Faculty of Engineering+

MASTER OF SCIENCE IN SOFTWARE ENGINEERING

**ZNA’s Telephone and Internet Usage-Bill Anomaly** **Detection Model (T&I-UADM)**

by

Wellington Murakata

(D23101121B)

A review of chapter (1-3) submitted to the Zimbabwe National Defense University in a partial fulfilment of requirements for the award of the Master of Science Degree in Software Engineering

HARARE, ZIMBABWE

RESEARCH SUPERVISOR: DR R CHIWARIRO

30 June 2024

Contents

**CHAPTER 1: INTRODUCTION** 8

**1.1** **Introduction** 8

**1.2** **Background to the study** 9

**1.3** **Statement of the problem** 10

**1.4** **Conceptual Framework** 10

**1.5** **Research Objectives** 11

**1.6** **Research Questions** 11

**1.7** **Significance of the Study** 12

**1.7.1** **The Research** 12

**1.7.2** **ZNA** 12

**1.7.3** **ZNDU** 12

**1.8** **Delimitations of the Study/Scope of the Study** 12

**1.9** **Limitations of the Study** 13

**1.10** **Document structure** 14

**1.11** **Chapter Summary** 14

**CHAPTER TWO: LITERATURE REVIEW** 15

**2.1** **Introduction** 15

**2.1.1** **Lack of Awareness Theory** 15

**2.1.2** **Insufficient Supervision Theory** 16

**2.1.3** **Telephone-internet Excessive Usage** 17

**2.1.4** **Impact of Telephone-internet Anomaly Detection** 18

**2.1.5** **Factors that affect Telephone-internet Abuse** 18

**2.2** **Explore Discoveries Obtained from Telephone-internet Usage-bill data** 19

**2.3** **Telephone-internet Usage** 20

**2.3.1** **Voice Calls** 20

**2.3.2** **Messaging** 21

**2.3.3 Internet Usage** 21

**2.3.4** **Email** 21

**2.3.5** **Social Media** 21

**2.3.6** **Online Communication** 21

**2.3.7** **Streaming and Entertainment** 21

**2.3.8** **Online Transactions** 21

**2.3.9** **Cloud Services** 22

**2.4** **Call Duration and Internet Usage Package Limits** 22

**2.4.1** **Call Duration** 22

**2.4.2** **Monthly or Daily Call Limits** 22

**2.4.3** **Data Caps** 22

**2.4.4** **Limiting Speed** 22

**2.4.5** **Time-Based Limitations** 23

**2.4.6** **Day of the Week and Month Telephone-internet usage** 23

**2.4.7** **Time of the day** 23

**2.5** **Machine Learning (Identification of Patterns-Relationships)** 23

**2.5.1** **Supervised Learning Telephone-internet Usage Prediction and Anomaly Detection** 24

**2.5.2** **Random Forest** 25

**2.5.3** **Gradient Boosting** 25

**2.5.4 Decision Trees** 25

**2.5.5 Gaussian NB** 26

**2.5.6 Support Vector Machines (SVM)** 26

**2.5.7 Logistic Regression** 26

**2.6** **Application of Ensemble Techniques On The Best Algorithms** 26

**2.6.1** **Max Voting / Voting Classifier** 27

**2.6.2** **Averaging** 27

**2.6.3** **Blending** 28

**2.6.4** **Bagging** 28

**2.6.5** **Boosting** 29

**2.6.6** **Stacking** 29

**2.7** **Implement the algorithm in detecting Telephone-internet Usage Anomalies** 30

**2.8** **Related studies** 30

**2.8.1** **In Cerda-Alabern, Iuhasz and Gemmi, (2023)** 30

**2.8.2** **In Cerdà-Alabern, Iuhasz and Gemmi, (2023)** 31

**2.8.3** **Jurisic, Tomicic and Grd, (2023)** 31

**2.8.4** **Jurisic, Tomicic and Grd, (2023)** 32

**2.8** **Comparison of results from related work** 33

**2.9** **Conceptual framework** 34

**2.10** **Chapter Summary** 35

**CHAPTER 3: RESEARCH METHODOLOGY** 37

**3.1** **Introduction** 37

**3.2** **Research philosophy** 37

**3.3** **Research approach** 37

**3.4** **Research design** 38

**3.5** **Data collection methods** 39

**3.6** **Data Collection Tools-Instruments** 39

**3.7** **Population of the study** 40

**3.9** **Sources of data** 40

**3.10** **Validity and Reliability** 41

**3.11** **Data Collection Procedures** 41

**3.12** **Data Analysis Framework** 41

**3.12.1** **Business Understanding:** 42

**3.12.1.1** **Define the business objectives** 42

**3.12.2** **Data Understanding** 43

**3.12.2.1 Data Preperatiom** 43

**3.12.2.2** **Data collection** 43

**3.12.2.3** **Importing Libraries** 44

**3.12.2.4** **Data Extraxction** 45

**3.12.3** **Data Exploration Analysis (EDA)** 45

**3.12.3.1**  **(Examining the Data in Python)** 45

**3.12.4.1** **Data Cleaning and Preparation** 47

**3.12.4.2** **Handling missing values** 47

**3.12.4.3** **Handling outliers** 48

**3.12.4.4** **Standardizing or Normalizing** 49

**3.12.4.5** **Encoding Categorical Variables** 49

**3.12.4.6** **Manipulating Dates and Times** 50

**3.12.4.7** **Feature Extraction** 50

**3.12.5** **Evaluation** 51

**3.12.6** **Deployment** 51

**3.12.7** **Monitoring and Maintenance** 52

**3.12.10** **Ethical Consideration** 52

**3.12.11** **Chapter Summary** 52

**CHAPTER 4: RESULTS PRESENTATION AND ANALYSIS** 53

**4.1** **Introduction** 53

**4.2** **Exploratory data analysis results** 53

**4.2.1** **Descriptive statistics of the dataset** 53

**4.2.2** **Correlations** 54

**4.2.3** **Transaction amount** 55

**4.2.4** **Gender in relation to misuse** 57

**4.2.5** **Spending Category in relation to Fraud** 57

**4.2.6** **Age in relation to fraud** 59

**4.2.7** **Cyclicality of Telephone-internet Abuse** 59

**4.2.8** **State in relation to Fraud** 61

**4.3** **Train unsupervised machine learning algorithms to detect telephone-internet usage anomaly detection** 62

**4.3.1** **Decision tree** 62

**4.3.2** **Logistic regression** 63

**4.3.3** **Random forest** 63

**4.3.4** **XGboost** 64

**4.3.5** **Naïve Bayes** 65

**4.3.6** **Evaluation of the algorithms** 65

**4.4** **Ensemble the best unsupervised machine learning algorithms to detect credit card fraud** 68

**4.5** **Implementing the algorithm in detecting credit card fraud** 68

**4.6** **Comparative analysis** 69

**4.7** **Chapter summary** 69

**CHAPTER FIVE: CONCLUSION AND RECOMMENDATIONS’** 70

**5.1** **Introduction** 70

**5.2** **Summary of findings** 70

**5.3** **Conclusion** 70

**5.4** **Recommendations** 70

**5.5** **Summary** 71

**CHAPTER 1: INTRODUCTION**

**• Introduction**

The widespread use of Telephone and Internet services come along with potential abuse and excessive usage, which can result in increased costs, compromised network performance and security risks, Carrillo-Mondejar, Martinez and Suarez-Tangil, (2022). The Design of Zimbabwe National Army (ZNA)'s Telephone and Internet Usage Anomaly Detection Model (T&I-UADM) aims to develop an advanced system that can detect anomalies associated with abuse and excessive use of telephone and internet services. The T&I-UADM project recognizes the importance of identifying and addressing misuse and anomalies in telephone and internet usage. This project designs a machine learning (ML) model tool, that predicts telephone-internet usage and detect anomalies or excessive usage of telephone and internet services. As noted by Ayamga, (2018), technologies with network administrators are there to ensure responsible and secure use of communication channels by every telephone and internet user.

Engaging in abusive activities have a negative impact on individuals, organizations and society as a whole. Network congestion is a common concern that affects telephone and internet users, it occurs when the demand for network resources exceeds the available capacity, leading to a significant impact on the performance, user experience and network security. Sigorta *et al.*, (2022) cited that, network misuse and abuse of network services such as conducting malicious activities, engaging in illegal file sharing and excessive use of bandwidth beyond acceptable usage policies, can lead to network congestion. Accurate billing and monitoring of telephone-internet usage is essential for financial management, security and operational efficiency.

The ZNA relies heavily on effective communication channels and secure information sharing to achieve mission success. Traditional monitoring methods in use are absolete and fall short in providing informative decision support tools that allows threat detection and efficient resource management. Telephone-internet abuse is the main cause of communication failure, network congestion and unnecessary financial loss. Therefore, the researcher proposes to introduce a ZNA’s T&I-UADM using ML Algorithm to predict excessive resource consumption and address telephone and internet abuse.

**• Background to the study**

Research on the design of a ML-based T&I-UADM began in the early 2000s, Carrillo-Mondejar, Martinez and Suarez-Tangil, (2022). During this period, organizations such as University of Cambridge Computer Laboratory, University of Oxford Department of Engineering Science and British Telecom (BT) Labs started exploring the potential ML techniques to analyze and detect patterns of abuse in telephone-internet data. According to S. Abella *et al.*, (2024) Researchers collaborated with telecommunications providers to collect and analyze usage-bill datasets. These datasets consisted of call records, network logs, internet traffic data and other relevant information. The datasets helped in training and evaluating ML models for predicting and detecting telephone-internet abuse.

In the European nations, where regulations and policies related to telephone-internet abuse were well-established, research focused on designing and implementing preventive measures aligned with existing legal frameworks, Persky, (2023). Organizations such as Department for Culture, Media and Sport (DCMS), BT and University of Cambridge Computer Laboratory collaborated to develop ML models and tools for identifying and mitigating excessive usage and malicious activities. These applications of ML-based preventive measures expanded to Africa, including North African, Central/East African, Southern African and Zimbabwean regions. Union and Sector, (2021), indicated that researchers applied preventive measures to suit specific regional countries, they considered factors such as infrastructure limitations, cultural dynamics, regulatory environments and socioeconomic conditions. The aim was to develop effective models and policies that address telephone-internet abuse in each region. In Zimbabwe, Gtel could track lost or stolen mobile smartphones by maintaining a record that relates their customer’s mobile smartphones and the time when a user is changed. This information could be used in investigating the whereabouts of the lost or stolen mobile smartphone. Throughout the research process, collaborative efforts and knowledge sharing played a vital role. Researchers from different organizations, including academia, industry, government and non-profit sectors, collaborates to share insights, exchange best practices and develop comprehensive approaches to telephone-internet usage prediction.

Within the ZNA, a T&I-UADM holds significant importance for maintaining operational security, safeguarding sensitive information and ensuring efficient communication. As discussed by Syariah and Ilmu, (2021) Such a system helps to manage and control telephone-internet usage within the military networks, enforce stipulated limits, detect and highlights misuse, security breaches and facilitate secure and cost-effective communication channels. The implementation of a T&I-UADM within the ZNA would involve aligning the system with national regulations, integrating it into existing military networks and ensuring interoperability with other military systems. The research on ML-based T&I-UADMs is an ongoing process. Organizations and researchers continuously explore new techniques, refine existing models and adapt advanced preventive measures to address emerging challenges and evolving patterns of abuse.

**1.3 Statement of the problem**

Telephone-internet services are becoming expensive to mantain due to lack of supervision, adminstators are finding it difficult to predict organisational telephone-internet usage pattens that facilitate effective resource management, optimize budgets and enhance network security, within ZNA communication networks. Some unsupervised users rapidly use up their allocated monthly telephone-internet usage limits, while others let their packages expire without being used up. The ZNA's incapacity to properly account for its telephone-internet usage packages results in unnecessary expenses and erratic communication networks. The current system requires the administrators to physically review and evaluate printed bill statements from the Telephone-internet Service Providers in order to identify anomalies. The manual inspection is time consuming, prone to errors and associated with many other challenges and concerns of telephone-internet usage. The researcher intends to design a T&I-UADM as a decision support tool for the adminstrators, which use ML algorithms to predict ZNA’s telephone-internet usage traffic data and detect telephone-internet usage anomalise in the ZNA communication networks.

