A HYBRID MODEL FOR DETECTING INSURANCE FRAUD USING K MEANS AND SUPPORT VECTOR MACHINE ALGORITHMS

BRIAN NDIRANGU MUTHURA

J57/26173/2019

Sign_____ **Date** _____

Department of Computing and Information Technology

A research project submitted in partial fulfilment of the requirements for the award of the degree of Master of Science (Computer Science) in the School of Pure and Applied Sciences of Kenyatta University

Supervisor

Dr. Abraham Matheka Mutua

Computing and Information Technology (CIT)

Kenyatta University

Sign……..……..Date……..

October 2024

ABSTRACT

Medical insurance fraud is a significant issue in the healthcare sector, commonly characterized by fraud patterns such as misrepresentation of services, false claims, and identity theft. These patterns contribute to severe data class imbalances, with legitimate claims vastly outnumbering fraudulent ones, complicating effective detection. Current fraud detection methods struggle to address these evolving patterns and manage imbalanced datasets. This study employs a mixed-methods approach, integrating an extensive literature review with quantitative analysis of historical medical claims data. The research develops and evaluates four machine learning models: a standalone Support Vector Machine (SVM), a tuned SVM, a hybrid model combining K-Means clustering with SVM, and a tuned hybrid model. The models were compared using key metrics, including accuracy, precision, recall, and F1 score. Results show that the tuned hybrid model achieved the highest performance with an accuracy of 97.49%, demonstrating its superior ability to detect fraudulent claims compared to the standalone and default hybrid models. Future work will focus on further improving the computational efficiency of the hybrid model and exploring its adaptability to new and evolving fraud patterns in real-time environments. This research significantly advances fraud detection by offering a robust solution that tackles class imbalances and adapts to evolving fraud schemes.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABBREVIATIONS AND ACRONYMS

SVM: Support Vector Machine

ACFE: Association of Certified Fraud Examiners

GDP: Gross Domestic Product

AKI: Association of Kenya Insurers

ECP: Electronic Claim Processing

LOF: Local Outlier Factor

EHR: Electronic Health Records

CRISP-DM: Cross Industry Standard Process for Data Mining

PCA: Principal Component Analysis

EDA: Exploratory Data Analysis

CHAPTER ONE

INTRODUCTION AND BACKGROUND OF THE STUDY

1.1 Introduction

Medical insurance fraud is a critical challenge in the healthcare sector, leading to significant financial losses for insurers due to fraudulent activities such as misrepresentation of services, false claims, and identity theft. These schemes not only impact profitability but also lead to higher premiums for policyholders. Traditional fraud detection methods, such as manual reviews by experts and rule-based systems, are no longer sufficient to handle the increasing volume and sophistication of fraud cases. Moreover, the data class imbalance - where legitimate claims vastly outnumber fraudulent ones - complicates accurate detection, as machine learning models often struggle to identify rare fraud cases.

To address these gaps, this study proposes a hybrid machine learning model that integrates K-Means clustering for grouping similar claims with SVM classification to enhance fraud detection. The hybrid approach aims to improve performance in handling imbalanced datasets and evolving fraud patterns. The study compares four models: a standalone SVM, a tuned SVM, a hybrid K-Means-SVM model, and a tuned hybrid model, using accuracy, precision, recall, and F1 score as evaluation metrics.

This chapter outlines the study's context and problem statement, emphasizing the limitations of traditional fraud detection approaches and the need for more dynamic machine learning solutions. The research objectives focus on understanding fraud

patterns, evaluating current machine learning techniques, and developing and comparing the performance of the proposed models. Key research questions guide the study in identifying the effectiveness of different models in detecting fraud.

1.2 Background to the Study

Insurance fraud ranks second only to tax fraud in terms of prevalence (Association of Certified Fraud Examiners, 2019). The nature of the insurance business renders it particularly vulnerable to fraudulent activities. Insurance entities primarily mitigate risks for policyholders by pooling resources and generating substantial cash flows through insurance premiums to cover loss claims. Fraud in insurance occurs when policyholders attempt to benefit from insurers while failing to adhere to the contractual terms and conditions of the policy, resulting in damages and losses for the insurer. Such fraudulent activities can manifest at any stage throughout the policy term (Association of Certified Fraud Examiners, 2019). These losses encompass both long-term policies, such as life insurance, and short-term policies, including motor and health insurance.

The prevalence of insurance fraud transcends geographical boundaries, affecting multiple countries globally. The Association of Kenya Insurers (2021) conducted an analysis of the financial impact of insurance fraud across various nations. For instance, the Coalition Against Insurance Fraud estimates annual losses of over $80 billion due to insurance fraud in the United States of America. In the United Kingdom, the Association of British Insurance recorded 107,000 fraudulent insurance claims in 2019 valued at over £1.2 billion, reflecting a 5% increase from 2018 (Association of Kenya

Insurers, 2021). France's National Health Fund estimated losses of $321.4 million due to fraudulent schemes and claims, with healthcare providers, such as doctors and practitioners, accounting for the highest percentage of fraudulent claims at 48%. Meanwhile, health institutions like hospitals and clinics accounted for 31% of fraudulent claims, and 21% were attributed to insured individuals. In India, annual losses from fraud amount to approximately $6 billion, nearly 8.5% of the total premiums collected. The South African Insurance Crime Bureau estimated that out of $2.4 billion in insurance claims paid in 2019, $497.86 million could have been related to fraudulent claims, constituting roughly 20% of the total claims made that year (Association of Kenya Insurers, 2021). Kenya's Insurance Fraud Investigation Unit identified 83 fraud cases valued at close to KES 386.34 million (Association of Kenya Insurers, 2021). The escalating instances of insurance fraud and substantial financial implications have led to increased insurance costs. However, these figures represent only estimations of potentially fraudulent claims and may not accurately depict the true magnitude of incurred losses (Association of Kenya Insurers, 2021).

Insurance industry fraud significantly impacts a company's financial performance, with more than 5% of global company revenues lost annually to fraudulent activities (Association of Certified Fraud Examiners, 2019). The Association of Kenya Insurers (2020), in its annual review of the insurance industry's performance, noted a net loss of KES 2.33 billion in 2020 within the short-term insurance sector. Effective fraud detection mechanisms and stringent claim validations could substantially enhance profitability for insurance firms. Apart from financial losses, fraudulent schemes tarnish insurers' reputations. Insurance fraud represents a global issue affecting

economies, states, communities, and individuals (Association of Certified Fraud Examiners, 2019).

Detecting and preventing fraud remain critical concerns in the insurance industry (Matloob & Khan, 2019). Hanafy & Ming (2021) highlighted the challenge posed by the substantial number of insurance claims, surpassing the capacity of experts tasked with their analysis in real-time. Additionally, varying experiences and perspectives among experts dealing with similar claim cases contribute to decision biases (Gupta et al., 2021). Matloob and Khan (2019) observed a shift in fraud detection from traditional domain expert analysis to rule-based systems in the medical insurance industry. These rule-based systems involve conditions that evaluate claim validity, significantly increasing claim analysis throughput compared to domain expert analysis. However, there remains a need for even more efficient insurance fraud detection in the medical insurance domain (Matloob & Khan, 2019).

Researchers have proposed machine learning as a sophisticated technology to analyze claim patterns in medical insurance claims (Matloob & Khan, 2019). Rawte & Anuradha (2015) emphasized that machine learning, applied to extensive datasets, can identify unknown patterns and predict outcomes crucial for fraud detection. Machine learning techniques are classified as supervised, unsupervised, semi-supervised, and reinforced learning (Raj & Lin, 2020). In supervised machine learning, models are trained using predefined class labels, while unsupervised machine learning identifies instances with anomalous behavior using unlabeled data. Rawte & Anuradha (2015) highlighted the use of supervised learning to classify claims into predefined labels

(fraudulent and legitimate), while unsupervised machine learning algorithms primarily address outlier detection and clustering within claim datasets. They proposed a hybrid algorithm combining supervised and unsupervised algorithms for fraud detection.

Consequently, the hybrid machine learning approach can address both classification and clustering issues, identifying both known and emerging forms of fraud. Raj & Lin (2020) identified K-nearest neighbor, Naïve Bayes, and SVM as commonly used supervised algorithms for classification problems. This research aims to evaluate the effectiveness of a hybrid machine learning algorithm integrating K-Means with SVM for fraud detection within a leading insurance firm. The study seeks to provide a comparative analysis between the performance of the hybrid model and that of a standalone SVM classification model.

1.3 Problem Statement

Medical insurance fraud detection remains a critical issue due to the increasing sophistication of fraud schemes such as service misrepresentation, fictitious claims, and identity theft. Traditional detection methods, including manual reviews, rule-based systems, and isolated machine learning models, are insufficient to keep pace with these evolving fraud patterns (Gupta et al., 2021; Matloob & Khan, 2019). These conventional approaches are often inadequate in handling the scale and complexity of modern fraud, leading to a high number of undetected fraudulent cases. Compounding this challenge is the issue of data class imbalance, where legitimate claims significantly outnumber fraudulent ones, resulting in models being biased towards legitimate claims and struggling to detect the rarer fraudulent ones (Alloghani et al., 2020).

