnt_ma_rech_chls30 medianamnt_ma_rech_chls90
    medianmarechprebal90 medianmarechprebal_chls30 Outgoing_call pcircle pdate sumamnt_ma_rech_chls30
    sumamnt_ma_rech_chls90 T_amnt_bal30 T_amnt_bal90 transfer30 transfer90=d0 d1 d2 d3 d4 d5 d6 d7 d8
    d9 d10 d11 d12 d13 d14 d15 d16 d17 d18 d19 d20 d21 d22 d23 d24 d25 d26 d27 d28 d29 d30 d31 d32 d33
    d34 d35 d36 d37)
  /FILE='C:\Users\samso\Desktop\merg\modify.sav'
  /RENAME (aon Avg_main_bal30 Avg_main_bal90 Billing_Number Called_Number Calling_Number
    cnt_bal_chls30 cnt_bal_chls90 cnt_da_rech_chls30 cnt_da_rech_chls90 cnt_ma_rech90
    cnt_ma_rech_chls30 daily_decr30 daily_decr90 fr_da_rech_chls90 fr_da_rech_cnls30 Fr_ma_rech_chls30
    Fr_ma_rech_chls90 label last_rech_amt_ma last_rech_date_da last_rech_date_ma maxamnt_bal30
    maxamnt_bal90 medianamnt_bal30 medianamnt_bal90 medianamnt_ma_rech_chls30 medianamnt_ma_rech_chls90
    medianmarechprebal90 medianmarechprebal_chls30 Outgoing_call pcircle pdate Service_type
    sumamnt_ma_rech_chls30 sumamnt_ma_rech_chls90 T_amnt_bal30 T_amnt_bal90 transfer30 transfer90=d38
    d39 d40 d41 d42 d43 d44 d45 d46 d47 d48 d49 d50 d51 d52 d53 d54 d55 d56 d57 d58 d59 d60 d61 d62 d63
    d64 d65 d66 d67 d68 d69 d70 d71 d72 d73 d74 d75 d76 d77)
  /DROP=d0 d1 d2 d3 d4 d5 d6 d7 d8 d9 d10 d11 d12 d13 d14 d15 d16 d17 d18 d19 d20 d21 d22 d23 d24
    d25 d26 d27 d28 d29 d30 d31 d32 d33 d34 d35 d36 d37 d38 d39 d40 d41 d42 d43 d44 d45 d46 d47 d48 d49
    d50 d51 d52 d53 d54 d55 d56 d57 d58 d59 d60 d61 d62 d63 d64 d65 d66 d67 d68 d69 d70 d71 d72 d73 d74
    d75 d76 d77.
EXECUTE.
```

Table 7-1 Training data

```
25405,15246,251922297401,251922297401,15246,similar,normal,normal,0.360070852,OFFPEAK,NSunday,local
22560,13536,251920075316,251920075316,13536,similar,normal,normal,0.36,OFFPEAK,Sunday,local,local,f
20992,12600,251928699682,251928699682,12600,similar,normal,normal,0.360137195,OFFPEAK,NSunday,local
20990,12594,251923138852,251923138852,12594,similar,normal,normal,0.36,OFFPEAK,NSunday,local,local,
17124,0,251924872417,251921498870,-10278,different,abnormal,abnormal,0,OFFPEAK,NSunday,local,local,
16218,0,251911407923,251926872125,-9732,different,abnormal,abnormal,0,OFFPEAK,Sunday,local,local,fr
15600,0,251924872466,251921020694,-9360,different,abnormal,abnormal,0,OFFPEAK,NSunday,local,local,f
13620,0,251913293602,251910516922,-8172,different,abnormal,abnormal,0,OFFPEAK,NSunday,local,local,f