**• Conceptual Framework**

The block diagram in figure 1.1 outlines the process of designing an effective T&I-UADM that predicts telephone-internet usage pattens, billing errors, or potential security breaches in communication networks. The process includes data collection, preprocessing, feature selection, model selection, model training, anomaly detection, threshold definition, alert generation and reporting and model evaluation and iteration.

Data collection involves collecting relevant data from organizational communication networks, while preprocessing involves cleaning and transforming the collected data. Feature selection helps reduce noise and dimensionality in the data, while model selection involves choosing an appropriate anomaly detection model based on the data's characteristics and objectives. Model training involves fitting the model to the training data and optimizing its parameters. Anomaly detection involves applying the trained model to new data, defining thresholds and generating alerts or notifications. Regular evaluation and iteration are necessary to address any limitations or false positives-negatives.

**Figure 1.1:** Conceptual Framework Source: Author

**• 1.5 Research Objectives**

The primary objective of this research is to:

To design a model that uses supervised machine learning to predict telephone-internet usage pattens.

The sub-objectives of the research are:

• To identify the appropriate usage plan among the ZNA users

• To clasify three distinct telephone-internet usage limits.

• To detect telephone-internet usage anomalies.

**• 1.6 Research Questions**

• What is the most appropriate machine learning classification algorithm that is required to predict telephone-internet usage?

• What are the most common reasons for telephone-internet anomalies?

• How will the model classify ZNA telephone-internet usage-bill limits?

• What metrics can be used to assess the model's efficiency?

**• 1.7 Significance of the Study**

Several organizations involved in the telecommunications sector are going to benefit from this research's scientific knowledge and framework for predicting telephone and internet usage. Those who are taking part are;

**• 1.7.1 The Research**

The study will improve the researcher's knowledge, qualifications and understanding in the field of machine learning, which could result in collaborations, funding and publications.

**• 1.7.2 ZNA**

The ZNA will gain a comprehensive understanding of T&I-UADM's enhanced communication security, identifying unique ZNA requirements and challenges, thereby optimizing resources and reducing costs within the military organization.

**• 1.7.3 ZNDU**

Research in communication technologies, cyber security and machine learning-based prediction model, will improve academic products and services at ZNDU, inspire innovative questions and promote collaboration in advancing knowledge.

**1.8 Delimitations of the Study/Scope of the Study**

This research will mainly focus on designing a T&I-UADM to be used by ZNA particularly at Josiah Magama Tongogara (JMT) Barracks, Harare, Zimbabwe. The respondents and participants of the research are telephone and internet service providers (Tel\*One), ZNA telephone and internet End-users and Administration (Zimbabwe Signals Corp). The research will be carried out from 29 January 2024 to 05 July 2024. The main variables are the implementation use of machine learning and big data analytics to develop the T&I-UADM. To develop the prototype, the researcher intends to use Python Jupyter Notebook. In this study, Logistic Regression, Decision Treee Classifyer, Gaussian NB, Random Forest Classifier and Gradient Boosting Classifier are the algorithms used. Since these algorithms have been applied to anomaly identification in the past, the researcher is using them in this research to develop a model that will identify anomalies in telephone-internet usage in a way that is both efficient and accurate.

**1.9 Limitations of the Study**

The researcher will expect to face some limitation during the research process, the most expected limitations are related to their possible solutions as indicated in the table below:

**Table: 1.1** challenges and possible solutions

**Limitation**

**Solution**

• Time constraint

Clearly defining the research objectives, priorities and key activities.

• Limited Resources

Conducting a thorough resource assessment to find opportunities for collaboration with external organizations or institutions.

• Complexities and Resistance to Change

Conducting a comprehensive analysis of the organizational structure and culture, identify potential barriers and develop change management strategies.

Engaging key decision-makers and stakeholders early on, promoting understanding and addressing concerns will help support and facilitate a smoother implementation process.

• Technical challenges

Conducting a thorough technology assessment, identify potential scenarios and develop appropriate mitigation strategies.

Collaboration with experienced technical partners or consultants will provide guidance and expertise in addressing technical challenges effectively.

• Ethical and Legal Matters

Establishing robust data privacy and security measures, obtaining appropriate consent and engaging legal experts can help mitigate ethical and legal concerns.

Addressing these limitations through careful planning, stakeholder engagement, resource optimization and methodological rigor, the researcher will enhance quality and impact of the research, contributing to the successful development and implementation of the T&I-UADM for the ZNA.

**1.10 Document structure**

There are five chapters in this research work. The detailed chapter breakdown is provided below:

**Chapter 1**: Introduction- provides the overview of the project.

**Chapter 2:** The second chapter pays attention to the literature review, which identifies, evaluates and summarizes the body of research on a number of T&I-UADM-related subjects, such as anomaly detection, billing accuracy, patterns of usage and network security.

**Chapter 3:** The third chapter is entitled Methodology. The process follows the Cross Industry Standard Process for Data Mining (CRISP-DM) model. It goes through phases which include business understanding, data comprehension, data preparation, modeling, assessment and implementation.

**Chapter 4:** Data presentation and analysis- This chapter will cover the findings from the analysis and show how the data is represented using graphs and charts.

**Chapter 5:** Conclusions and recommendations - the chapter will present the summary of findings of the research, the conclusions which are drawn from the findings and the recommendations based on the findings.

**1.11 Chapter Summary**

The chapter is aimed at enlightening the T&I-UADM by outlining the background information on the research area. The researcher revealed the statement of the problem from which he drew the research objectives. The research questions were raised to guide him in digging deep into literature review of the research. The significance of the study is outlined to identify the beneficiaries, stakeholders and relate them to the benefits and outcomes of the research. Finally, the researcher outlines limitations and delimitations of the research thus setting the boundaries of carrying out the research.

**CHAPTER TWO: LITERATURE REVIEW**

**•**

**•**

**•**

**• Introduction**

According to Snyder (2019), a literature review is an essential component of any research study. It involves carefully reading, selecting and combining a collection of recent literature in order to have a solid understanding of the research problem. A review of the literature in full will help the researcher build on existing knowledge, discover areas that need further research and establish a strong foundation for his own research work. In particular, the chapter will explore the theoretical overview, the concepts of supervised machine learning and the methods used to construct the framework for the in-depth supervised machine learning that classify telephone-internet usage limits into distinct categories, providing adminstrators with decision support infomation. The researcher will analyse several T&I-UADMs to identify possible research gaps or opportunity for further research. The researcher will depend on the study of finished research by other academics.

**2.2 Theoretical framework**

A large number of organizations are worried about telephone-internet budget projections. Improper accounting will result in poor resource distribution and increased bills, which impose financial strain on to organization. Through predicting and detecting excessive usage, organizations can take proactive measures to optimize resource allocation, identify cost-saving opportunities and prevent unnecessary expenses. This research thoroughly explores two well-known telephone-internet abuse hypotheses, the lack of awareness and insufficient supervision theories.

**• Lack of Awareness Theory**

A philosophyer named Halio made a comment about the awareness study in 1962. According to Halio, suggested that telephone-internet abuse may occur due to lack of awareness and education among individuals regarding the consequences and policies associated with telephone-internet usage. Ehret *et al.*, (2020) revealed that, lack of awareness hypothesis proposes that individuals may engage in abusive behaviors unintentionally because they are not fully aware of the rules, limitations and potential negative outcomes related to their usage.

When employees are not informed, they can not fully appreciate how their activities affect their organizations or the potential costs involved. The organization's detailed rules and regulations that governs telephone-internet usage, such as limitations on international calls, unnecessary downloads and having access to restricted content may not be known to them.

The theory emphasizes the significance of training and guidance for proper usage procedures. Those who lack a formal education might not possess the knowledge and skills required to make responsible decisions about how they engage with the telephone-internet. The possible risks relating to abusive conduct, optimum resource usage, or recommended procedures may be unclear to them.

Moigne, (2023) cites Halio's foundational study, which states that Unintentional abuse can occur when individuals engage in activities that exceed their allocated usage limits without realizing the consequences. For example, they may unknowingly download large files or use streaming services excessively, resulting in increased costs and network congestion. Without understanding the impact of their actions, individuals may not recognize their responsibility to use the services responsibly and efficiently.

**• Insufficient Supervision Theory**

The theory holds that insufficient monitoring and supervision of telephone-internet usage results in abuse. Conforming to this theory, users may act abusively without bothering about being detected or facing penalties when there is minimal supervision and control. According to, Mikhailova, Savina and Savin, (2022) it is difficult to effectively analyze and evaluate telephone-internet usage trends if there is no any reliable prediction tools in place. It is simpler for individuals to engage in abusive acts without being detected when there is no enough supervision. According to, Lin, Gen and White, (2022) If policies are in place, inadequate supervision may result in inconsistent or weak enforcement. If usage policies and guidelines are not constantly enforced, users may believe that there are no penalties for abusive behavior, which could lead to a rise in exploitation incidents. As another example, Dastres and Soori, (2021) revealed that abuse can be discouraged by routine auditing of telephone-internet usage. However without regular audits, organizations may not recognize and respond to cases of excessive or unauthorized use, which would let abusive behaviors continue.

According to Sagiroglu and Kilinc, (2021), uncontrolled telephone-internet use might result in a lack of personal accountability. Users may act more abusively because they see a lower danger of being held accountable when they do not feel that their actions are being observed or even examined. These pressures are the same as those expressed by Thomas and Fischer, (2023) in the lack of awareness hypothesis. The ideology component indicates that encouraging accountability and discouraging abusive behavior can be achieved through educating users about the control measures in place and the significance of responsible usage. Shirgave *et al.*, (2019) ephsised that, awareness of the potential implications of misuse and the advantages of adhering to usage regulations can be increased through educational efforts and public relations campaigns. Adopting exhaustive awareness and supervisory techniques that record and examine telephone-internet usage information can assist in detecting anomalous trends and possible abuse cases, allowing for swift action and remedial measures.

According to Lin and Hu, (2020), the theory of insufficient supervision illustrates the need for organizations to set precise regulations and rules around telephone-internet usage, as well as regular enforcement methods. This involves defining implications for violating usage specifications, making sure users are aware of them and putting in place suitable penalties. Finding gaps, excessive use and possible misuse can be aided by constantly reviewing telephone-internet usage. Organizations can move proactively to quickly resolve problems through applying the knowledge about patterns and trends that audits can reveal.

**• Telephone-internet Excessive Usage**

According to Price, (2018), excessive use of telephone-internet services beyond what is considered appropriate is referred to as telephone-internet misuse, more particularly, excessive usage. It is going above and beyond what is reasonable, performing activities that hinder productivity or violate regulations, like utilizing resources in a way that is undesirable to both individuals and organizations. When individuals or users ignore established policies and guidelines about ethical usage, it is known as telephone-internet misuse. Examples of this include accessing websites or content that are prohibited, exploiting business resources for private purposes and participating in activities that are specifically prohibited.

One of the top five T&I-UADM experts, Kount, estimates that 40% of all telephone-internet abuse involves excessive usage and that the loss from telephone-internet misuse worldwide in the month of August 2022 was $5.55 billion. The goal of any technique used for telephone-internet anomaly detection is to increase detection rates while lowering false alarm rates. Several methods, including the Logistic Regression, Decision Treee Classifyer, Gaussian NB, Random Forest Classifier and Gradient Boosting Classifier are used to identify anomalies. Statistical excessive usage detection methods fall into two broad categories, supervised and unsupervised. Supervised anomaly detection methods create models based on samples of unlawful and legal telephone-internet usage in order to classify future pattens as anomalies or ilegitimate In unsupervised anomaly detection, outliers or irregular usage are identified as probable instances of misuse. Both of these anomaly detection methods forecast the likelihood of excessive usage in various communication networks.

**2.2.4 Impact of Telephone-internet Usage Prediction and Anomaly Detection**

T&I-UADM involves finding patterns in telephone-internet usage-bill data that do not follow the expected behavior, as cited by Ioannidou, (2021). The patterns that do not follow the expected behavior in telephone-internet usage-bill data, are often referred to as anomalies but has also more reference and could also be called outliers, surprises, discordant or contaminants in different application domains. The most used terms for anomaly detection are anomalies and outliers. Anomaly detection in telephone-internet usage-bills serves multiple important purposes, including ensuring billing accuracy, detecting fraud, reducing costs, improving customer satisfaction, enhancing service quality, ensuring regulatory compliance, strengthening network security, support budget projections and resource allocation decisions.