Previous research has focused on either supervised or unsupervised machine learning techniques in isolation. Supervised models like SVM often rely on balanced datasets for effective classification but perform poorly when tasked with detecting rare fraud cases in imbalanced datasets (Waghade & Karandikar, 2018). On the other hand, unsupervised models such as K-Means clustering excel in identifying anomalies but lack the precision needed for reliable classification of fraudulent claims. This gap between detection accuracy and adaptability to imbalanced data highlights the need for a more effective solution (Rawte & Anuradha, 2015).

This study addresses the gap by proposing a hybrid machine learning model that integrates K-Means clustering with SVM classification. The hybrid model aims to improve fraud detection by leveraging K-Means clustering to group claims with similar features, thereby reducing noise in the dataset. SVM is then used to classify the claims into fraudulent or legitimate categories. The key parameters being studied and enhanced by this hybrid model include accuracy, precision, recall, and F1 score, which are critical metrics for evaluating the effectiveness of fraud detection systems.

The issue of data class imbalance is addressed using a stratified sampling technique that ensures both fraudulent and non-fraudulent claims are adequately represented in the training data. This helps prevent the model from being biased towards the majority class (legitimate claims) and enhances its ability to correctly classify the minority class (fraudulent claims) (Hanafy & Ming, 2021). By ensuring that both classes are balanced during training, the model is less likely to overfit and more likely to generalize well

when deployed on real-world data, thus improving its ability to detect rare fraudulent cases.

1.3.1 Objectives

1. To investigate how medical insurance fraud occurs.
2. To investigate the current machine learning techniques used in fraud detection.
3. To develop a hybrid machine learning model to detect fraud in medical insurance claims.
4. Evaluate the performance of the hybrid model to detect fraud and compare it with a lone SVM model.

1.3.2 Research Questions

1. How does medical insurance fraud occur?
2. What are the current machine learning techniques used for fraud detection by medical insurance firms?
3. How is a hybrid machine learning model developed for detecting fraud in medical insurance claims?
4. How effective is the performance of the hybrid machine learning model in detecting medical insurance claims in comparison to a lone SVM model?

1.4 Justification

The need for a more effective fraud detection model in medical insurance is critical due to the limitations of current systems, which struggle to accurately detect fraudulent

claims. Traditional methods, such as rule-based systems and manual reviews, have proven inadequate in handling the growing complexity and volume of fraud. For example, the Association of Kenya Insurers (2021) reported that KES 386.34 million worth of fraudulent claims were identified, yet this figure only represents detected fraud, indicating the potential for a larger, undetected problem. In the U.S., the Coalition Against Insurance Fraud estimates annual losses of over $80 billion due to fraud, further highlighting the scale of the issue.

Current machine learning models, such as standalone SVM, while effective in controlled, balanced datasets, have shown significant limitations when applied to real-world, imbalanced datasets where fraudulent claims are vastly outnumbered by legitimate ones. These models are prone to misclassifying fraud cases, leading to a high number of false negatives and undetected fraud. Unsupervised models, like K-Means clustering, though useful for anomaly detection, lack the precision needed for accurate fraud classification, further underscoring the inadequacy of existing solutions.

The proposed hybrid model combining K-Means and SVM addresses these shortcomings by leveraging the strengths of both approaches. The model is designed to improve accuracy in detecting fraud, even in imbalanced datasets, by utilizing stratified sampling techniques to balance the representation of fraudulent and non-fraudulent claims. This ensures more reliable detection of fraud, which current models are unable to achieve at scale. The hybrid model's ability to adapt to evolving fraud patterns offers a more robust solution, thereby reducing the financial losses caused by undetected fraudulent claims and improving overall profitability for insurers.

The justification for this study is rooted in the significant financial impact of fraud and the inadequacy of existing detection models. The development of a more accurate, adaptive, and efficient fraud detection model is essential for mitigating these losses and improving the financial sustainability of insurance firms.

1.5 Scope

The study's scope was confined to the examination of claim records sourced exclusively from a prominent health insurance firm in Kenya. Spanning the timeframe of 2011 to 2022, these claim records were meticulously classified into distinct categories of either fraudulent or non-fraudulent cases. This classification process was facilitated through the utilization of a sophisticated hybrid machine learning model specially designed for this purpose.

1.6 Significance of the study/(Rationale)

The study's findings hold significant promise for aiding insurance firms in discerning intricate patterns and irregularities within claim data, leveraging the insights garnered from the trained machine learning model. The integration of machine learning algorithms and models is poised to notably enhance the efficacy of fraud detection techniques when juxtaposed with prevailing conventional methods.

The surge in fraudulent claims within the realm of medical insurance directly impacts the pricing structure of premiums for policyholders. Frequently, insurance firms adjust premium prices to mitigate the financial strains arising from escalating fraudulent activities (Hanafy & Ming, 2021). However, the adoption of machine learning models

is anticipated to markedly improve the efficiency of fraud detection techniques. Consequently, this enhancement is likely to curtail claim expenditures for insurance firms, subsequently resulting in lowered premiums for insurance services. Furthermore, the reduction in premiums extended by insurance firms has the potential to augment the accessibility of insurance in Kenya, where the current penetration rate stands at 2.3% of GDP.

The research outcomes bear significance for the broader research community, advancing the collective understanding regarding the application of machine learning in detecting insurance fraud. Within Kenya, this study is positioned to serve as a benchmark, emphasizing the imperative need for insurance firms to integrate machine learning models into their fraud detection frameworks.

CHAPTER TWO

LITERATURE REVIEW

2.1 Introduction

This chapter provides a comprehensive review of the literature on fraud detection methodologies within the insurance industry, with a particular focus on medical insurance. Fraud detection has become increasingly sophisticated as fraudulent schemes evolve, demanding more advanced methods beyond traditional approaches such as rule-based systems and manual reviews. In recent years, machine learning algorithms have emerged as powerful tools for detecting fraud, with a particular emphasis on supervised, unsupervised, and hybrid approaches.

The chapter is structured to systematically explore different aspects of medical insurance fraud, and the methods used for detection. Section 2.2 categorizes the various types of medical insurance fraud, distinguishing between service-availing and service-providing fraud patterns. Section 2.3 delves into the existing healthcare fraud detection methods, including traditional and technological approaches. Section 2.4 discusses data mining techniques and the application of machine learning in fraud detection, broken down into three parts: supervised learning (2.4.1), unsupervised learning (2.4.2), and hybrid models (2.4.3). This is followed by Section 2.5, which identifies the research gap in the current body of knowledge and highlights the need for more adaptable models, particularly those that address the issue of data class imbalance.

By critically evaluating existing research, this chapter lays the groundwork for the subsequent development of a hybrid machine learning model that combines the strengths of supervised and unsupervised approaches to better detect fraud in medical insurance.

2.2 Types of Medical Insurance Fraud

The study conducted by Matloob et al., (2020) delineate healthcare fraud into two primary categories based on distinct servicing patterns: service-availing patterns and service-providing patterns. Service-availing patterns encompass fraudulent activities perpetrated by the insured parties, while service-providing patterns pertain to misrepresentations and fraudulent actions by healthcare professionals.

Waghade & Karandikar (2018) systematically classify the various types of fraud prevalent within the healthcare industry based on the entities involved in fraudulent activities. Their comprehensive categorization identifies healthcare fraud as comprising service provider fraud, insurance subscriber fraud, insurance provider fraud, and conspiracy fraud. Service provider fraud encompasses a spectrum of fraudulent actions, including charging for services not rendered, overbilling, unbundling medical procedures into multiple stages, falsifying patient diagnoses, and manipulating treatment histories to justify unnecessary medical procedures. Insurance subscribers may engage in fraud by filing claims for non-received medical services, providing false information during enrollment to secure lower premium rates, or illegitimately utilizing another policyholder's coverage (Waghade & Karandikar, 2018). Conversely, insurance providers might engage in fraudulent practices by

misrepresenting benefits associated with a specific scheme or product, issuing counterfeit reimbursements to service providers or policyholders. Conspiracy fraud involves collusion among multiple entities to illicitly gain benefits (Waghade & Karandikar, 2018).

Zhou & Zhang's (2020) related study focuses on the driving forces behind medical insurance fraud. Their analysis highlights a complex interplay among medical insurance providers, medical service providers, and insured individuals, each motivated by distinct interests within this tripartite dynamic. For instance, medical service providers often aim to bolster hospital revenues by over diagnosing, inflating bills, and unbundling medical treatments (Zhou & Zhang's, 2020). In parallel, insured individuals may exploit illegal reimbursements, fabricate medical treatments, or undergo unnecessary consultations to curtail personal medical expenses. Consequently, insurance providers might manipulate information regarding offered products to minimize operational and claims costs (Zhou & Zhang's, 2020)..