13126,0,251926126111,251931319102,-7878,different,abnormal,abnormal,0,OFFPEAK,NSunday,local,local,f
12960,7776,251924154641,251925828075,777,different,normal,abnormal,0.36,OFFPEAK,Sunday,local,local,
12618,17668,251926980702,251932289035,1766,different,normal,abnormal,0.840133143,PEAK,NSunday,local
12582,7554,251913293602,251910516922,755,different,normal,abnormal,0.360228898,OFFPEAK,NSunday,loca
12561,17598,251932100338,251913271394,1759,different,normal,abnormal,0.840601863,PEAK,NSunday,local
12472,7488,251917827851,251917631252,748,different,normal,abnormal,0.360230917,OFFPEAK,Sunday,local
12110,7266,251911798632,251927549148,7266,different,normal,normal,0.2,OFFPEAK,NSunday,local,local,f
11280,6768,251913293602,251910516922,6768,different,normal,normal,0.22,OFFPEAK,NSunday,local,local,
```

Table 7-2 testing Data

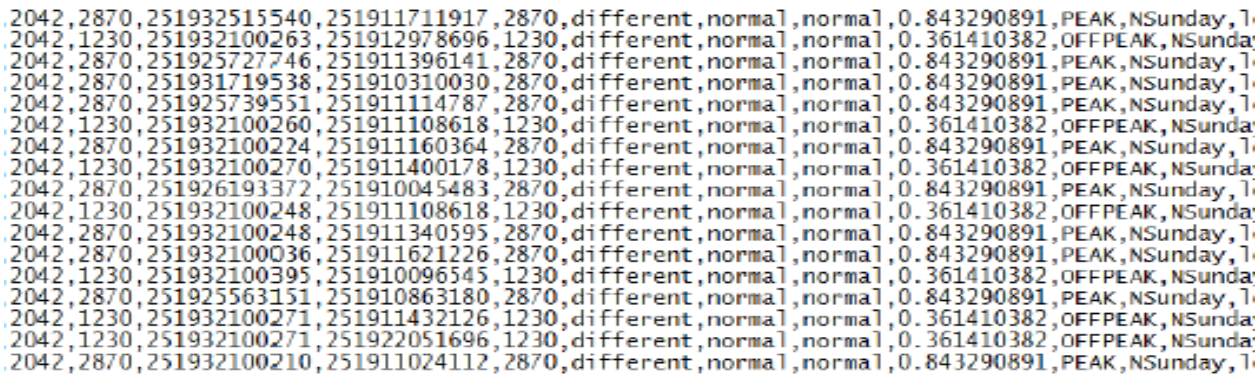


Table 7-3 normalization

	label	age	daily_spent_1m	daily_spent_5m	Avg_main_bal_1m	Avg_main_bal_5m	last_rech_date_ma	incoming_call	last_rech_amt_ma	Time_ma_rech_		Time_ma_rech_1m	cnt_bal_rec_1m	Tot_bal_rec1m	maxmnt_bal_rec1m	med_bal_rec1m	cnt_bal_rec5m	Tot_bal_rec5m	maxmnt_bal_rec5m	Tot_bal_rec5m	Calling_Number
0	0	272.0	3055.050000	3065.150000	220.13	260.13	2.0	0.0	1539	2	...	2	12	6.0	0.0	2.0	12	6	0.0	29.000000	29.000000
1	1	712.0	12122.000000	12124.750000	3691.26	3691.26	20.0	0.0	5787	1	...	1	12	12.0	0.0	1.0	12	12	0.0	0.000000	0.000000
2	1	535.0	1398.000000	1398.000000	900.13	900.13	3.0	0.0	1539	1	...	1	6	6.0	0.0	1.0	6	6	0.0	0.000000	0.000000
3	1	241.0	21.228000	21.228000	159.42	159.42	41.0	0.0	947	0	...	2	12	6.0	0.0	2.0	12	6	0.0	0.000000	0.000000
4	1	947.0	150.619333	150.619333	1089.90	1089.90	4.0	0.0	2309	7	...	7	42	6.0	0.0	7.0	42	6	0.0	2.333333	2.333333
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
205407	1	492.0	5829.000000	7438.860000	600.39	2117.82	14.0	0.0	1539	3	...	2	12	6.0	0.0	6.0	36	6	0.0	0.000000	5.333333

2054108	1	99.0	2794.780000	2799.400000	1481.55	1562.19	9.0	0.0	1924	2	...	2	12	6.0	0.5	2.0	12	6	0.5	0.0000000	0.0000000
2054109	1	1298.0	48840.000000	57872.820000	6674.09	7561.57	1.0	0.0	80000	7	...	5	30	6.0	0.0	8.0	48	6	0.0	1.5000000	2.444444
2054110	1	197.0	29.531667	29.531667	78.00	78.00	6.0	0.0	2309	1	...	1	6	6.0	1.0	1.0	6	6	1.0	0.0000000	0.0000000
2054111	1	260.0	4894.000000	6849.700000	300.26	680.39	20.0	0.0	3178												