**2.2.5 Factors that affect Telephone-internet Abuse**

A comprehensive strategy including awareness-raising, education, policies and enforcement is required to address telephone-internet misuse. One way to reduce the circumstances that lead to abuse is to encourage responsible usage, provide advise on security and internet proper conduct and create an appropriate and encouraging community on the internet. Setting up strong regulations, standards and enforcement procedures can also stop abusive behavior and keep individuals secure. The researcher reviews various factors that affect misuse.

**2.2.5.1 Accessibility**

Mitnick and Inge, (2022), indicated that individuals find it easier to do abusive acts since telephone-internet services are readily available and accessible. Telephone-internet services have become increasingly prevalent globally, resulting in enhanced communication, information access and financial opportunities. However, the availability of internet connectivity along with convenient access to computers, cellphones and tablets may lead to a rise in potential for misuse because of their anonymity and wide audience reach, exploitation. Emotional distress, reputational damage and psychological distress can all result from online harassment. To deter and lessen this kind of activity, governments, groups and internet service providers are putting laws, user education programs, awareness campaigns and usage restrictions guidelines into place.

**2.2.5.2 Lack of Awareness and Education**

Insufficient knowledge or awareness about responsible telephone-internet usage can contribute to abuse. Without understanding the potential risks, individuals may engage in inappropriate or harmful behaviors, such as cyberbullying, online harassment, or sharing sensitive information.

**2.2.5.3 Lack of Regulation and Enforcement**

Inadequate regulation and enforcement of policies related to telephone-internet usage can contribute to abuse. Without proper guidelines, individuals may engage in inappropriate activities without fear of consequences, leading to misuse and harm.

**2.2.5.4 Technological Advancements**

Rapid advancements in technology, including the introduction of new communication platforms and features, can create new opportunities for abuse. Emerging technologies may present challenges in terms of regulation, moderation and user protection.

**2.3 Discoveries Obtained from Telephone-internet Usage-bill data**

Utilizing data analytics technologies and knowledge, organizations can derive significant facts and guide their decision-making over telephone-internet services It is critical to identify and extract the precise data as well as the analytical methods used from telphone-internet usage-bill data pattens that will produce quality, detailed and in-depth analysis tool to provide useful and informative prediction conclusions to assit organizations’ budgets and optmal distribution of their limited telephone-internet resources. The researcher is going to explore….

**2.3.1 Usage Patterns**

Usage-bill data can reveal patterns in telephone-internet usage, such as peak usage hours, days of the week with higher usage, or seasonal fluctuations. This information can be useful for optimizing resource allocation, capacity planning, and network management.

**2.3.2 Usage Distribution**

Analyzing usage-bill data can help identify how telephone-internet resources are distributed among different user groups or departments within an organization. This can provide insights into potential areas of overutilization or underutilization, allowing for resource optimization and cost control.

**2.3.3 Excessive Usage**

Usage-bill data can flag instances of excessive usage that may result in additional charges or strain network resources. Identifying users or accounts with consistently high usage can prompt further investigation and intervention to address potential misuse or inefficiencies.

**2.3.4 Cost Analysis**

Analyzing usage-bill data, organizations can gain insights into cost drivers related to telephone-internet services. This includes identifying the most significant cost components, understanding usage patterns that contribute to higher bills and identifying opportunities for cost optimization and negotiation with service providers.

**• Compliance Monitoring**

Telephone-internet usage-bill data can be used to monitor compliance with usage policies, regulatory requirements and service agreements. Through data analysis, organizations can identify instances of policy violations, unauthorized usage, or non-compliance with regulatory guidelines.

**• Service Quality**

Usage-bill data can provide insights into service quality metrics, such as network performance, downtime incidents, or service-level agreement (SLA) compliance. Monitoring these metrics, organizations can identify areas for improvement, address service issues promptly and ensure a satisfactory user experience.

**• User Behavior Analysis**

Analyzing telephone-internet usage-bill data can help understand user behavior and preferences. It can reveal trends in the types of applications or services accessed, the duration and frequency of usage and the impact of different devices or platforms on usage patterns.

**• Predictive Analytics**

Using historical usage-bill data, predictive analytics techniques can forecast future usage trends, cost projections, or network capacity requirements. These insights can assist in proactive planning, budgeting and resource allocation.

**• Telephone-internet Usage**

Using the phone and internet for a variety of reasons, such as communication, information access, entertainment and engaging in online activities, is referred to as telephone-internet usage. It includes the use of both data and voice communication via internet-based networks and ordinary telephone networks.

**2.4.1 Voice Calls**

Traditional telephone services allow individuals to make and receive voice calls to communicate with others locally, nationally and internationally. This includes both landline and mobile phone calls.

**2.4.2 Messaging**

Telephone services often include messaging capabilities, such as SMS (Short Message Service) or MMS (Multimedia Messaging Service), enabling users to send text messages, pictures, or videos to other phone numbers.

**2.4.3 Internet Usage**

Internet usage involves accessing websites and browsing web pages using web browsers. Users can search for information, read news, shop online, access social media platforms and interact with various online services.

**2.4.4 Email**

Internet services enable users to send and receive emails, providing a convenient means of communication for personal and professional purposes. Emails can include text, attachments and multimedia content.

**2.4.5 Social Media**

Social media platforms allow users to connect, share content and communicate with others. Users can post updates, photos and videos, interact with friends or followers, join communities and participate in discussions.

**2.4.6 Online Communication**

Internet-based communication tools, such as instant messaging applications, voice over IP (VoIP) services and video conferencing platforms, enable real-time communication with individuals or groups over the internet.

**2.4.7 Streaming and Entertainment**

Internet usage includes streaming audio and video content, such as music, movies, TV shows and online gaming. Users can access streaming platforms, watch videos on platforms like YouTube, or listen to music through online services.

**2.4.8 Online Transactions**

Internet usage involves conducting various online transactions, such as online banking, e-commerce purchases, bill payments and money transfers. Users can securely perform financial transactions using internet-based services.

**2.4.9 Cloud Services**

Internet usage extends to cloud-based services, such as file storage, document collaboration and data backup. Users can store and access their files, documents and other data remotely through cloud service providers.

**• Call Duration and Internet Usage Package Limits**

The term "call duration and internet usage package limits" describes the restrictions placed by telecommunications service providers and organozational adminstrators on the length of telephone conversations and the total quantity of data that can be used in a given period of time. The service provider and adminstrators, allots the user's particular package, organisational role and the location may affect these limitations. Users should be informed of restrictions and call durations linked to their individual service plans, excessive use may lead to further fees, slower internet, or other limitations. Users must get comprehensive data from their service providers and administrators who define and supervise their usage, including details about the restrictions, contracts and any related repercussions.

**• Call Duration**

Voice calls have a maximum time that certain telephone service packages may enforce. Calls could be restricted to not more than 10 minutes each, for instance, under certain organizational packages. There can be additional costs if the call lasts longer than expected.

**• Monthly or Daily Call Limits**

The amount of time that can be spent on voice calls over the course of a day or a month is limited by the service providers and organizational policies. These caps may change significantly throughout packages and groups.

**• Data Caps**

Through organizational policies, internet service providers and network administrators frequently place data caps, or limits, on the quantity of data that can be consumed within a certain billing cycle. A plan might, for instance, have a monthly data cap of 100 GB. The internet connection speed may be lowered or extra charges may be incurred if the user  exceeds this limit.

**• Limiting Speed**

In certain instances, service providers may choose to reduce internet speed after an established data threshold is met rather than enforcing stringent data limitations. This means that the internet speed may be considerably lowered for the balance of the payment cycle after using a specific amount of data.

**• Time-Based Limitations**

Certain internet packages could have time-based limitations, including unrestricted access during off-peak hours or restricted access during certain hours. Usually, the purpose of these limitations is to effectively distribute resources and control network congestion.

**• Day of the Week and Month Telephone-internet usage**

Despite the day of the week or month, using telephones-internet services has facilitated accessibility and convenience for communication, information access and other activities on the internet. It completely altered the way individuals work, spend money, interact with others and communicate throughout the year.

**• Time of the day**

It is significant to keep in mind that many factors, such user behavior, geographic location and network capacity, can affect how the time of day specifically affects telephone-internet usage. Furthermore, other tactics may be used by service providers and network adminstrators to control peak usage times and guarantee satisfaction with clients throughout the day.

**2.6 Machine Learning (Identification of Patterns-Relationships)**

ML is a subcategory in AI, Chairman (2020). when an algorithm is constantly updated by a computer through experience, typically by a particular training procedure. Statistical analysis is used by a Machine Learning system to predict a result from a set of input data. After analyzing the input data, the algorithm will either predict an action or produce a class, which is a representation of what it considers the input data should represent, as indicated by Chairman, (2020). The algorithm makes use of an underlying mathematical model to learn and build its experience. This is done through training, where the algorithm makes changes to the mathematical model with the experience it gets from the training process. Below are the four basic categories of ML.

Figure 2.7: Four ML categories Source: Alzahrani, (2023)

**2.6.1 Supervised Learning Telephone-internet Usage Prediction and Anomaly Detection**

According to Ines and Ferreira, (2020), Supervised learning techniques can be applied to anomaly detection and telephone-internet consumption prediction employment. The process of preparing data include gathering historical data on the use of the telephone and the internet, including details about the type of communication, duration and time. After feature selection, the data is divided into testing and training sets. The properties of the data and the needs of the task determine which supervised learning algorithm is best. The training set, learning patterns and correlations between the input characteristics and the target variable are used to train the model. The model's performance is assessed using performance metrics such as accuracy, Mean Absolute Error, Root Mean Squared Error and Mean Squared Error. After training, the model is applied to fresh, untainted data to generate predictions and insights regarding usage in the future. Cerda-Alabern, Iuhasz and Gemmi, (2023). Gathering historical telephone-internet usage logs, both typical and unusual is a necessary step in anomaly detection. Metrics such as Precision, Recall, F1-score, or Area Under the Receiver Operating Characteristic (ROC) Curve are used to assess the model's performance. To maintain the model's accuracy and dependability over time, regular monitoring, model changes and feedback loops are required. Various aspects, including specific prediction goal, dataset features, interpretability requirements, computational performance, and the requirement to manage non-linear relationships, determine the choice of algorithm. Testing a variety of algorithms is advised so as to determine which one most effectively meets the demands of the current telephone-internet usage prediction problem. The algorithms' performances should be compared. Keeping in remember that any supervised learning algorithm has its advantages and can be applied to the prediction of phone and internet usage. The researcher is going to explore the following: Random Forest, Gradient Boosting, Decision Trees, Gaussian NB, Support Vector Machines (SVM), and Logistic Regression.

**2.6.2 Random Forest**

An algorithm called Random Forest uses ensemble learning to produce predictions that are more reliable and accurate, Carrillo-Mondejar, Martinez and Suarez-Tangil, (2022). Clustering is a way to separate unlabeled data set into a natural hidden data structure, instead of trying to find the accurate characterization of the new data set that has been provided to this method Raman, Narayanan and Velmurugan, (2020). Multiple decision trees' predictions are combined using the "wisdom of the crowd" idea. Random subsampling increases model robustness and decreases overfitting by introducing variation into the training process, Ioannidou, (2021). Decision trees are separately trained on bootstrapped data via recursive binary splitting, random feature selection increases variation among trees. The class with the majority vote is chosen for classification tasks and the average prediction is chosen for regression tasks, after the final prediction is combined. Large datasets, categorical and numerical features, high-dimensional data and feature relevance estimates can all be handled by Random Forest. It finds the ideal ratio of variance to bias through hyperparameter adjustment, which optimizes performance. Among its many uses are image processing, marketing, finance and healthcare.

**2.6.3 Gradient Boosting**

Another ensemble technique is called gradient boosting, which merges weak predictive models (such decision trees) one after the other in a sequential fashion with the goal of fixing the flaws of the preceding models. Gradient Boosting techniques, such as XGBoost, AdaBoost, or LightGBM, are capable of handling missing data, accurately predicting telephone and internet usage, and capturing complex usage patterns. They are renowned for their capacity to manage big datasets, deal with non-linear relationships, and carry out automated feature selection.Jurišić, Tomicic and Grd, (2023).