These seminal studies collectively contribute to a comprehensive understanding of the multifaceted nature of healthcare fraud, elucidating distinct patterns, motivations, and stakeholders involved in perpetrating fraudulent activities within the medical insurance landscape.

2.3 HealthCare Fraud Detection Methods

The landscape of fraud detection in medical insurance has undergone significant evolution in recent years. Traditionally, as highlighted by Zhou & Zhang (2020),

conventional methods heavily relied on rule-based techniques for fraud identification, where claims were scrutinized based on pre-defined rules set by domain experts. However, the effectiveness of this rule-based evaluation was contingent upon the accuracy and completeness of these rules, consequently constraining its efficacy (Zhou et al., 2020). Furthermore, Waghade & Karandikar (2018) underscored the limitations of this approach, emphasizing its inefficiency and time-consuming nature, particularly due to the reliance on a limited number of experienced auditors tasked with handling a vast volume of claims.

Recent developments have seen the integration of electronic claim management systems into healthcare frameworks (Kose et al., 2015). These systems have been increasingly utilized for audit purposes, review, and automating claim processing. Ai et al., (2018) elaborate on the effectiveness of Electronic Claim Processing systems, categorizing them based on two principal inference rule engines: medical and insurance rules. Electronic Claim Processing systems have demonstrated superior efficiency and higher throughput in claim analysis compared to traditional expert domain analysis (Ai et al., 2018).

Moreover, the advent of Electronic Health Records within medical insurers and service providers has led to the accumulation of extensive healthcare-related data, necessitating the implementation of data analytics solutions for proficient fraud detection (Khoshgoftaar & Seliya, 2017).

A pivotal transition in fraud detection methodology is evident in the widespread adoption of big data technology within the medical insurance domain, as discussed by

Zhou & Zhang (2020). Mining algorithms rooted in big data analytics have gradually permeated the healthcare sector for insurance fraud detection purposes. This shift toward leveraging big data for fraud analysis traces back to earlier research in 1999, wherein studies identified potential data patterns associated with fraudulent activities (Zhou et al., 2020).

According to Joudaki et al. (2015), advancements in artificial intelligence, machine learning, and deep learning have paved the way for novel automated fraud detection methods. Data mining and regression emerge as primary approaches utilized in medical insurance fraud detection leveraging big data and machine learning techniques.

Collectively, these advancements underscore a paradigm shift in fraud detection methodologies within medical insurance, emphasizing the crucial role played by technological innovations, big data analytics, and machine learning in fostering more efficient and automated fraud detection processes.

2.4 Data Mining Approaches in Healthcare Fraud Detection

The domain of data mining techniques, as initially introduced by Segal (2016), offers a robust framework for identifying intricate patterns within extensive datasets. These methodologies not only afford insights into underlying models and trends but also excel in detecting anomalies and outliers by scrutinizing data against established models and profiles.

Building upon this foundation, Bauder et al. (2017) delve deeply into the application of data mining techniques within the specific context of fraud detection in the sphere of medical insurance. Within this milieu, they discern between structured and unstructured data. Structured data adheres to standardized formats amenable to tabular representation in conventional databases, whereas unstructured data lacks formal organization. Structured data, owing to its systematic nature, facilitates more straightforward analysis and modeling using data mining algorithms. Conversely, unstructured data demands additional preprocessing steps, such as parsing, to enable comprehensive analysis.

Extending the discourse, Zhou & Zhang (2020) provide an intricate and comprehensive exploration of data mining's instrumental role in healthcare fraud detection. Their study emphasizes the use of data mining techniques to discern potentially fraudulent medical treatments by scrutinizing aberrant data records. This approach frames fraud detection as a classic outlier detection problem, wherein different methodologies are employed. This includes classification-based anomaly detection, wherein datasets are categorized into normal and abnormal types using labeled data for training, effectively transforming the detection process into a binary classification problem. Distance-based detection methods gauge dataset distances through metrics like the local outlier factor (LOF), serving as an indicator of the likelihood of datasets being outliers. Statistical-based approaches operate under the premise that outlier datasets deviate from the distribution law of normal data. Additionally, cluster-based methodologies discern normal data, typically part of multi-point clusters, from outliers found in clusters with fewer or no data points.

Furthering the exploration of machine learning's role, Abdalla et al. (2016) categorize machine learning algorithms into supervised, unsupervised, and semi-supervised categories for effective fraud detection. Moreover, Bauder et al. (2017) advocate for the fusion of supervised and unsupervised algorithms to form hybrid machine-learning approaches, a direction that Zhou & Zhang (2020) endorse. They argue for the adoption of hybrid models due to their potential to mitigate the risks associated with creating new dataset labels while harnessing the combined strengths of supervised and unsupervised learning algorithms. These hybrid models, as elucidated, signify a burgeoning frontier in the enhancement of fraud detection methodologies within the domain of medical insurance (Abdalla et al., 2016).

2.4.1 Supervised Machine Learning

Supervised machine learning algorithms remain a cornerstone in the domain of data mining, particularly in the discernment between fraudulent and legitimate claims (Abdallah et al., 2016). This classification method relies heavily on a training dataset, using its structured information to create models capable of assessing novel claims against established class labels. Notably, the efficacy of these models heavily relies on their adaptability to evolving fraud patterns, necessitating frequent updates to reflect emerging fraudulent strategies within the intricate landscape of medical insurance (Joudaki et al., 2015).

A comprehensive review by Larnyo et al. (2018) delves extensively into the array of supervised algorithms instrumental in detecting and combating fraudulent activities within health insurance. Highlighting Bayesian Networks, K-Nearest Neighbor, SVM,

and decision trees among the roster of prominent techniques, their discussion underscores the versatility and varied applications of supervised learning in this domain. Furthermore, specific algorithms like Random Forest, SVM, and Gradient Boost demonstrate impressive accuracy rates in identifying fraudulent transactions, leveraging historical claim data to discern intricate patterns in the dataset (Abdallah et al., 2016). This reliance on labeled data epitomizes the machine learning's efficacy, yielding high recall and precision measures crucial in minimizing both false positives and negatives.

Carcillo et al. (2021) showcase the effectiveness of supervised learning models, citing the example of an SVM-based model achieving a commendable 95% accuracy in detecting credit card fraud. Additionally, models like Decision Trees and Logistic Regression not only offer interpretable insights but also empower analysts to delve deeper into the underlying factors influencing prediction outcomes. The transparent and interpretable nature of these models not only bolsters confidence in the decision-making process but also facilitates intervention by domain experts when required (Carcillo et al., 2021). Decision Trees visually articulate the decision-making trajectory, thereby enhancing the model's interpretability.

Nevertheless, challenges persist within the domain of supervised learning. Dou & Xiong (2017) underscore the inherent reliance of these algorithms on consistent human intervention for model updates, potentially compromising their proficiency in identifying previously unseen or novel fraudulent schemes. Moreover, the fundamental reliance on labeled data for training poses significant hurdles; the scarcity

of high-quality labeled datasets often impedes the creation of accurate and robust fraud detection models (Dou & Xiong, 2017). The scarcity of fraudulent cases in comparison to legitimate claims poses inherent risks of overfitting, potentially leading to diminished performance in identifying new fraud patterns (Joudaki et al., 2015). Biases or imbalances within the labeled data further complicate predictions, underscoring inherent limitations within supervised learning paradigms. These challenges warrant continued research efforts in refining supervised algorithms to circumvent these limitations for more effective fraud detection methodologies in the medical insurance sector (Dou & Xiong, 2017).

2.4.2 Unsupervised Machine Learning

Unsupervised learning, as delineated by Rawte & Anuradha (2015), operates without the reliance on predefined class labels, making it a pivotal methodology in discerning abnormal patterns and uncovering both established and emerging fraud types. The efficacy of unsupervised learning in healthcare fraud detection was rigorously investigated by Joudaki et al. (2015), who compared its applicability against supervised learning methods. Their findings, hinging upon data availability and accuracy, advocated the superiority of unsupervised learning in the realm of healthcare fraud detection. The unsupervised approach scrutinizes claim characteristics against the dataset, identifying correlations among features and outliers autonomously, and facilitating adaptability to evolving fraud patterns (Bergstra & Bengio, 2012). This adaptability diminishes the requisite constant updates and training mandated by supervised models.

Moreover, the intrinsic lack of labels in unsupervised algorithms renders them highly suitable for real-time or near-real-time fraud detection scenarios, where instantaneous analysis and responses are pivotal (Enders, 2010). The scalability and adeptness in handling voluminous datasets empower unsupervised algorithms to operate seamlessly in batch processing and high-throughput environments, facilitating continuous fraud monitoring and detection.

However, despite their myriad advantages, unsupervised learning algorithms exhibit a susceptibility to higher false positive rates compared to their supervised counterparts due to the absence of labeled training transactions (Dou & Xiong, 2017). The augmented false alerts pose additional analytical burdens on fraud analysts, necessitating comprehensive investigations into the root causes of flagged transactions. Furthermore, the inherent "black box" nature of clustering algorithms presents challenges in comprehending the rationale behind the identification of transactions as fraudulent (Joudaki et al., 2015). This lack of interpretability is particularly problematic in heavily regulated sectors like medical insurance, where transparency and explicability in autonomous decision-making processes are imperative (Enders, 2010). Addressing these limitations remains a crucial avenue for refining and enhancing the applicability of unsupervised learning algorithms in healthcare fraud detection, especially within the intricate landscape of medical insurance Joudaki et al. (2015).

2.4.3 Hybrid Machine Learning

The integration of supervised and unsupervised learning, as highlighted by Bauder et al. (2017), represents a pivotal approach in data mining termed hybrid learning. This fusion combines the advantages while mitigating the shortcomings inherent in singular supervised or unsupervised techniques. Notably, Lawand & Kulkarni's (2019) examination of insurance fraud prediction involves solving the classification problem within the input space, deploying a suite of algorithms such as Random Forest, decision trees, Naïve Bayesian Classification, and SVM. Their robust model, evaluated through recall and precision metrics derived from the confusion matrix, showcases heightened accuracy.

However, the impact of imbalanced datasets on machine learning accuracy is a pertinent concern, as demonstrated by Hanafy & Ming (2021) in their comparison of 13 machine learning methods in insurance fraud detection. Addressing this issue, they implement resampling techniques like Random Over Sampler and Random Under Sampler, showcasing how these methodologies enhance classifier algorithms' accuracy, with SVM and C5.0 standing out as top performers under specific sampling methods.

The prowess of SVM in outlier detection is underscored by Naik & Laxminarayana (2017), who explicate SVM's representation of data items in an n-dimensional space and its robustness in discerning hyperplanes that differentiate classes. Similarly, Rawte & Anuradha (2015) delve into a hybrid machine learning model's development,

leveraging the Evolving Clustering Model and SVM to detect outliers and duplicate claims, albeit with limitations in detecting other forms of medical fraud.

The mechanics behind the K-Means algorithm for clustering, elucidated by Naik & Laxminarayana (2017), involve iteratively defining centroids that represent clusters within a given input space, iteratively associating data with the nearest centroids until convergence. Ogbuabor & Ugwoke (2018) compare the performances of K-Means and DBSCAN algorithms, revealing K-Means' superiority in accuracy and execution time due to its solid inter-cluster separation and intra-cluster cohesion.

Practical applications of clustering algorithms like K-Means in medical claim record analysis, as conducted by Wakoli et al. (2014) and Zhang et al. (2020), exhibit the efficacy of these methodologies in identifying suspicious claims. While traditional rule sorting demonstrated a 24% detection rate, clustering algorithms like DBSCAN, K-Means, Isolation Forest, and Local Outlier Factor displayed detection rates ranging from 33% to 47%, underscoring their potency in identifying potential fraud cases within medical insurance claims.

2.5 Research Gap

Despite significant advancements in fraud detection methodologies, a critical gap remains in the ability of current systems to accurately detect medical insurance fraud in large, imbalanced datasets. Traditional rule-based approaches, although effective in specific cases, lack the scalability and adaptability required to handle the evolving nature of fraudulent schemes (Zhou & Zhang, 2020). These systems rely heavily on

predefined rules, which must be continuously updated to reflect emerging fraud patterns, often leading to inefficiencies in fraud detection (Matloob & Khan, 2019).

Supervised machine learning algorithms, such as SVM, are commonly employed due to their accuracy in structured datasets. However, their performance is significantly hindered by data class imbalances, where legitimate claims vastly outnumber fraudulent ones (Alloghani et al., 2020). As a result, these models tend to misclassify fraudulent claims, leading to a high rate of false negatives. This is problematic in real-world applications where fraudulent claims represent only a small portion of the total claims dataset (Waghade & Karandikar, 2018).

Unsupervised models like K-Means clustering offer the flexibility of detecting unknown patterns and anomalies without the need for labelled data. However, these models are less effective at providing the precision necessary for fraud detection, often flagging many false positives due to their inability to differentiate between benign anomalies and actual fraud cases (Rawte & Anuradha, 2015). Furthermore, unsupervised models are not designed to handle complex fraud schemes that require nuanced classification, limiting their effectiveness in identifying more sophisticated fraud patterns (Raj & Lin, 2020).

Recent research has shown promise in hybrid machine learning models, which combine both supervised and unsupervised approaches to address the limitations of each method (Bauder et al., 2017). However, studies focusing on hybrid models often overlook the issue of data class imbalance, which remains a major obstacle in fraud detection. While hybrid models improve detection accuracy by integrating

classification and clustering techniques, many studies do not apply proper resampling techniques to balance the dataset, resulting in continued misclassification of fraud cases (Hanafy & Ming, 2021). Additionally, most research does not explore the potential for tuning these models to optimize their performance in real-world settings.

This study seeks to fill these gaps by developing a hybrid machine learning model that integrates K-Means clustering with SVM classification and addresses the issue of data class imbalance through stratified sampling techniques. By leveraging the strengths of both supervised and unsupervised methods and properly balancing the dataset, the proposed model aims to significantly improve fraud detection accuracy, precision, and recall, even in large, imbalanced datasets. Table 2.1 below highlights the summary of references and their deficiencies.

Table 2.1: Summary of the references and their deficiencies

Reference	Methodology	Deficiencies
Matloob & Khan (2019)	Rule-based systems	Inefficient at handling evolving fraud patterns; requires constant rule updates.
Zhou & Zhang (2020)	Rule-based systems	Lack of scalability and adaptability to new fraud schemes; performance diminishes as fraud becomes more sophisticated.
Alloghani et al. (2020)	Supervised (SVM)	Performance hindered by data class imbalance; high rate of false negatives when detecting rare fraudulent cases.
Waghade & Karandikar (2018)	Supervised (SVM)	Struggles with large, imbalanced datasets; overfits legitimate claims and misses minority fraud cases.
Rawte & Anuradha (2015)	Unsupervised (K-Means)	High false positive rates due to inability to differentiate between anomalies and actual fraud cases; lacks precision for fraud classification.
Raj & Lin (2020)	Unsupervised (K-Means)	Effective at anomaly detection but lacks the complexity needed for nuanced classification of fraudulent and legitimate claims.
Bauder et al. (2017)	Hybrid (Supervised + Unsupervised)	Improvement in accuracy but neglects addressing data class imbalance; continued misclassification of fraud cases.
Hanafy & Ming (2021)	Hybrid (Supervised + Unsupervised)	Fails to implement proper resampling techniques to balance datasets, leading to biased detection towards legitimate claims.

2.6 Theoretical Model for Implementing a Hybrid Model for Fraud Detection

To effectively detect fraud in medical insurance claims, the proposed hybrid model integrates both supervised and unsupervised machine learning techniques. This section outlines the theoretical framework for implementing a hybrid model combining K-

Means clustering (unsupervised) and SVM classification (supervised). The model leverages the strengths of both approaches to address data class imbalance, enhance detection accuracy, and adapt to evolving fraud patterns.

2.6.1 Overview of the Hybrid Model

The hybrid model begins by employing K-Means clustering to group similar data points (claims) based on their feature characteristics. This clustering process helps reduce noise in the dataset, making it easier for the subsequent classification algorithm to focus on relevant patterns and anomalies. The SVM, a supervised learning algorithm, is then applied to classify each claim as either fraudulent or legitimate.

The key steps in implementing the hybrid model are as follows:

1. **Data Preprocessing and Feature Selection**:
    a. **Data Cleaning**: Remove incomplete or irrelevant data to ensure data quality.
    b. **Feature Engineering**: Identify the most relevant features that influence fraud detection (e.g., claim amount, patient demographics, treatment duration).
    c. **Normalization/Scaling**: Standardize the dataset to ensure that all features are on the same scale, which is crucial for algorithms like SVM and K-Means.

2. **Handling Data Imbalance**:

    a. **Stratified Sampling**: To address the class imbalance, a stratified sampling technique is applied. This ensures that both fraudulent and legitimate claims are proportionally represented in the training dataset, preventing the model from overfitting to the majority class (legitimate claims).
    b. **Oversampling/Undersampling**: Alternatively, oversampling techniques (e.g., SMOTE) or undersampling can be used to artificially balance the dataset.

3. **K-Means Clustering (Unsupervised Learning)**:

    a. **Clustering Claims**: Apply the K-Means algorithm to group similar claims into clusters based on feature similarities. This step helps in reducing noise by clustering similar claim patterns, which makes the dataset more structured and prepares it for classification.
    b. **Cluster Optimization**: The number of clusters (k) is determined using techniques like the elbow method, ensuring that the dataset is divided into optimal clusters.
    c. **Assigning Cluster Labels**: Each claim is assigned a cluster label, which becomes a new feature used by the SVM classifier in the next step.