**2.6.4 Decision Trees**

Models that are easy to understand and apply for both regression and classification applications are decision trees. Decision trees can be used to find important features and develop rules for predicting patterns of telephone and internet usage. Decision Trees offer an easy-to-understand hierarchical framework for use prediction interpretation by repeatedly dividing the data according to distinct feature thresholds. If proper regularization techniques are not used, they may not generalize well to new data and may even suffer from overfitting..

**2.6.5 Gaussian NB**

A probabilistic classifier based on Gaussian distribution and feature independence is called Gaussian Naive Bayes. Gaussian NB can be used to classify usage behavior based on the probability of several usage groups or predict usage patterns in telephone-internet usage prediction. It can effectively handle big datasets and is especially helpful when working with continuous features. It does, however, operate under the assumption that the traits are independent, which might not always hold true in practical situations.

**2.6.6 Support Vector Machines (SVM)**

Support vector machines (SVM) is an effective supervised training technique that can be applied to regression and classification problems alike, Pinto and Sobreiro, (2022). The algorithm can be utilized to predict consumption patterns, categorize usage behavior, or estimate usage volumes in relation to predicting telephone-internet usage. Finding an area that effectively divides several groups or forecasts usage volume is the objective the algorithm of. SVM use of kernel functions, it can manage non-linear relationships and is particularly useful when handling high-dimensional data. Raman, Narayanan and Velmurugan, (2020).

**2.6.7 Logistic Regression**

For conditions requiring binary classification, a frequent statistical model is logistic regression. When attempting to predict binary outcomes, such as whether or not a user will exceed a specific usage threshold, in telephone and internet usage prediction, logistic regression can be utilized. Logistic Regression offers information about the likelihood of particular usage behaviors by calculating the probabilities of various classes based on input features. It is capable of handling both continuous and categorical features, is interpretable, and is computationally efficient., Bjornerud, (2021). (Sai, 2022).

**2.6.8 Hyperparameter Optimization**

Hyperparameters in machine learning models must be set in order for the model to be customized for the dataset. Although the overall effects of hyperparameters on a model are frequently understood, it can be challenging to choose the right hyperparameter or set of interacting hyperparameters for a particular dataset. There are frequently broad heuristics or rules of thumb for configuring hyperparameters. A more effective method would be to objectively compare various values for the model hyperparameters and choose the subset that produces the model that performs the best on the given dataset. The Python machine learning toolkit scikit-learn supports this technique, which is also referred to as hyperparameter optimization or hyperparameter tuning. A single set of high-performing hyperparameters is generated by a hyperparameter optimization and can be used to configure your model. There are many optimization algorithms that can be used, but random search and grid search are the most common and straightforward. A search space is defined by Random Search as a bounded domain of randomly sampled hyperparameter values. Every point in a search space is regarded by Grid Search as a grid of hyperparameter values.

**2.7 Application of Ensemble Techniques On The Best Algorithms**

By ensemble numerous models rather than relying just on one, ensemble approaches seek to increase the accuracy of findings in models. The integrated models considerably improve the results' accuracy. As a result, ensemble approaches for machine learning have gained prominence. Bagging, boosting, stacking, voting classifiers, averaging and blending are some of the combination techniques used in data mining. Both the bagging and boosting techniques aggregate the output of distinct models of the same kind using voting (the same algorithm is used in constructing the models). To create a prediction model, boosting is an iterative technique that uses weighted instances that concentrate on a certain group of instances. Stacking is distinct from both boosting and bagging because it mixes the results of various algorithmic kinds to provide a final prediction. (Austen, 2024)., (Bernard, 2022).

Every learning system operates by adjusting to its environment. Learning algorithms produce predictions based on their own set of presumptions when given instances that have never been encountered. Inductive bias is the term used to describe these presumptions (Mitchell 1980). The use of numerous algorithms allows for the exploration of various search spaces and the possibility of obtaining results that are possibly diverse because different algorithms have various representations and search heuristics. Combinations of learning algorithms can be used to evaluate complex databases because no one method performs optimally on all types of datasets.

**2.7.1 Max Voting / Voting Classifier**

For categorization issues, the max voting approach is typically utilized. With this method, predictions are made for each data point using a variety of models. Each model's predictions are regarded as a "vote." The majority of the models' forecasts serve as the basis for the final projection. A voting classifier is a machine learning model that gathers training data from a large ensemble of models and forecasts an output (class) based on the class with the highest likelihood of being the output. To predict the output class based on the highest majority of votes, it merely averages the results of each classifier that was passed into the voting classifier. The concept is to build a single model that learns from these models and predicts output based on their aggregate majority of voting for each output class, rather than building separate dedicated models and determining the accuracy for each of them.

**• Averaging**

Each data point is subject to multiple predictions when averaging. In this approach, the final prediction is made by averaging the results of all the models. When computing probabilities for classification problems or making predictions in regression problems, averaging can be applied. Using multiple models to be trained on the same dataset and then combining their predictions is the simplest technique to create a model averaging ensemble in Keras. To illustrate a model averaging ensemble, we will use a straightforward multi-class classification problem as the starting point.

A multi-class classification issue with the required number of samples, input variables, classes, and variance of samples within a class can be created using the make blobs () function, which is included in the scikit-learn class. With input variables to represent the x and y coordinates of the points and a standard deviation of 2.0 for points within each group, we use this issue with 500 samples. To guarantee that we always receive 500 points, we shall employ the same random state.

**• Blending**

Similar to stacking, blending makes predictions using only a holdout (validation) set from the train set. In contrast to stacking, the forecasts are based solely on the holdout set. A model is created using the holdout set and the forecasts and tested on the test set.

**• Bagging**

Combining bootstrapping and aggregating is referred to as bagging. By resampling data from the training set with the same cardinality as the original set, a technique known as bootstrapping can assist reduce the classifier's variance and overfitting. Compared to a single individual model, the produced model ought to be less over fitted. It is undesirable for a model to have a high variance because this indicates that the model's success depends on the training set. Therefore, the model can still perform poorly even if extra training data are provided. Additionally, it might not even lessen the variance of our model. When you just have a small amount of data, bagging is a useful technique. By using samples, you may estimate the results by averaging the scores across several samples.

If the ensemble accuracy is significantly higher than the base models, the simplest bagging strategy is to use a few tiny subsamples and bag them; if not, use larger subsamples. It's not a given that using larger subsamples will lead to better findings. The accuracy of the base model and the benefit from bagging are trade-offs in bagging. As you have an unstable model, aggregation from bagging may significantly improve the ensemble; however, when your base models are more stable and have been trained on larger subsamples with higher accuracy, advantages from bagging diminish. A weighted average is then used to get the final score after the bagging is complete and all the models have been built using (mainly) distinct sets of data.

**• Boosting**

The fundamental goal of boosting is to successively add new models to the overall ensemble model. Previously, when bagging, we averaged every single model that was produced. This time, a new model is generated with each iteration of boosting, and the new base-learner model is trained (updated) using the mistakes made by the prior learners. An overall prediction is obtained by adding the result from each weak model that the program builds. Ensemble modelling from before is being used here. Similar to how gradient descent goes toward the correct values, the now-boosted gradient adjusts the current prediction, nudging it toward the true target. The different models' outputs, rather than their individual parameters, are subject to gradient descent optimization. Boosting algorithms can be optimized using a variety of techniques. Contrary to the bagging examples above, the generation of the subset in classical boosting is not random, and performance is based on the success of earlier models. Since each new subset that is iterated upon contains components that earlier models might have mistakenly categorised. Additionally, to ensemble the models together, we will continue to use hard voting as we did before.

**• Stacking**

Stacking is an ensemble learning method that creates a new model by combining predictions from other models (such as decision trees, KNNs, or SVMs). On the test set, predictions are made using this model. Stacking, often referred to as stacked generalization, is an ensemble technique that employs a meta-classifier or meta-regressor to merge many classification or regression models. The meta-model is trained on the features that are outputs of the base-level models after the base-level models have been trained on the entire training set. Because the base-level frequently consists of various learning methods, stacking ensembles are frequently diverse.

Since the research is on credit card detection models the ensemble technique used is Stacking, using the stacking classifier. The steps involved in the ensemble stacking are shown below,

Below is a step-wise explanation for a simple stacked ensemble:

**• First:** The train set is divided into 10 pieces.

**• Step 2:** Predictions are made for the 10th part after a base model (let's say a decision tree) is fitted to the first nine parts. Each piece of the train set receives this treatment.

**• Step 3:** The entire train dataset is then fitted with the basic model (in this case, a decision tree).

**• Step 4:** Predictions are performed on the test set using this model.

**• Step 5:** Repeating steps two through four with a different base model (let's say knn) will produce a different set of predictions for the train set and test set.

**• Step 6:** A new model is constructed using the predictions from the train set as features.

**• Step 7**: Final predictions are made using this model on the test prediction set

**• Telephone-internet Usage Prediction and Anomaly Detection Algorithm Implementation**

The following procedures can be used to use machine learning algorithms for anomaly identification and telephone-internet usage prediction:

**• Data preparation:** Compile necessary details and labels about telephone-internet activity.

**• Feature engineering:** Carry out any required preprocessing.

**• Spliting Dataset:** Create training and testing sets from the dataset.

**• Model evaluation and training:** Utilizing the Python scikit-learn module, train a (Random Forest, Gradient Boosting, Decision Trees, Gaussian NB, Support Vector Machines (SVM) and Logistic Regression) models on the training set.

**• Performance comparison:** Using evaluation metrics, compare each algorithm's performance.

**• Algorithm Selection:** Based on performance comparison, select the best algorithm.

**• Hyperparameter Tunning:** Use sophisticated approaches such as ensemble methods or optimization algorithms, or change hyperparameters and feature selection to fine-tune and optimize the algorithm.

These procedures enables application and assessment of several algorithms for anomaly detection and telephone-internet consumption prediction.

**• Related studies**

There have been numerous studies conducted on telephone-internet usage abuse . To identify previously published works, we shall summarize some of the articles in this review. This section covered machine learning using (supervised approaches) like Random Forest, Gradient Boosting, Decision Trees, Gaussian NB, Support Vector Machines (SVM) and Logistic Regression. Researchers like Cerdà-Alabern, Iuhasz and Gemmi, (2023), Cerda-Alabern, Iuhasz and Gemmi, (2023), Jurisic, Tomicic and Grd, (2023), have identified supervised machine learning method as the most common method.

**• In Cerda-Alabern, Iuhasz and Gemmi, (2023)**

Investigated and evaluated different unsupervised learning algorithms to see which of them has the better detection rate and lowest false-positive rate. The tested algorithms are K-Means, SOM, DAGMM, and Adversarial Learned Anomaly Detection (ALAD). They set two different benchmark data sets for the network flow-based anomaly detection, while they test different parameters and neural network settings for each UL algorithm. The results they found during their work was that the DAGMM algorithm gains the lowest false-positive rate and a high result for anomaly detection rate, it worked better on one of the data sets compared to the other. Where the SOM algorithm gained the best result on the other data set.

When they compared the anomaly detection rate for each algorithm, they found out that K-means works better than the other algorithms when the unseen attacks have no instances in the training set. When the attacks have a few instances on the training set, the algorithm ALAD worked better because it uses an adversarial sample generation mechanism. They concluded that the algorithms they used had stable performances and were easy to implement and the ALAD algorithm showed excellent results in the detection of minority attacks. Then they also concluded that to be able to achieve better results they should integrate the algorithms for excessive telephone-internet usage-bill anomaly detection.

**• In Cerdà-Alabern, Iuhasz and Gemmi, (2023)**

Swartling, and P. Hanna, investigated DBSCAN and the LOF algorithms for identifying anomalies on unlabeled data, they focused on the damage identification for industries in their production process. They observed the pump-generated data during execution and used high frequency sampled current and voltage time series data for identification. The collected data was split up into five different phases, the startup phase, three duty point phases and the shutdown phase. They concluded that the DBSCAN and LOF algorithm identifies unexpected data points within their data set and they state that their problem was around the validation of the points. Their validation method had two parts to it, first, they needed to investigate if the data point that has been found was an outlier or not compared to the rest of the data in the set, the second part to the validation was if the outliers were indeed an anomalous behavior. Their study resulted in that they found out when increasing the number of dimensions then the number of outliers will increase rapidly for both DBSCAN and LOF. With this result, they stated that the algorithms are less confident when the outlier number increases, because the graph curve is getting flatter and is visible for the duty points. They concluded the methods they used could identify unexpected data points within their set.