4. **SVM (Supervised Learning)**:

    a. **Classification**: The SVM is trained on the clustered dataset to classify each claim as either fraudulent or legitimate. The SVM separates the two classes by finding the optimal hyperplane in an n-dimensional feature space.
    b. **Kernel Selection**: The SVM's performance can be enhanced by selecting an appropriate kernel (linear, polynomial, or radial basis function) based on the dataset's characteristics.
    c. **Tuning Hyperparameters**: Hyperparameters such as the regularization parameter (C) and kernel parameters are optimized using grid search or random search to maximize classification performance.

5. **Evaluation Metrics**:

    a. The model is evaluated using metrics such as accuracy, precision, recall, and F1 score. These metrics assess the effectiveness of the model in correctly classifying fraudulent claims while minimizing false positives and false negatives.
    b. A **confusion matrix** is used to further analyze the model's performance by evaluating the true positives, true negatives, false positives, and false negatives.

6. **Model Optimization and Cross-Validation**:

    a. **Cross-Validation**: Perform k-fold cross-validation to validate the model's performance across multiple subsets of the data, ensuring robustness and generalizability.
    b. **Principal Component Analysis (PCA)**: If the dataset has many features, PCA can be applied for dimensionality reduction. This step reduces computational complexity and enhances model efficiency by selecting the most significant components.

7. **Deployment and Real-Time Detection**:

    a. After training and evaluation, the model can be deployed for real-time fraud detection in medical insurance systems. The hybrid model continuously classifies new claims and flags suspicious transactions for further investigation.
    b. **API Integration**: The model can be integrated with existing insurance management systems via APIs, allowing for real-time fraud detection as claims are processed.

2.6.2 Theoretical Justification for the Hybrid Approach

The combination of unsupervised and supervised learning provides several advantages:

1. **Noise Reduction**: By clustering similar claims with K-Means, the model reduces noise in the data, allowing the SVM to focus on classifying more relevant patterns.
2. **Handling Imbalanced Data**: The hybrid model addresses the imbalance in the dataset by ensuring the clustering process captures both fraudulent and legitimate claims, and by using stratified sampling to prevent the SVM from being biased toward the majority class.
3. **Improved Fraud Detection**: The hybrid approach enhances accuracy by leveraging unsupervised clustering to detect novel fraud patterns, while the SVM provides precise classification of known fraud instances.
4. **Scalability and Adaptability**: The hybrid model can adapt to changing fraud schemes by retraining with updated data and adjusting the clustering and classification parameters, making it highly scalable for real-world applications.

CHAPTER THREE

RESEARCH METHODOLOGY

3.1 Introduction

This study adopted the Cross-Industry Standard Process for Data Mining (CRISP-DM) methodology to guide the development of the hybrid fraud detection model. CRISP-DM is a flexible, iterative approach that is particularly effective for machine learning projects dealing with complex datasets. It involved several key phases: understanding the business problem, examining and preparing the data, developing and testing models, and evaluating their performance.

In this study, these phases included gaining insights into medical insurance fraud, cleaning and transforming the dataset, building a hybrid model combining K-Means clustering with SVM classification, and assessing the model using metrics such as accuracy, precision, and recall. This structured process ensured that the model was robust and adaptable to evolving fraud patterns.

3.2 Research Design

The research exclusively adopted an experimental research design to address the primary objectives concerning medical insurance fraud detection. It focused on utilizing numeric data derived from historical medical claims to accomplish two specific aims:

1. Development of a Hybrid Machine Learning Model: The experimental analysis was primarily centered on constructing a hybrid machine learning model for detecting fraudulent medical insurance claim. This involved the utilization of data exclusively from past medical claims to integrate K-Means and SVM algorithms.
2. Evaluation of Model Performance: Another crucial aspect of the experimental analysis was the evaluation of the hybrid machine learning model's effectiveness. The assessment specifically compared the performance of this hybrid model against that of a standalone SVM model.

The experimental methodology employed in this research was oriented towards the analysis of numerical data sourced from historical medical claims.

3.3 Data Sampling

The study focused on the medical claim data from a leading health insurance firm in Kenya. Private and public insurers can provide the medical insurance line of business and proper detection techniques and findings can be applied across the industry. The study focused on historical claim data from 2011 to 2022. This provided the research with a large sample size to analyze trends in fraud over the years. The dataset included both inpatient and outpatient claim data. The dataset contained claims lodged by service providers from January 2011 to October 2022 with the insurance company, equating to 12,301,761 rows with 5,734,156 unique claims recorded. The sample dataset used for the model was selected with a random seed and limited to 350,000 unique invoices. The purpose of seeding the dataset was to produce a consistent

random input resulting in 1,194,833 rows of data and 83 columns of data. Features containing personally identifiable data were truncated before outputting the seeded data to a CSV file. Moreover, there was a need to choose a robust sampling method to address class imbalance and reduce overfitting. The study used the stratified sampling technique to deal with the class imbalance concerns. Stratified sampling involved dividing the dataset based on the class labels and identifying the total number of records classified as fraudulent claims.

3.4 Data Collection

The study utilized secondary data on past medical claims provided by an insurance firm. To obtain a relevant dataset, the researcher first prepared a formal data request form specifying the attributes needed to align with the study objectives. The requested data attributes included: Claimant metadata (e.g., age, gender, employment status), claim location, treatment history, cost of treatment, treatment facility, and risk profile of insured.

The data request form was submitted to the insurance firm's Chief Data Officer. After reviewing the request, the insurance firm provided a dataset extracted from their claims database covering the period of January 2011 to October 2022. Once the dataset was received, several validity checks were conducted. The researcher verified the dataset included the requested attributes, had the expected number of records based on the specified timeframe, and contained realistic/valid values. Basic statistical summaries were generated, and outliers examined to ensure the data was clean and suitable to address the study aims. Minor data cleaning steps were taken as needed, such as

handling missing values and excluding any claims with inaccurate or erroneous data. The final validated dataset contained 1,194,833 rows of data and 83 columns of data, providing a robust set of secondary data to train and test the machine learning models for fraud detection.

3.5 Data Mining Methodology

The Cross-Industry Standard Process for Data Mining (CRISP-DM) methodology was used for the research. Developed by a consortium of data mining agents through an initiative sponsored by the European Union, CRISP-DM depicted data mining as a six-phase cycle (Schröer, Kruse, & Gómez, 2021). The methodology consisted of the following phases: business understanding, data understanding, data preparation, modeling, evaluation, and deployment. Ordering the phases in the CRISP-DM methodology is flexible (Schröer, Kruse, & Gómez, 2021). Figure 3.1 shows the phases of CRISP-DM.

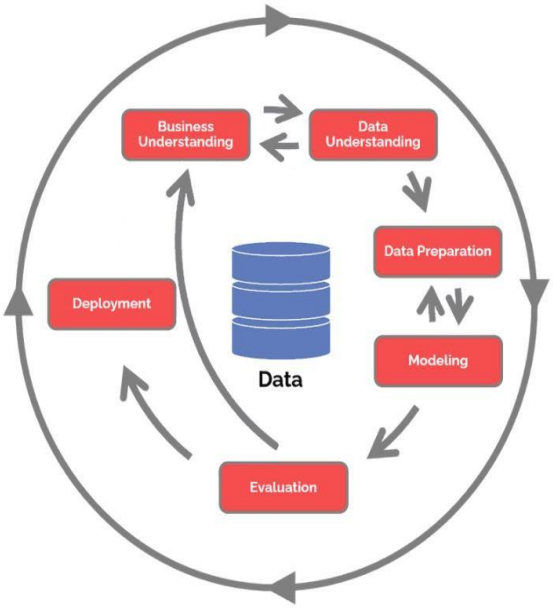

Figure 3.1: Phases of CRISP-DM **(Schröer, Kruse, & Gómez, 2021)**

The business understanding phase of CRISP-DM involved a review of the journals and research papers on the various types of medical insurance fraud, how each fraud type is detected, and the current mitigation strategies for each, which assisted the research to gain deeper domain knowledge on the current controls, the evolution of fraud schemes, and which datasets can be mined for fraud detection in medical insurance. Moreover, this phase helped in selecting the appropriate technique to be used in answering the research questions.

3.5.2 Data Understanding

The data understanding phase involved the collection of relevant medical claims datasets in tandem with the acquired domain knowledge. Subsequently, the research sought to familiarize with the claim datasets and ascertain the quality of datasets in developing the fraud detection model.

3.5.3 Data Preparation

In this phase, the raw data provided was transformed into a standard acceptable format. This involved several activities, such as selecting relevant attributes and removing irrelevant attributes, handling null or missing values, and removing duplicate entries. The data cleansing step used a process called imputation which identified inaccurate and incomplete datasets, substituted missing data with a placeholder, and noise reduction by removing data that did not relate to the research questions and objectives. Moreover, several steps to prepare the data for training using the models were implemented. The study dataset contained features with white spaces, empty values, inconsistent flag data types and null values. The data cleaning phase involved handling missing data. Missing data can either be imputed or removed from the feature depending on the extent to which the data is not present (Enders, 2010). Yuan (2014) posits that machine learning algorithms cannot handle missing or incomplete data directly. Imputation is a data cleaning technique used to handle missing data by estimating the missing data with likely substitutes.

3.5.4 Modeling

The modeling phase involved developing a hybrid machine learning that implemented both supervised and unsupervised learning. The k-means algorithm was applied to the dataset to cluster similar features, and the SVM was harnessed to classify fraudulent and non-fraudulent claims.

3.5.5 Evaluation

After training the K-Means and SVM algorithms, the confusion matrix and the classification metrics were used to evaluate the efficiency and performance of the model on claims data on the insured. The model tested how many claims are categorized as false positives and false negatives (recall measure). Additionally, the model's performance was gauged by the percentage of correct classification of fraudulent claims (precision measure). To improve the accuracy of the model and reduce the false negatives and false positives, the input space used a random resampling technique which will rebalance class distribution for the imbalanced dataset.

3.5.6 Deployment

After successfully evaluating the hybrid data mining model, Python was used to deploy the model to detect fraud in medical insurance claims.

3.6 Ethical Considerations

Before the analysis and data mining process commenced, the researcher sought permission to follow the company's procedures and policies. The claim data and policyholder’s treatment history are considered the organization's intellectual property, and any analysis or access must be done within the confines that preserve the confidentiality of the data. Moreover, the research was in line with the Data Protection Act of Kenya 2019, which protects the privacy and rights of the data subjects, that is, policyholders and service providers. The claimants’ personal details, contacts, policy details, and treatment history were anonymized in the research to adhere to the Data Protection Act of Kenya 2019.

CHAPTER FOUR

RESULTS AND DISCUSSION

4.1 Introduction

The chapter builds upon the research methodology and details the implementation of the hybrid machine learning model. This section focuses on the practical aspects of the project and seeks to link the theoretical foundation elucidated in the previous chapters. We begin the chapter by describing the data cleaning process and the characteristics of the dataset using Exploratory Data Analysis. Next, we discuss the results based on the formulated research question which includes; how medical insurance fraud occurs, the current machine learning techniques used for fraud detection by medical insurance firms, how the achieved data is used to develop a hybrid machine learning model for detecting fraud in medical claims, and lastly, demonstrate the effectiveness of the hybrid learning model in detecting medical insurance claims compared to a lone SVM model.

4.2 Data Cleaning

The study evaluated the percentage of missing values per feature and plotted instances where over half of the features were missing, as depicted in Figure 4.1. These included QA_VERIFIED, SECOND_DIAGNOSIS, PRIMARYSPECIALIST, SECONDARYSPECIALIST & REASONCODEDESC. Duplicated features that could be derived from an existing feature set or represented by another feature name were dropped. Identifying these features aided in identifying the strategy for feature

engineering, feature relevance, and imputation strategies. The study adopted a dual approach to cleaning data. Features duplicated or depicted with other features were dropped during the data cleaning phase. Features such as PRIMARYSPECIALIST and SECONDARYSPECALIST, which depicted flag values that are null for absent or not applicable and inverse, were transformed to Boolean features indicating the presence/occurrence of the feature.

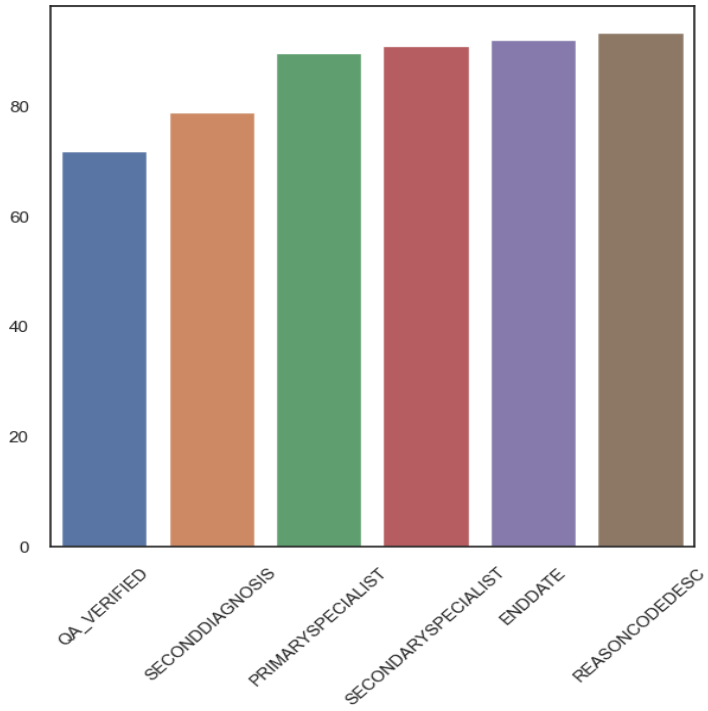

Figure 4.1: Columns with over 50% missing feature values

The dataset contained date-related features such as the date of the invoice, the date of the claim, the date of admission, the date of discharge, and the date of claim assessment, among other date features. These date features were initially extracted as string outputs to ensure compatibility and handle errors during the data collection stage. Date-related features denoted with string data types were converted to the date data type, as indicated in Figure 4.2. The conversion ensured that date-related features, such as the number of days admitted, could be engineered. Examples of engineered features include the beneficiary's age and the type of admission when the claim was made.

```
train['INVOICEDATE'] = pd.to_datetime(train['INVOICEDATE'])
train['POLICYSTARTDATE'] = pd.to_datetime(train['POLICYSTARTDATE'])
train['RENEWALDATE'] = pd.to_datetime(train['POLICYSTARTDATE'])+ pd.offsets.DateOffset(years=1)
train['DOB'] = pd.to_datetime(train['DOB'], errors = 'coerce')
train['TREATMENTDATE'] = pd.to_datetime(train['TREATMENTDATE'])
train['RECEIVEDDATE'] = pd.to_datetime(train['RECEIVEDDATE'])
train['CAPTUREDATE'] = pd.to_datetime(train['CAPTUREDATE'])
train['RECEIVEDDATE'] = pd.to_datetime(train['RECEIVEDDATE'])
train['DISCHARGEDATE'] = pd.to_datetime(train['DISCHARGEDATE'])
train['ADMISSIONDATE'] = pd.to_datetime(train['ADMISSIONDATE'])
train['PAYMENTDATE'] = pd.to_datetime(train['PAYMENTDATE'])
train['EFFECTIVEDATE'] = pd.to_datetime(train['EFFECTIVEDATE'])
```

Figure 4.2: Conversation of date columns from string dtype to date dtype.

The beneficiary age at the time the claim was recorded was derived by subtracting and rounding the number of years between the year of birth and the admission date. The admission type, inpatient or outpatient, was determined by the number of days the patient was hospitalized, which was derived by subtracting the discharge date from the admission date.

4.3 Exploratory Data Analysis

The Exploratory Data Analysis (EDA) phase aided primarily in understanding and summarizing the patterns within the dataset. The EDA, often characterized by visual depictions and correlations of the features, is important in validating the dataset's quality. The study deployed EDA primarily to provide a global view of the dataset in terms of the predefined labels. The initial step in EDA plotted the two main classifications of the study, fraudulent and non-fraudulent claims, as shown in Figure 4.3. Results from the bar graph reveals that most of the assessed claim (85%) cases we.re non-fraudulent, while 15% were fraudulent.

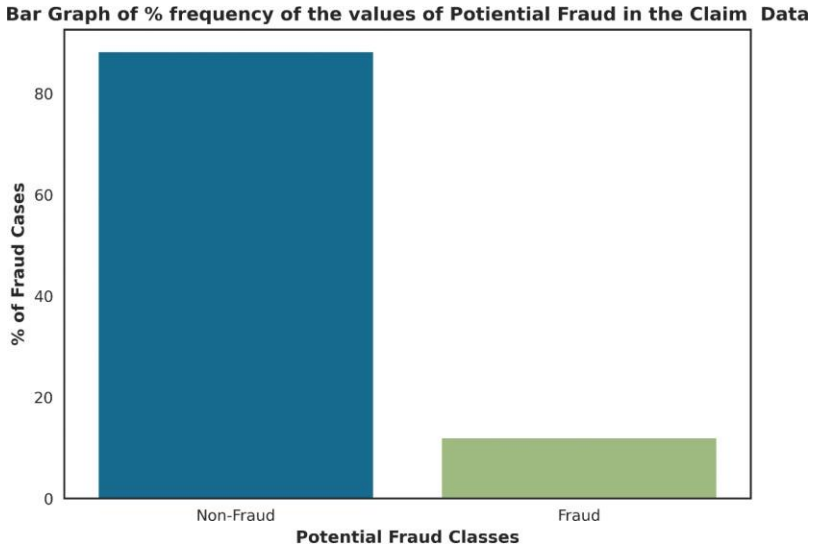

Figure 4.3: Bar graph showing the percentage of fraudulent claims in the dataset.