**• Jurisic, Tomicic and Grd, (2023)**

B. Georgescu, I. Shimshoni, and P. Meer investigate how to reduce the computational complexity of adaptive mean shift. They have investigated the most popular techniques for clustering and listed what kind of limitations they hold. The first clustering method they investigated was the k-means algorithm and they have stated the limitation the technique beholds. The limitation they have mentioned is, k-means need to know the number of clusters before execution, and the clusters are constrained to be spherically symmetric. The second method they investigated was nonparametric clustering methods that are based on mean shift, and they stated that the limitations of this method will be eliminated, and when the dimension of space increases then the amount of computation becomes prohibitively large. They want to find a way to reduce the computation complexity and during the time of the report, there was a recently proposed approximation technique that could help with the reduction of the computation complexity. The method they used was locality-sensitive hashing (LSH), but they have changed the implementation of the method and states that with their way of implementing the method they find the optimal parameters of the data structure and will be determined by a pilot learning procedure, with the help of data-driven partitions. Their conclusion to this study was that with the help of a data structure that was based on a locality-sensitive hashing method they obtained a significant decrease in the run time and at the same time they maintained the quality of the result. The results they got during this study for k-means and the adaptive mean shift algorithm was that the k-means algorithm could classify 97.32% and the adaptive mean shift (AMS) could classify 98.66% for the Brodatz database.

**2.9.4 Jurisic, Tomicic and Grd, (2023)**

In the event of inconsistent communication, it is important for a military organization to estimate how much phone and internet usage it will require as well as the usage patterns of its clients. There has been interest in studying if machine learning techniques could offer improved usage assessment due to recent developments in the sector. For a long time, this kind of usage patterns has been determined using statistical methods. The purpose of this thesis is to determine the optimal machine learning method for telephone-internet prediction, based on a set of model evaluation standards. The techniques examined included K-Means Clustering, GMM, Hierarchical Clustering, DBSCAN, and SOM. SMOTE oversampling was used to address the disparity between classes for the response variable. The findings demonstrated that, in terms of the selected model assessment metric, K-Means Clustering without SMOTE implementation produced the best results.

**2.10 Comparison of results from related work**

Table 2.1: Comparison of related work

**Researcher**

**Work carried out**

**Recommendations**

Abrahamsson and Granstrom, 2019

The best default prediction performance for selected model assessment criteria was identified among Logistic Regression, Random Forest, Decision Tree, AdaBoost, XGBoost, Artificial Neural Network, and Support Vector Machine. SMOTE oversampling was used to address the disparity between classes for the response variable.

The potential future work for this project will involve developing the model further by conducting a more in-depth analysis of the variables used in the models and by developing new variables to improve predictions.

Zhu et al., 2019

The development of a loan default prediction model based on the Random Forest algorithm used real-world customer loan data from Lending Club. The issue of imbalanced classes in the dataset is addressed using the SMOTE method.

Future studies should focus on conducting trials on larger data sets or modifying the model to achieve cutting-edge performance.

Kisutsa, 2021

Analyzed how credit worthiness scores and limits are determined by mobile lending platforms using predictive analytics, including transaction history, call logs, text messages, contact lists, age, education, and income. This study investigates the use of machine learning algorithms to increase the precision of loan default prediction. Mobile lending companies will use this model to assess the credit risk of their customers. Decision trees, the model that performs the best in the research, have an accuracy rate of about 64%.

Different techniques for parameter tuning and feature selection can be used to increase performance.

In order to compare the model's performance with other open dataset sources as they become more accessible, it may also be useful to perform cross validation.

On the basis of the borrower's characteristics, loan characteristics, and loan recollection processes, further investigation may be made in estimating the expected return of the loan.

Das, 2022

The author conducted research on and discussed a number of machine learning models that have been applied to the issue of lending money to individuals. Research explored Logistic Regression as one of its methods, but it also employs Decision Trees, Support Vector Machines, and a variety of other classification-based algorithmic models to perform a predictive analysis based on the relationship between two factors known as Recall and Precision to determine the F1 score.

Use more evaluation metrics to measure performances of the models used for predictive analysis

Tariq et al., 2019

The research focused on using data mining techniques for the prediction and classification of loan defaults. In this work, KDD, CRISP-DM, and SEMMA techniques were employed. During the experimentation phase, three different data mining approaches were used for the suggested model and their effectiveness was evaluated according to a number of criteria. Based on these qualities, the best technique was chosen, described, and recommended due to its major characteristics in predicting loan defaults in the financial industry.

Use of data mining techniques which include Neural Networks, Support Vector Machine, Linear Regression, Random Forest, Decision Tree, Logistic Regression, Fuzzy Logic, Genetic Programming, Discriminant Analysis, Bayesian Networks, Hybrid Methods and Ensemble Methods.

**2.11 Conceptual framework**

What the researcher hopes to learn from the investigation is depicted in a conceptual framework. It outlines potential relationships between them and establishes the pertinent factors for the researcher's investigation. The three-phase architecture of the proposed framework for T&I-UADM. The architecture consists of the (X) input, the prediction models and a Excessive usage detection phase(Y) output. The (X) input are the variables of the dataset which are used to predict excessive telephone-internet usage and normal telephone-internet usage. The prediction models are the algorithms which process the dataset to come up with a prediction or detection. The (Y) is the predicted output that is either normal or excessive telephone-internet usage.

**Conceptual Frame of a Telephone-Anomaly Detection Model**

**INPUT (X)**

1. **Call Logs:** (caller ID, recipient ID, call duration, timestamp, and call type).

2. **Internet Usage Records:** (browsing history, website visits, data usage, timestamps of internet sessions, and the amount of data transferred).

3. **User Information**: (user IDs, account information, demographics, subscription plans, and user behavior metrics).

4. **Geographic Information:** (geographic location of the user/device during a call or internet session).

**5. Device Information:** (device IDs, device types, operating systems, and device-specific metrics).

6. **Time-related Features:** (Timestamps of call and internet activities, day of the week, time of day, and seasonal variations).

7. **Aggregated Metrics:** (average call duration, total data usage per user, call frequency, or browsing behavior metrics).

**OUTPUT (Y)**

• Normal telephone-internet Usage

• Excessive Telephone-internet Usage (**Anomaly**)

**Prediction Models**

• Random Forest

• Gradient Boosting

• Decision Trees

• Gaussian NB

• SVM

• Logistic Regression

**Figure 2.1:** Conceptual Framework

**2.12 Chapter Summary**

The chapter highlighted the theoretical framework, explaining the concepts of telephone-internet abuse detection, the impact of telephone-internet abuse detection. The mitigation methods currently being used to deal with the problems and their shortcomings were discussed. Machine learning, a branch of artificial intelligence was introduced and supervised machine learning algorithms were highlighted focusing more on the algorithms that can be employed for classification. The chapter was concluded with a look at some of the related researches that have conducted in the detection of telephone-internet abuse using machine learning. The coming chapter will deal with the methodology to be employed in carrying out the research.

**CHAPTER 3: RESEARCH METHODOLOGY**

**3.1 Introduction**

Research design or methodology refers to the overall plan or framework which guides the research process, as defined by Martinez, Viles and G. Olaizola, (2021). It encompasses the overall strategy and techniques employed to gather, analyze and interpret data in order to answer the research objectives or test hypotheses. In other words, it defines the flow of activities executed by the researcher in carrying out the research process. Sigorta et al., (2022) revealed that a well-defined research methodology gives researchers a road map for gathering authentic and reliable data, guaranteeing study accuracy and production of significant results. In this section the researcher decides to adopt to a combination of quantitative and qualitative approaches, known as a mixed methods research methodology, it is the optimal methodology which allows the researcher to be involved in taking decisions regarding a research design of a T&I-UADM, data collection strategies, sample tactics, data processing protocols and ethical issues.

**3.2 Research philosophy**

The design of a T&I-UADM is a problem that requires a scientific approach in coming up with a comprehensive solution, therefore the optimal research philosophy to adopt would be a combination of positivism and interpretivism, known as a pragmatic research philosophy. According to Sigorta et al, (2022), pragmatism is a research philosophy that allows researchers to make decisions regarding various research-related problems while considering both the objective and subjective aspects of the study. Pragmatism allows for flexibility in research design, enabling researchers to integrate both quantitative and qualitative elements. The philosophy places emphasis on addressing practical problems and finding solutions. It encourages researchers to engage with real-world challenges, collaborate with stakeholders and develop actionable recommendations based on research findings.

**3.3 Research approach**

According to Chen, Sharma and Munoz, (2023) a research approach refers to the specific methods and techniques employed by researchers to conduct their study within a chosen research paradigm. For this research the optimal research approach taken by the researcher was a sequential exploratory research approach. Njera, (no date), The sequential exploratory research approach is a two-phase approach that begins with qualitative data collection and analysis followed by quantitative data collection and analysis. This approach allows researchers to make decisions of the research plan, including research design, data collection strategies, sample tactics, data processing protocols and ethical issues, effectively. The sequential exploratory approach involves an initial qualitative phase followed by a quantitative phase. In the qualitative phase, researchers conduct interviews, surveys and focus groups to explore users' experiences and perceptions to telephone-internet usage anomalies. This qualitative exploration helps inform the subsequent quantitative phase, where statistical models and algorithms for anomaly detection can be developed. This approach ensures a comprehensive understanding of the research problem and enhances the validity of the findings. In the quantitative phase, usage data from telecommunication service providers, billing records, call logs and internet traffic logs were collected for further analysis.

**3.4 Research design**

Snyder, (2019), defined a research design as a referal to the overall structure and plan of action that guides the entire research process, including the selection of research methods, data collection strategies, sample tactics, data processing protocols and ethical considerations. Explanatory research design, descriptive design, experimental design and correlational design are some of the designs available. The experimental research design was used in this study, which is a scientific research approach, it involves investigating how one or more independent factors affect a dependent feature, one or more dependent variables are modified and applied to one or more independent variables. In order to calculate the statistical significance of the observed effects, experimental designs frequently involve the gathering and analysis of quantitative data for making deductions, this offers solid and neutral evidence. The situational dataset's features were employed as independent variables, impacting the target feature's reliability. Through adjustments to the input space and model parameters during training, it enables the researcher to modify the qualities of the study. This gave the researcher total control over the model's quality and applicability for the research. Through adjustments to the input space and model parameters during training, the researcher created classification models that are able to predict telephone-internet usage at reasonable accuracies. Different models were evaluated using different and 2 matrices were chosen for hyper tuning to try to improve models performance. The researcher then selected one final model and evaluated the model, compared it with a sample model to see how well the model performed comparing to a baseline model and also determined the important predicting features for the model.

**3.5 Data collection methods**

Research methods refer to the specific techniques, procedures and tools used by researchers to collect, analyze and interpret data within their chosen research paradigm and approach. In this research, the optimal research methods included a combination of both qualitative and quantitative methods. Predicting telphone-internet usage was done through the sequential research approach. Data was gathered by researchers using a variety of techniques, such as user logs, mobile apps, web analytics, call detail records (CDRs), surveys, interviews and sample datasets.

User logs offer real-time data, surveys gather self-reported information about telephone-internet usage. Mobile apps gather usage data, collaborating with providers enables access to anonymous data, though they might not record telephone usage. Web analytics tools might offer insights into user activity, although they need access agreements with providers. CDRs offer precise data on telephone usage, but they require a lot of resources, field studies enable the collecting of data in real time. To concentrate on collection of quantitative data the researcher downloaded a sample datasets with necessary matching characteristics and features that correspond with study objectives of telephone-internet usage patten prediction from Kaggel.com.

**3.6 Data Collection Tools-Instruments**

Data collection tools can be used individually or in combination to gather comprehensive data related to telephone-internet usage predictions. It is essential to define the scope of data collection, consider privacy implications and adhere to ethical guidelines throughout the study. Table 3.1 below contains a list of the study's data collection resources.