Further, EDA was used to elucidate the relationship between features and classification of fraudulent claims. These features pertained to critical domain indicators of possible fraudulent claims. The study sought to identify the link between the claimant and the likelihood of the claim being declined on fraudulent grounds (Figure 4.4). Results from the study show that most of the declined claims on fraudulent grounds were from unmarried children, followed by the company employees and lastly their spouses.

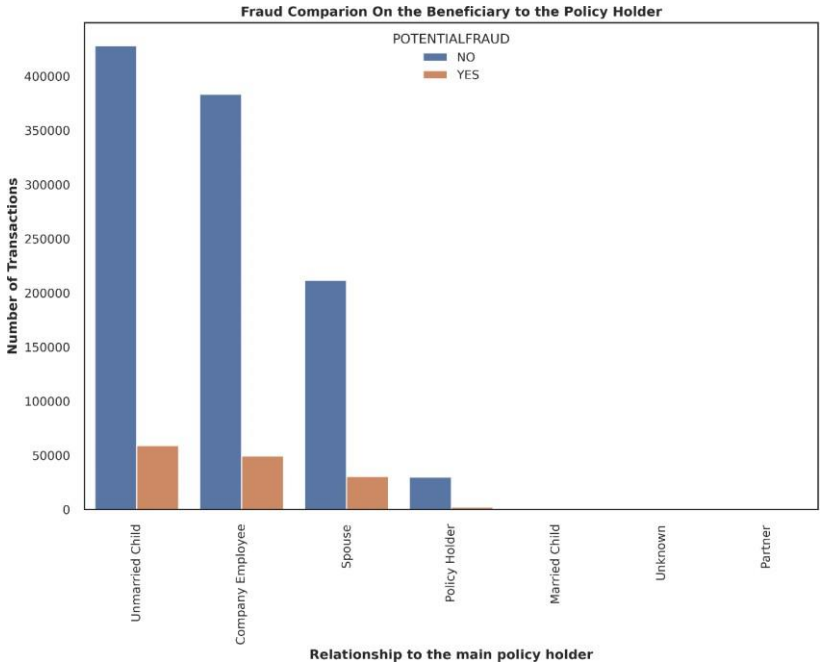

Figure 4.4: Bar graph showing classification based on beneficiary relationship

From the data analysis, other important positive and direct correlations with the predefined labels included the age of the claimants (Figure 4.5), type of admission,

and number of admission days (Figure 4.6). The link between the features was plotted on a correlation heatmap indicated in Figure 4.7.

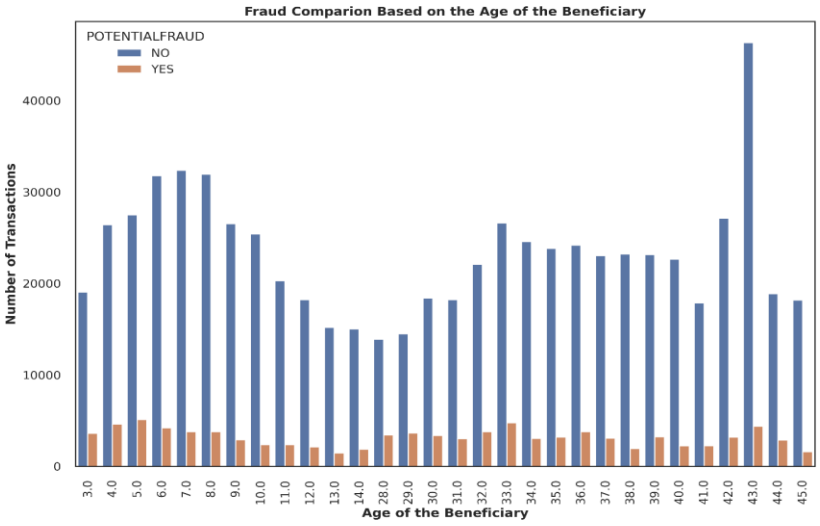

Figure 4.5: Comparison of claim classification based on the beneficiary age.

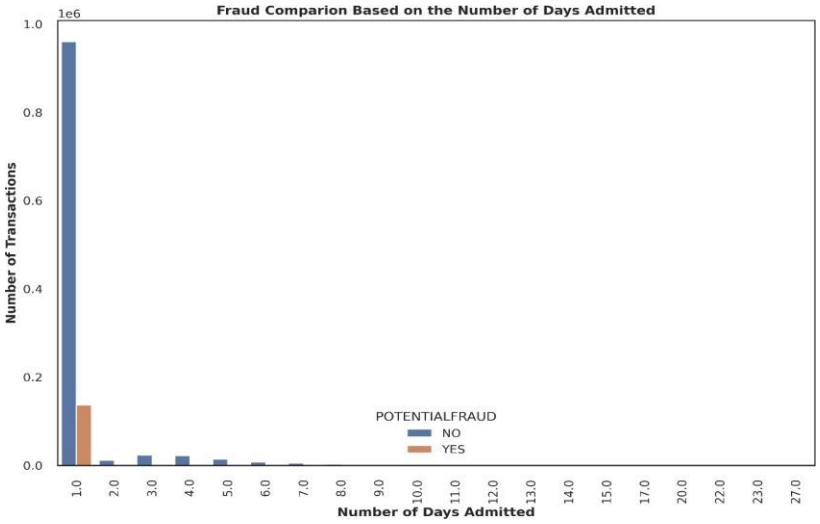

Figure 4.6: Comparison of claim classification based admission days.

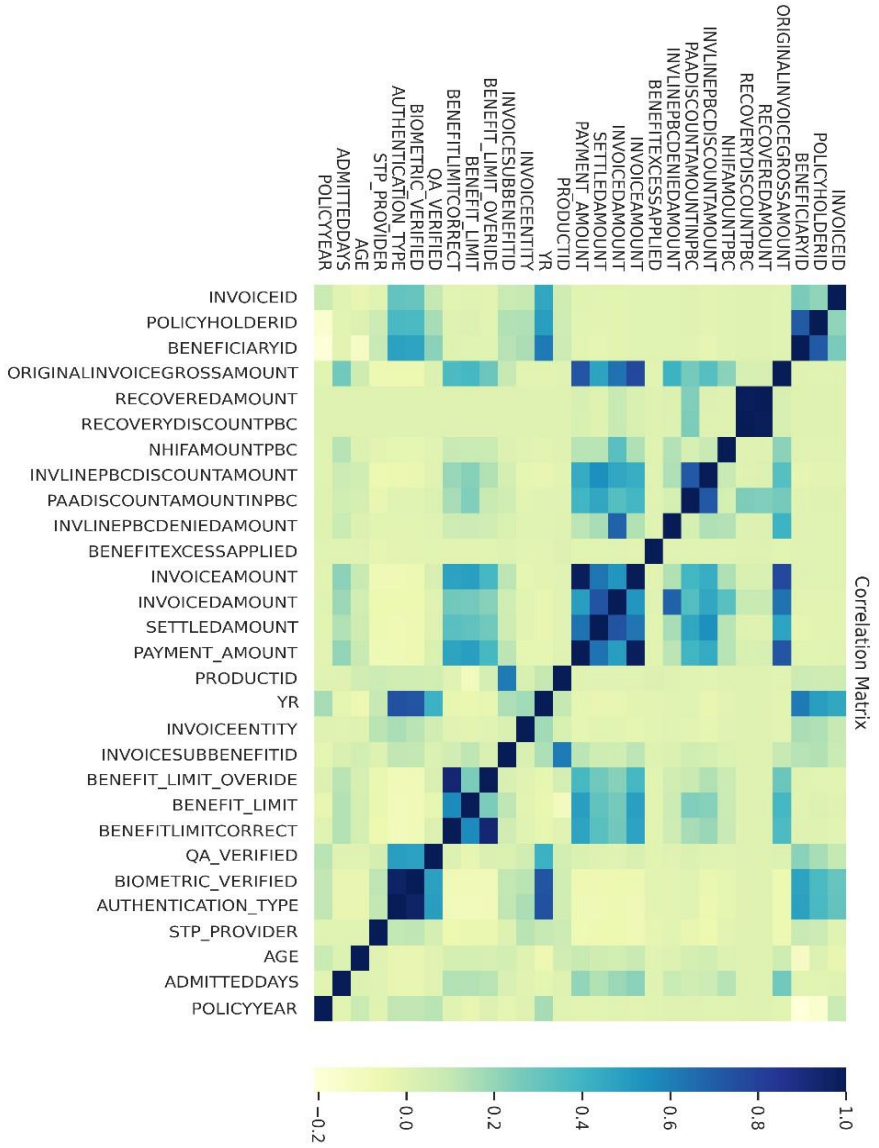

Figure 4.7: Heat map showing the correlation of the features.