**Table 3.1:** Research materials and tools

**Tool**

**Purpose**

**Source**

Anaconda

Application software

 https://docs.anaconda.com/

Python 3.7.3

Programming language

http://www.dartgo.org/pyii

Jupyter notebook

Programming environment

https://jupyter-notebook.readthedocs.io/

CSV

Dataset

Kaggel,

• Random Forest

• Gradient Boosting

• Decision Trees

• Gaussian NB

• Support Vector Machines (SVM)

• Logistic Regression.

Classification algorithms

https://scikit-learn.org/

**3.7 Population of the study**

Reliability of the demographic definition to the specific research issues being addressed and the study aims is an important point to remember as noted by Kaushik and Walsh, (2019). A telephone-internet usage prediction model requires identifying the population of entities that use telecommunication services, such as telephone internet and charging them for use. Included in the list of important elements are the sample approach, demographic parameters, service providers, geographic scope, user categories and accessibility and availability. Inferences or conclusions from ZNA telephone-internet usage behaviors are the objectives of the research. Three main platforms includes mobile phones, ADSL and optical fiber were the focus of the research, each with its own class. Aside from demographic variables like gender, age, occupation, financial level and educational background, the study also took accessibility and availability of communications services into account. Taking into account the objectives and available resources of the research, the sampling plan was selected by methods such stratified, cluster, systematic, convenience, purposive and basic random sampling.

**3.9 Sources of data**

Customer behavior and billing data can be gathered using the telephone-internet usage prediction system. Internet and telecommunications service providers, billing systems, usage tracking apps, mobile applications, surveys, questionnaires and Internet of Things (IoT) gadgets are some examples of data sources.

For this particular research the researcher downloaded a dataset from kaggel. The dataset contained 7043 Rows and 21 Columns of customers usage pattens within the last month the column is called Chun services that each customer has signed up for phone, multiple lines, internet, online security, online backup, device protection, tech support and streaming TV and movies. Customer account information, on how long they have been a customer, contract, payment method, paperless billing, monthly charges and total charges Demographic info about customer’s gender, age range and if they have partners and dependents

**3.10 Validity and Reliability**

Many datasets are available through open-source platforms that maintain ethical norms, such as GitHub, Data.gov, Kaggle, UCI Machine Learning Repository, Open ML, Google Dataset Search and Data.gov. These platforms ensure that data is properly obtained, anonymized and complies with privacy and regulatory regulations. The researcher thoroughly examined the documentation and licensing details linked to the dataset.

**3.11 Data Collection Procedures**

To collect data for qualitative and quantitative analysis the researcher employed Several strategies. Primary data collection involved directly obtaining data from original sources, such as surveys, interviews, or experiments. Primary and secondary data sources are combined in a mixed-methods approach to provide a thorough analysis of the research question. Whereas secondary data collection the researcher made use of already present data from government agencies, research groups, or publicly accessible databases.

**3.12 Data Analysis Framework**

Data analysis refers to the process of inspecting, cleaning, transforming and modeling data to discover meaningful patterns, extract insights and support decision-making. The Cross Industry Standard Process for Data Mining (CRISP-DM) was employed to conduct data analysis in this research. According to Alliance, (2021), CRISP-DM is a widely used structured approach to guide the entire data analysis process through exploring the six stages, as shown in Figure 3.1, these steps are business understanding, data preparation, data understanding cleaning, modelling, evaluation and deployment, the framework used allows for a more business-oriented approach and better results in the shortest amount of time.

Figure 3.1: CRISP-DM source:

**3.12.1 Business Understanding:**

The phase involves fully understanding the business objectives, goals and requirements for ZNA communication networks. Priority lies on resource allocation, service quality assurance, network security and operational efficiency to assure dependable services, Technology seek to increase security and manage user access to network resources. ZNA networks should be developed to provide quick and easy access to resources that are allowed while offering a smooth user experience. They require integration with current technological advancement and would want to accommodate a variety of network environments since they are adaptable and scalable.

**3.12.1.1 Define the business objectives**

The model's objectives are to discover abnormalities in telephone-internet usage data, maintain network security, detect over consumption or billing problems, assist with cost optimization and comply with relevant rules and regulations. In order to preserve moral and legal standards, it also aids in the enforcement of organizational policy, standards and regulation compliance . It also helps with discrepancies in billing and cost optimization.

**3.12.1.2 T&IUADM Requirements**

The model construction process involves collecting and integrating data from various sources, preprocessing it to handle missing values and extracting relevant features to capture meaningful usage patterns. Suitable for patten prediction, anomaly detection algorithms and techniques should be selected based on the data's characteristics and objectives. The model should be trained using labeled data and evaluated using metrics like precision, recall, F1-score, or AUC. Real-time monitoring is required for prompt detection and alerting of anomalies. Comprehensive reports and visualizations should be generated to provide actionable insights for decision-making. The model should be scalable to handle large data volumes and adapt to changing usage patterns.

**3.12.2 Data Understanding**

Gathering the data required to meet the project's input resource requirements is the second step of the model. This stage of data collecting lays the basis for understanding the data. Comprehending data also drives the search for datasets to gather and examine so as to achieve the goals of the project.

**3.12.2.1 Data Preperatiom**

For the purpose of gathering information, identify patterns and develop hypotheses, the researcher perfomed exploratory data analysis (EDA) processes. EDA was useful in formulation of new varriables that create new predictions, working with unstructured or qualitative data and exploring new occurrences. Typical approaches included reviews of relavent literature, data exploration and descriptive statistics, exploratory factor analysis, content analysis, qualitative coding and thematic analysis, EDA and grounded theory.

**3.12.2.2 Data collection**

The researcher searched for a dataset from https://www.kaggle.com/datasets/reyhanarighy/data-telco-customer-churn. The dataset was saved on the local machine, the file contains data on the customer loss rate of a telecom provider, which is an important business metric for client retention. Dependents, duration, internet service, device protection, online security, online backup, contracts, paperless billing, monthly fees and churn are among the variables that are taken into consideration. To help with user connectivity and satisfaction decisions, the dataset was utilized for feature engineering, user classification, predictive modeling, customer churn research and business strategy. Below is the dataset in csv fomat.

**Figure 3.1:** Dataset in Csv Fomat

**3.12.2.3 Importing Libraries**

To access pre-defined functions, classes and modules in Python, it is typical to import libraries. There are three ways to import a libraries which includes using an alias, directly importing certain functions or classes and importing the library's whole contents. Due to potential nomenclature conflicts and code maintenance concerns, the third strategy is generally not advised. Use package managers such as pip or conda to install required libraries. Data manipulation, numerical calculations, charting and machine learning techniques are just a few of the features that can be accessed via importing libraries

**Figure 3.3** Importing Libraries

**3.12.2.4 Data Extraxction**

Extracting data from databases requires utilizing the right techniques to communicate with the databases. This may involve developing SQL queries to retrieve relevant data from views or tables. Utilizing programming languages and frameworks, read and process data saved in files (such as CSV or Excel). Depending on the dataset format, the researcher used different libraries in Python to load the dataset. Pandas is frequently used to load data from SQL databases, Excel, JSON and CSV files. Use the read\_csv() function, then the read\_excel() function with the file location and sheet name to load data from CSV. The read\_json() function is used for JSON. Use a library such as sqlite3 or pymysql to establish a connection for SQL databases. Depending on the installed libraries and dataset format, select the appropriate approach.

**Figure 3.3: Loading the Dataset**

**3.12.3 Data Exploration Analysis (EDA)**

Data quality evaluation, data classification, pattern recognition, single variable relationship analysis, computational descriptive statistics, graphic design and dimensionality reduction are all part of the critical data analysis phase referred to as "data exploration." Model building, data preprocessing, hypothesis testing and research design are all facilitated with this process.

**3.12.3.1 (Examining the Data in Python)**

**data.head()**: The python command explore the first five rolls of the dataset

**Figure 3.4:** Show the dataset's initial few rows using

**3.12.3.2 data.tail()** Excution of the command outputs the dataset, including column names and data types, then display the last five rows of the dataset

**Figure 3.5:** Get details about the last five raws

**3.12.3.3 data.info().form:** Determine the dataset's dimensions (number of rows, number of columns)

**Figure 3.6:**

**• 3.12.4.1 Data Cleaning and Preparation**

Evaluating the dependability, correctness, consistency, completeness and relevance of the data that has been gathered or retrieved is an essential phase in the preparation of data. The following are important factors:

• data validity and integrity

• documentation and metadata

• data governance and compliance

• accuracy

• completeness

• consistency

• relevance and timeliness

Checking for inconsistencies, errors, or outliers involves comparing accurate data with a reliable source. Every needed field and variable should be verified for completeness and any missing information should be noted. Relevance should be evaluated for alignment with project objectives and consistency between variables or fields should be verified. It is important to evaluate timeliness to make sure the data is current and satisfies analytical needs. To guarantee ethical principles and compliance with laws and regulations, data governance procedures should be evaluated. Ensuring the integrity and reliability of outcomes requires regular inspection of the data.

Essential steps in data preprocessing include cleaning and data preparation, which includes handling outliers and missing values, normalizing or standardizing data and converting variables for analysis. Handling missing values, handling outliers, standardizing or normalizing, encoding categorical variables, managing text or categorical data and manipulating dates and times are examples of common Python procedures.

**• 3.12.4.2 Handling missing values**

Handling missing values is essential in development of a telephone-internet usage prediction model with supervised machine learning methods. Python library functions such as isnull() and isna() are used to find missing values, they determine underlying patterns and dependencies by analyzing the missingness pattern. Computation methods encompasses model-based computation, hot-deck computation and mean/median/mode computation.

**Figure 3.1:** Count the number of missing values in each column

**• 3.12.4.3 Handling outliers**

A non-parametric method called kernel density estimation (KDE) was used to calculate the probability density function of a random variable. Duration analysis, or how long a person has held an activity or status, is one its uses. KDE, gather data, examines it and create a KDE plot. The kernel function's width, which affects the ratio of underfitting to overfitting, is controlled by the bandwidth parameter. Displaying core patterns, modes and possible outliers, the KDE plot sheds light on the duration value distribution. Overlaying different KDE plots allows for comparative examination, emphasizing variations in duration patterns. To better comprehend the distribution and features of data, KDE was used with additional statistical methods like box plots or summary statistics.

**• 3.12.4.4 Standardizing or Normalizing**

Numerical data can be transformed to a common scale using two preprocessing techniques: normalization and standardization. Transforming data to have a mean of 0 and a standard deviation of 1, standardization also referred to as Z-score normalization ensures that every feature is given equal weight. It works well for optimizing models that rely on feature scaling or for distributions that resemble Gaussian distributions. Normalization, often called Min-Max scaling, on the other hand, preserves relationships and proportions between values by transforming data to a predefined range, usually between 0 and 1. For Gaussian-like distributions or to make sure all features have a similar scale, standardization is appropriate. The decision between normalization and standardization is based on the chosen techniques, data properties, and analytic requirements. The researcherto tried out the two methods to determine how they influence model performance.

**• 3.12.4.5 Encoding Categorical Variables**

Frequently, machine learning algorithms convert categorical variables into numerical values in order to accurately train and forecast models. Typical methods encompass hash encoding, binary encoding, ordinal encoding, label encoding and one-hot encoding. Using one-hot encoding, new dummy variables are created for every category, converting them into binary columns. The algorithm comprehended the ordinal relationship between categories since label encoding gave each category a distinct number value. More freedom in defining the order is provided by ordinal encoding, which assigns number values based on the chosen order of categories. With binary encoding, every category was represented as a binary code that is first converted to its corresponding numerical value and then to its binary representation. Hashing encoding reduces the dimensionality of data by using a hash function to translate categories into numerical values.

**• 3.12.4.6 Manipulating Dates and Times**

It is usual in data analysis to manipulate dates and times, which can involve parsing, formatting, component extraction, arithmetic computations and time zone conversions. Among the methods are time zone conversion, handling Daylight Saving Time (DST), parsing and structuring strings into a structured format, extracting particular components and doing arithmetic computations. Timezone conversions, component extraction, arithmetic computations, parsing and formatting are all handled by built-in functions in programming languages and libraries. Use of reputable libraries and stay up to date on any upgrades or modifications to DST regulations is important because time zone conversion libraries frequently automatically adapt for DST transitions.

**3.12.4 Modeling**

To model telephone-internet prediction, the researcher initially defined the prediction objective, then preprocessed the data, utilized feature engineering, chose an appropriate model, trainned and fine-tuned it, assessed and validated it and finally deployed and continues to monitor it. Metrics including F1-score, mean squared error, accuracy, precision and recall were used to assess the model's performance. The model's applicability in handling the prediction task is evaluated. Subsequently, the model's performance is implemented in a real-time prediction environment. The model's performance is constantly observed and it is updated or retrained as necessary. The modeling strategy may change according to the particular task, the available data and prediction context. It is necessary that the modeling process be iterated, experimented with and refined in response to input and knowledge obtained during the development cycle. (recommentation)

**3.12.4.7 Feature Extraction**

To extract relevant patterns and information from data for telephone-internet prediction, features must be extracted. Domain knowledge, prediction goal and accessible data all play a role in feature selection, when examining data. User behavior, call and internet usage, user demographics, user profiles, and temporal aspects are all important components. Predicting usage requires knowing the duration, quantity, average, peak hours and frequency of calls. Trends in data usage, popular apps, peak internet usage hours and overall data usage were all considered aspects of internet usage. Application usage patterns, roaming status, call duration distribution and time intervals between calls are examples of user behavior. Factors such as age, gender, plan type, day of the week and location were considered. Days of the month, holidays and special occasions are examples of temporal aspects.

**3.12.5 Evaluation**

Several standards and methodologies are available for evaluating telephone-internet prediction models. Metrics including accuracy, precision, recall, F1-score, AUC-ROC, and regression evaluation were utilized for classification jobs. The mean absolute error (MAE), mean squared error (MSE), root mean square error (RMSE), and R-squared (R2) were utilized in regression projects. To evaluate the model's generalization performance and reduction of overfitting. The researcher employed the cross-validation technique. Learning curves revealed how the performance of the model changed against the function of the size of the training set, making it easier to determine whether the model is underfitting or overfitting and whether more data might be beneficial

**3.12.6 Deployment**

A number of steps such as testing and validation, continuous integration and deployment (CI/CD), infrastructure setup, API development, scalability and performance considerations, data preprocessing, monitoring and logging, security and privacy were involved in the process. The was packaged in a deployable style and offered with an API for consistent communication. Performance optimization, appropriate data pretreatment and handling the anticipated demand should all be possible with this approach. To keep track of the model's behavior and performance, logging and monitoring systems were put in place. To safeguard the model and its data, security precautions were also considered. The deployment procedure was automated by creating a continuous integration and deployment pipeline. To guarantee the model's scalability, security and dependability in a production setting, regular upgrades and maintenance are required and will be carried out at regular intervals.

**3.12.7 Monitoring and Maintenance**

A telephone-internet prediction model's continuous accuracy, dependability and performance depend on regular maintenance and monitoring. Performance tracking, data drift detection, model retraining and updates, error and issue monitoring, security and privacy auditing, user feedback and validation, version control and rollback, knowledge sharing and documentation, routine maintenance and updates and continuous improvement are some of the important components.

Error tracking techniques, data drift detection, retraining or updating the model and finding degradations or anomalies in key performance indicators (KPIs) are all part of performance monitoring. Compliance with privacy laws and industry standards is ensured through security and privacy auditing. Assessing the model's performance and resolving issues requires user validation and input. Mechanisms for version control and rollback are used to monitor modifications and guarantee reproducibility. Thorough documentation makes knowledge exchange easier.

**3.12.10 Ethical Consideration**

Achieving justice requires ongoing interaction and seeking out alternative perspectives it is important to carefully address ethical issues when developing and implementing telephone-internet prediction models. A few of these are responsibility, accountability, openness, consent from the user, fairness-bias and user privacy protection. The researcher prioritized data protection and reduced bias and unfairness. Users are allowed to keep control of their data and forecasts, as well as transparency and clarity. Besides monitoring the model's outputs, accountability and responsibility was observed. There should be multidisciplinary teams involved in the governance and ethical reviews. Assessing the social impact is necessary to find any unintentional biases or adverse effects. All throughout the model's lifecycle, regular reviews must consider changing stakeholder expectations, regulatory frameworks and society norms into consideration. (reccommentation)

**3.12.11 Chapter Summary**

This chapter presents various elements of the research methodology that facilitates the collection of data to address the research objectives. The main elements highlighted in the chapter included research design, Data collection methods (Experiments and simulations), data collection instruments, validity and reliability issues on methods, instruments and tools used, data collection procedures and data analysis framework. The methodology's analysis of the data and conclusions will be covered in the next chapter.

**CHAPTER 4: RESULTS PRESENTATION AND ANALYSIS**

**•**

**•**

**• Introduction**

The researcher illustrates an effective telephone-internet usage anomaly detection model that predicts usage patterns and detects anomalies based on historical data. Usage levels are determined among two prepaid plans, Surf and Ultimate. The model analyses historical patterns, day of the week and time of day to provide visualizations presented through intuitive and comprehensive reports. These visual representations includes charts, graphs and dashboards that depict usage trends, predicted usage levels and detected anomalies. Resource planning and distribution is made possible by the identification of seasonal variations and recurring pattens that may impact the projected demand. Also, early detection and preventive steps are made simple by the model's ability to detect anomalies, such as illegal activity or network disruptions. Decision-makers can benefit from the prediction model's results, which are supplied in the form of machine learning techniques, throughout this research project.

**4.2**

**•**

**• Exploratory data analysis results**

The first objective is to explore insights from telephone-internet usage which is the process of exploratory data analysis of the available data frames. Exploratory data analysis is the crucial process of doing preliminary analysis on data in order to find patterns, identify anomalies, and test hypotheses with the aid of summary statistics and graphical representations. The insights deducted from the research are as follows.

**• Descriptive statistics of the dataset**

The information is spread across five files: "megaline\_calls.csv," "megaline\_internet.csv," "megaline\_messages.csv," "megaline\_plans.csv," and "megaline\_users.csv." The objective of this task is to identify cost-effective prepaid plans, allocate resources in accordance with policy categories, as well as analyze users' routines of telephone-internet usage. ZNA network administrators will use this information to initiate awareness campaigns against users who abuse the resources unknowingly.

**Figure 4.1:** Descriptive Statistics

**• Correlations**

**Figure 4.2:** Columns

The relationship between features is shown by the correlation figure 4.3 below.

**Figure 4.3:** Correlations

The correlations of the features in the data shows that there are some features which are strongly co-related for example the relationship between InternetService and InternetService\_Fiber optic. Such features cause noise in the model due to multi-collinearities. As such the removing of other features among the correlated was done to clean the data of the noise resulting from such. Amount and the class column has a correlation of 0.21 although amount has a high influence on determining excessive usage which shows their relationship is not direct but also influenced by other factors also.

**• Transaction amount**

When making a withdrawal transaction, a telephone-internet user specifies this sum of money. In this specific investigation. The variances in the total number of transactions represent the distinctive usage trends of each user. This relies on a number of variables, including the user's monthly limit, the types of activities they indulge in, and their financial obligations. A user might, for instance, have the following spending habits: ($200, $120, $80, $180, $110).However, at the fraud transactions they will be following a different trend of amount being withdrawn which is the anomaly in the transactions. The bar graph below shows the transaction distribution in dataset of the distribution of the frequencies of normal usage against those of misuse.

**Figure 4.3:** Transaction distribution

The result shown by the graph in figure 4.4 below, while normal transactions tend to be around $200 or less, we see fraudulent transactions peak around $300 and then at the $800-$1000 range. There is a very clear pattern here that higher amounts are misuse.

**Figure 4.4:** Amount vs fraud

**• Gender in relation to misuse**

Second, we will examine whether one gender is more susceptible to misuse than the other since the transactions which were recorded included both males and females.

**Figure 4.5:** Gender vs misuse

From the figure 4.5 above do not see a clear difference between both genders. Data seem to suggest that females and males are almost equally susceptible (50%) to telephone-internet misuse. Gender is not very indicative of abuse.

**• Spending Category in relation to Fraud**

Third, we examine in which spending categories fraud happens most predominantly. To do this, we first calculate the distribution in normal transactions and then the distribution in abuse activities. The difference between the 2 distributions will demonstrate which category is most susceptible to abuse. For example, if 'grocery\_pos' accounts for 50% of the total in normal transactions and 50% in fraudulent transactions, this does not mean that it is a major category for misuse, it simply means it is just a popular spending category in general. However, if the percentage is 10% in normal but 30% in misuse, then we know that there is a pattern.

**Figure 4.6:** Spending Category vs fraud

From the above graph figure 16, more often iSome spending categories indeed see more misuse than others! abuse tends to be happening 'Shopping net', 'Grocery\_pos', and 'misc\_net' while 'home' and 'kids pets' among others tend to see more normal telephone-internet usage than misuse ones.

**• Age in relation to fraud**

**Figure 4.7:** Age vs Fraud

The age distribution is visibly different between 2 transaction types. In normal transactions, there are 2 peaks at the age of 37-38 and 49-50, while in fraudulent transactions, the age distribution is a little smoother and the second peak does include a wider age group from 50-65. This does suggest that older people are potentially more prone misuse.

**• Cyclicality of Telephone-internet Abuse**

**a. Hourly Trend**

A very sharp contrast, while normal transactions distribute more or less equally throughout the day, fraudulent payments happen disproportionately around midnight when most people are asleep. To a certain extent the hour at which transactions are done especially at midnight are credit card fraud transactions. The distributions are shown in the figure 4.8 below.

**Figure 4.8:** Time in a day Vs fraud

**b. Weekly Trend**

**Figure 4.9:** Day of the week Vs fraud

Normal transactions tend to happen more often on Monday and Sunday while abusive ones tend to spread out more evenly throughout the week.

**c. Monthly Trend**

Figure 4.10: Month Vs fraud

Very interesting results, while normal payments peak around December (Christmas), and then late spring to early summer, fraudulent transactions are more concentrated in Jan-May. There is a clear seasonal trend.

**• State in relation to Fraud**

Now that we have examined fraud on the temporal level, let us also explore which geographies are more prone to Misuse. We will use the same methodology, where we calculate the difference in geographical distribution between the 2 transaction types.

**Figure 4.11:** State Vs fraud

As can be seen, NY and OH among others have a higher percentage of fraudulent transactions than normal ones, while TX and MT are the opposite. However, it should be pointed out that the percentage differences in those states are not very significant but a correlation does exist.

**• Train Supervised \Machine Learning Algorithms to Predict Telephone-internet usage and Detect Anomalies**

This research uses the suggested strategies to identify telephone-internet misuse. Different machine learning algorithms, including Logistic Regression, Decision Trees, Random Forest, Naive Bias, and SVM, are compared in order to ascertain which algorithm best fits three models that can be combined by ZNA telephone-internet adminstrators to identify excessive resource consuptions in their communication networks.

**• Decision tree**

The simplest machine learning techniques are decision trees. Because they are a perfectly clear way of grouping observations that resemble a tree-like arrangement of if-then sentences. The decision tree training's outcomes are as follows:

**Figure 4.12**: Decision Tree

**• Logistic regression**

It is a binary classification where the logistic function-modified linear combination of two or more input variables is used to calculate the conditional probability of one of the two potential perceptions of the response variable. The performance's outcomes are depicted below 4.13.

**Figure 4.13:** Logistic Regression

**• Random forest**

The Random Forest is a collection of a very large number of different decision trees. Each separate tree predicts a class.

**Figure 4.14:** Random Forest

**• XGboost**

Extreme Gradient Boosting is referred to as XGboost. The gradient boosted decision tree is implemented by an ensemble method algorithm that is built for high momentum and exceptional performance.

**Figure 4.15:** XGboost

**• Naïve Bayes**

Naive Bayes is a statistical technique that uses Bayes' theorem to calculate the likelihood that a characteristic belongs to a particular class. The performance of the Naive Bayes method is seen in the findings below.

**Figure 4.16:** Gaussian Naïve Bayes

**• Evaluation of the algorithms**

**a. Accuracy**

The evaluation of the algorithms is the process of comparing the performances of the algorithms to come up with a ranking of their performance. The first element to be evaluated was the use of accuracy on their performance as shown by the graph below, figure 4.17.

**Figure 4.17:** Accuracy in percentages in training

**Figure 4.18:** Accuracy percentages in testing

**b. Classification report**

Table 4.1: Table of the Classification report

**Model**

**precision**

**recall**

**F1 score**

**ROC**

SMV

0

0.92

0.94

0.93

0.92

1

0.94

0.92

0.93

Random Forest

0

0.87

0.95

0.91

0.90

1

0.94

0.86

0.90

Decision Tree

0

0.87

0.88

0.87

0.87

1

0.87

0.87

0.87

Logistic Regression

0

0.87

0.88

0.87

0.84

1

0.93

0.75

0.83

Gaussian Naive Bayes

0

0.69

0.85

0.76

0.73

1

0.81

0.62

0.70

The oversampling strategy, which is a technique that copies new or occasionally simulates cases in the minority class, was used to get the results of the models. It increases the instances, which improves the model's training. The model performance is shown to be excellent across the board with the oversampling dataset, and we can see how the trees' algorithms outperformed others, particularly the random forest algorithm, which displays an AUC score of 0.90%, and the XGboost algorithm, which displays an AUC score of 0.92%.

We also observed that the Scores of other algorithms had improved, indicating that the majority of algorithms perform well with oversampling datasets. The AUC measure demonstrates that XGboost is the superior algorithm when compared to the other metrics. Accuracy has a good recall and precision rate, therefore we can also rely on its prediction outcomes for telephone-internet excessive consumption.

**• Ensemble the Best Supervised Machine Learning Algorithms to Detect Telephone-internet Miuse**

Stacking is an ensemble learning method that creates a new model by combining predictions from other models (such as decision trees, KNNs, or SVMs). On the test set, predictions are made using this model.

**• Implementing the Algorithm in Detecting Telephone-internet Misuse**

The system views the incoming telephone-internet usage pattenstransactions and amount as credit card transactions. The ensemble algorithms are fed new incoming Transactions as input. By looking at the data, spotting patterns, and using machine learning ensemble algorithms, the output will indicate whether a transaction is legitimate or fraudulent. In the event that a fraudulent transaction has occurred, the card may be blocked to prevent further financial losses to the user and the credit card company by setting off an alarm. The typical transactions include the actual transactions that are recorded and have been performed.

**• Comparative analysis**

The table below shows the results of the previous studies that have been done before the research. Some of the models such as local factor isolation, Naive Bias, KNN have a very high accuracy but very low precision which shows the ability to ‘identify the true positives and true negatives is very low.

**Table 4.2:** Comparative analysis table

Classifier

Accuracy

Precision

AUC

Local factor isolation

99.6

33%

0.54

Naïve bias

97.2

30%

0.86

KNN

97.7

36%

0.9

Logistic regression

54.86

20%

0.86

ANN

91

70%

0.9

Random Forest

98.6

86%

1

Ada boost

80

54%

0.7

The research that has been done with the use of oversampling techniques, there was the improvement of precision and AUC without affecting the accuracy of the models. For example, the logistic regression model improved to 0.93% accuracy, 0.75% precision and 0.84 AUC.

**• Chapter summary**

The research recognized that when it comes to use of labeled data for anomaly detection and telephone-internet usage pattern prediction, the XGboost Classifier is the top-performing classifier. In terms of anomaly detection and telephone-internet usage pattern prediction, the stacking ensemble technique outperformed individual models.

**CHAPTER FIVE CONCLUSION AND RECOMMENDATIONS**

**•**

**• Introduction**

This chapter summarizes the research study's results and provides recommendations derived from them, and discusses the possible effects of the results of the research for various parties involved in the telecommunications industry and for telephone-internet users widely. Results of research, observations learned, and recommendations concerning utilizing such systems in the actual world.

**• Summary of findings**

The five machine learning models used in the study were Logistic Regression, Gaussian NB, Decision Trees, Random Forest, Gradient Boosting and Support SVM. Applying methods of evaluation including accuracy, recall and F1 score, three of the top models were chosen after they were successfully trained. Use of stacking, was also evaluated and revealed good performance, the three models were ensembled.

Rapid telephone-internet pattern prediction usage and detection of anomalies were the goals of the research. The introduction of the ensemble technique aimed to improve recollection in predicting patterns of telephone and internet usage and identifying abnormalities, consequently minimizing false alarms and failures to detect anomalies in the pattens. A 0.92% recall was obtained from the stacking..

**• Conclusion**

Telephone-internet prediction and anomaly detection is becoming one of the areas that require extensive research due to the significant demand for telephone-internet services, individuals, and every organization. Any observation that does not follow the predicted distribution or pattern of the other items is referred to as an anomaly in data evaluation. This could be used to describe a negative event in the age of computers, such as a network intrusion or telepphone-internet abuse. Telephone-internet misuse can be exposed using supervised learning approaches, which can also work together to improve performance. In this study, we utilized the stacking method and checked the model's efficacy using the labels on the available data.

**• Recommendations**

• The research mainly examined supervised machine learning. However, unsupervised machine learning techniques can be implemented in subsequent research to predict telephone usage patterns and detect anomalies.

• Also, only one ensemble stacking methodology was used, alternative ensemble strategies, such as voting Classifier, can be adopted in various systems to enhance machine learning models' performance in anomaly detection and telephone-internet prediction.

• The research's datasets came from India, in order to make more deductions and enhance detection of anomalies and telephone prediction, datasets from other areas and locations must be used.

• The noise data was not included in the training and testing phases of the prototype. It is necessary to introduce noise data into the prototype's training and testing processes for the purpose of  improving the models' performance in detecting anomalies and predicting telephone-internet usage.

**• Summary**

The results of the research are brought up in this chapter and are built on the research objectives' discoveries. The effects of the analysis for various interested parties are addressed as well, along with ideas drawn from the results. The research concludes with lessons learned, how to apply the results to everyday events and proposals for further studies.

**References:**

Abualhaj, M.M. *et al.* (2024) ‘MVF : A NOVEL TECHNIQUE TO REDUCE THE VOIP PACKET PAYLOAD LENGTH’, 25(1), pp. 43–53. Available at: https://doi.org/10.2478/ttj-2024-0005.

Aouini, Z. and Pekar, A. (2022) ‘NFStream: A flexible network data analysis framework’, *Computer Networks*, 204, p. 108719. Available at: https://doi.org/10.1016/j.comnet.2021.108719.

Austen, J. (2024) ‘RNNs and LSTMs’.

Bernard, S. (2022) ‘Walden University’.

Björnerud, P. (2021) ‘Anomaly Detection in Log Files Using Machine Learning’, (February).

Carrillo-Mondéjar, J., Martinez, J.L. and Suarez-Tangil, G. (2022) ‘On how VoIP attacks foster the malicious call ecosystem’, *Computers and Security*, 119. Available at: https://doi.org/10.1016/j.cose.2022.102758.

Cerdà-Alabern, L., Iuhasz, G. and Gemmi, G. (2023) ‘Anomaly detection for fault detection in wireless community networks using machine learning’, *Computer Communications*, 202(January), pp. 191–203. Available at: https://doi.org/10.1016/j.comcom.2023.02.019.

Chairman, V. (2020) ‘Docteur de l uni ersité du lu embourg en informatique’.

Chen, S., Sharma, G. and Muñoz, P. (2023) ‘In Pursuit of Impact: From Research Questions to Problem Formulation in Entrepreneurship Research’, *Entrepreneurship: Theory and Practice*, 47(2). Available at: https://doi.org/10.1177/10422587221111736.

Dastres, R. and Soori, M. (2021) ‘A Review in Recent Development of Network Threats and Security Measures Computer Integrted Manufacturing View project Modern CAD/CAM Systems View project’, *Article in International Journal of Computer and Information Sciences* [Preprint]. Available at: https://www.researchgate.net/publication/348676738.

Ehret, S. *et al.* (2020) ‘A grounded theory on the relation of time awareness and perceived valence’, *Timing and Time Perception*, 8(3–4), pp. 316–340. Available at: https://doi.org/10.1163/22134468-bja10014.

Ioannidou, P. (2021) ‘Anomaly Detection in Computer Networks Anomaly Detection in Computer Networks’.

Jurišić, M., Tomičić, I. and Grd, P. (2023) ‘User Behavior Analysis for Detecting Compromised User Accounts : A Review Paper’, 23(3), pp. 102–113. Available at: https://doi.org/10.2478/cait-2023-0027.

Kaushik, V. and Walsh, C.A. (2019) ‘Pragmatism as a research paradigm and its implications for Social Work research’, *Social Sciences*, 8(9). Available at: https://doi.org/10.3390/socsci8090255.

Lin, H., Gen, B. and White, M. (2022) ‘The Cyber Defense Review Vol 7, No 4, Fall 2022’, 7(4).

Lin, W. and Hu, S. (2020) ‘Design and implementation of an offline face recognition locker’, *Journal of Physics: Conference Series*, 1634(1), pp. 1–66. Available at: https://doi.org/10.1088/1742-6596/1634/1/012131.

Mikhailova, E.V., Savina, V.V. and Savin, I.M. (2022) ‘Artificial Intelligence as a Tool for Optimizing Resource Schedules’, *Строительное Производство*, (2), pp. 52–56. Available at: https://doi.org/10.54950/26585340\_2022\_2\_52.

Mitnick, K. and Inge, D.W. (no date) *Telecom System Security*.

Moigne, J.J. Le (2023) ‘2023 Advanced Information Systems Technology ( AIST ) Novel Observing Strategies ( NOS ) Grouped Annual Reviews’.

Njera, M. (no date) ‘Proposed Methodology’.

Persky, D. (no date) ‘Interested in learning more ? In sti tu te ho ins ll r igh’, (Security 503).

Pinto, S.O. and Sobreiro, V.A. (2022) ‘Literature review: Anomaly detection approaches on digital business financial systems’, *Digital Business*, 2(2), p. 100038. Available at: https://doi.org/10.1016/J.DIGBUS.2022.100038.

Price, B. (2018) ‘Telephone Expense Audit’.

Raman, S., Narayanan, B. and Velmurugan, K. (2020) ‘A Big Data Analytics Literature Survey Using Machine Learning Algorithms’, *International Journal of Computer Science and Software Engineering (IJCSSE)*, 9(7), pp. 39–42.

S. Abella, D.F. *et al.* (2024) ‘Optimizing Resource Usage in Cagayan de Oro City Institutions, Philippines’, *International Journal of Social Science and Human Research*, 07(03), pp. 1747–1752. Available at: https://doi.org/10.47191/ijsshr/v7-i03-32.

Sagiroglu, S. and Kilinc, H. (2021) ‘Telecom fraud detection with big data analytics Telecom fraud detection with big data analytics Duygu Sinanç Terzi \* and Ş eref Sa ğ ı ro ğ lu Hakan K ı l ı nç’, (March 2022). Available at: https://doi.org/10.1504/IJDS.2021.121090.

Sai, K.V.N. (2022) ‘K.venkata naga sai’, (38110278), pp. 1–58.

Shirgave, S.K. *et al.* (2019) ‘A Review On Credit Card Fraud Detection Using Machine Learning’, (October).

Sigorta, A. *et al.* (2022) ‘Kocaeli Journal of Science and Engineering SOBE : A Fraud Detection Platform in Insurance Industry’, 5, pp. 25–31.

Snyder, H. (2019) ‘Literature review as a research methodology: An overview and guidelines’, *Journal of Business Research*, 104(July), pp. 333–339. Available at: https://doi.org/10.1016/j.jbusres.2019.07.039.

Syariah, K.B. and Ilmu, G. (no date) ‘No 主観的健康感を中心とした在宅高齢者における 健康関連指標に関する共分散構造分析Title’, (september 2016), pp. 1–6.

Thomas, M. and Fischer, A. (2023) ‘Visual Analytics for Communication Analysis submitted by’.

Zawra, M.A., Emam, O.E. and Shehab, M.E. (2023) ‘A Survey of Call Detail Records (CDR) Analysis’, *International Journal of Computer Applications*, 185(13). Available at: https://doi.org/10.5120/ijca2023922808.