4.4 Discussion of Research Findings

The results of the discussions are presented as below:

Objective 1: Develop a hybrid machine learning model to detect fraud in medical insurance claims.

Objective 2: Evaluate the performance of the hybrid model to detect fraud and compare it with a lone SVM model.

4.4.1 Development of hybrid machine learning model to detect fraud in medical insurance claims.

4.4.1.1 Feature Engineering

After a visual depiction of the dataset through EDA, the dataset features were engineered to improve the performance and accuracy of the model. Feature engineering involved feature selection, feature transformation, feature creation and dealing with missing data as informed by the domain knowledge elucidated by the previous processes. The feature selection step involved dropping features with weak correlation to reduce the noise of the dataset. Categorical feature values, which were denoted as either true, false, or null, were mapped into a dictionary and represented with 1 for true or yes, and 0 for null, false, or no values. From the business rule, fraudulent claims were denoted with a null payment date, while valid and approved claims had a date value for when the payment was completed, which was later used in the classification of the algorithms.

New features relating to the dataset were derived from the mean or sum of the existing claim records. These included total beneficiary visits to a hospital in a policy year, total number of invoices received per beneficiary in a policy year per service provider, total number of invoices per beneficiary in the policy year, average cost of a service provider per year, total claims submitted by a service provider per year, average number of procedures conducted per claim and the average admission days per service provider.

The study aggregated the dataset using the mean of the features grouped by service provider, beneficiary, and fraud categorization. The service provider and the beneficiary identifiers were dropped from the dataset to remain with qualitative and quantitative features of the dataset. The data transformation stage aggregated all the feature columns using the mean, unique value, or sum, resulting in one row for every invoice. The transformation further reduced the row set to 140 953 records. Features denoted as arrays after the transformation were iterated to the data cleanup step. These features in the array were joined into a comma-separated value string using a Python lambda function. A new feature was created in the data frame to record the count of the comma-separated string values. A Boolean feature was created for values with a flag instead of one hot encoding.

4.4.1.2 Data Sampling

Most of the claims in the dataset were genuine claims. The EDA phase observed that 88.56% and 11.44% of the claims were denoted as non-fraudulent and fraudulent respectively. Based on the latter, the dataset was randomly resampled on each stratum

to produce two balanced datasets which were combined to one dataset. The merged dataset split into training and testing subset using a 7:3 ratio for evaluation of the performance of the models.

4.4.1.3 Implementation

The implementation phase sought to compare the performance of the hybrid fraud detection model vis a vis the use of a sole supervised machine learning algorithm – SVM. Additionally, these two models were tuned as the last step of their iterations and the performance metrics recorded. This resulted in four models, that is, the lone SVM model, the tuned SVM model, the hybrid model (SVM and K-Means), and the tuned hybrid model.

4.4.1.4 SVM Classifier

The transformed dataset was loaded to the SVM model with default hyperparameters for SVC as shown in Table 4.1.

Table 4.1: Default hyperparameters for the SVC model.

Parameter	Default Value
C	1
Kernel	'rbf'
Degree	3
Gamma	'scale'
coef0	0
Shrinking	TRUE
Probability	FALSE
Tol	1.00E-03
class_weight	None

The default hyperparameters were then tuned using the Grid Search Cross Validation library in Python to obtain the optimal parameters for the SVM classification algorithm. These optimized and non-optimized predictions were later used as the benchmark for evaluating the classification performance of the hybrid model. Performance evaluation metrics for the study included accuracy, precision, recall and F1 score. These metrics were plotted in a Confusion Matrix to provide a detailed breakdown of the true positive, true negative, false positive, and false negative predictions made by the algorithm.

4.4.1.5 Hybrid Machine Language Model

The first iteration of the implementation of the hybrid model involved the use of a pipeline with the K-Means and SVM. The pipeline workflow was designed to run the

standardized dataset with the default K-Means for clustering and classifying the output using SVM. The clustering and classification were performed using the default kernel hyperparameters of the K-Means (shown in Table 4.2) and SVM algorithms (shown in Table 4.1) respectively.

Table 4.2: Default values for the K-Means algorithm for model three

Hyperparameter	Default Value
n_clusters	8
init	'k-means++'
n_init	10
max_iter	300
tol	1.00E-04
precompute_distances	'auto'
verbose	0
random_state	None
copy_x	TRUE
algorithm	auto

The second iteration of the hybrid model applied the grid search library to exhaustively obtain the optimal parameters for the scaler, principal component analysis components, k-means clusters, and the hyperparameter C in SVM. The parameter grid for the scaler parameter evaluated the Standard Scaler, Robust Scaler, and Quartile

Transformer. The Standard Scaler independently transforms input features to a standard deviation of 1 and a mean of 0, ensuring that the feature values are centered around 0. Conversely, the Robust Scaler transforms features using the median value and the interquartile range (IQR) (Altman & Krzywinski, 2017). This is performed by subtracting the median of each feature value and dividing the centered values using the interquartile range, thus making the scaled data less sensitive to outliers. The Quartile Transformer transforms each feature column by estimating the input data's cumulative distribution function (CDF) to result in features that follow a Gaussian distribution (Altman & Krzywinski, 2017). The parameter grid for the PCA components ranged from 14 to 22 with an increment of 2. Similarly, the number of k-means clusters ranged from 6 to 12 with an increment of 2. The performance metrics of the four models were recorded and evaluated, as discussed in the next chapter.

4.4.2 Evaluation of the performance of the hybrid model to detect fraud in comparison to the lone SVM model.

Several researchers have posited that hybrid machine-learning algorithms are more advantageous than standalone supervised or unsupervised algorithms (Zang & Ma, 2020; Bauder et al., 2017; Abdallah et al., 2017). This section detailed the steps taken to set benchmarks using an SVM model, optimize the SVM model, prototype the hybrid model, and optimize the hyperparameters of the hybrid model. The section also evaluates the effectiveness of the hybrid model against the benchmark (a lone SVM) model in the classification of fraudulent claims.

4.4.2.1 Result Evaluation

The evaluation benchmarks were set by prototyping a classification model using an SVM classifier and optimizing the hyperparameters of the SVM classifiers. Similarly, the hybrid algorithm was modeled with default hyperparameters in the first iteration and tuned using grid search for the second iteration with the introduction of PCA. The study deployed several evaluation metrics, including confusion matrix, classification accuracy, and classification report.

4.4.2.2 Classification Accuracy

Classification accuracy, calculated by dividing the number of correct predictions by the total number of predictions, is a common performance metric used to evaluate machine learning models by measuring the classified instances of true positives and true negatives (Altman & Krzywinski, 2017).

Table 4.3: Summary of the accuracy metrics of the models.

Iteration	Model Description	Accuracy Level
Model one	SVM with default hyperparameters	91.31%
Model two	SVM with optimal hyperparameters	97.05%
Model three	Hybrid model with default hyperparameters	68.08%
Model four	Hybrid model with tuned hyperparameters	97.49%

4.4.2.3 Confusion Matrix

The study summarized the performance of the four prototypes using a confusion matrix to evaluate the classification performance by categorizing the predicted and actual

labels into True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN).

True Positives refer to the instances where the model predicted the positive class correctly with the actual label being positive. True Negatives occur when the model predicts a negative class correctly, thus matching with the negative label. False Positive (Type I errors occur when the model mispredicts a positive class, whereas the label is negative. False negative (Type II) error occurs when the model predicts a negative class, whereas the actual label is positive.

The table 4.4 below summarizes the confusion matrix of the four models depicted in Figure 4.12, Figure 4.13, Figure 4.14, Figure 4.15.

Table 4.4: Summary of the confusion matrix for the four models

Summary	TP	FP	FN	TN	Type I error rate	Type II error rate
Model one	1138	106	109	1122	4.28%	4.40%
Model two	1219	45	28	1183	1.82%	1.13%
Model three	784	328	463	900	13.25%	18.71%
Model four	1217	32	30	1196	1.29%	1.21%

The SVM model with default parameters noted a 4.28% Type I error and a 4.40% Type II error rate from the entire dataset. With the second iteration, both Type I and Type II error rates reduced to 1.82% and 1.13%, respectively, with an overall accuracy of 97.05% noted. The third prototype recorded increased Type I and Type II error rates

at 13.25% and 18.75%. This reduced the accuracy of the model to 68%. The fourth model had the best accuracy rate of 97.45, mainly attributed to the lowest Type 1 error at 1.29%. The Type II error rate for the fourth model stood at 1.21%

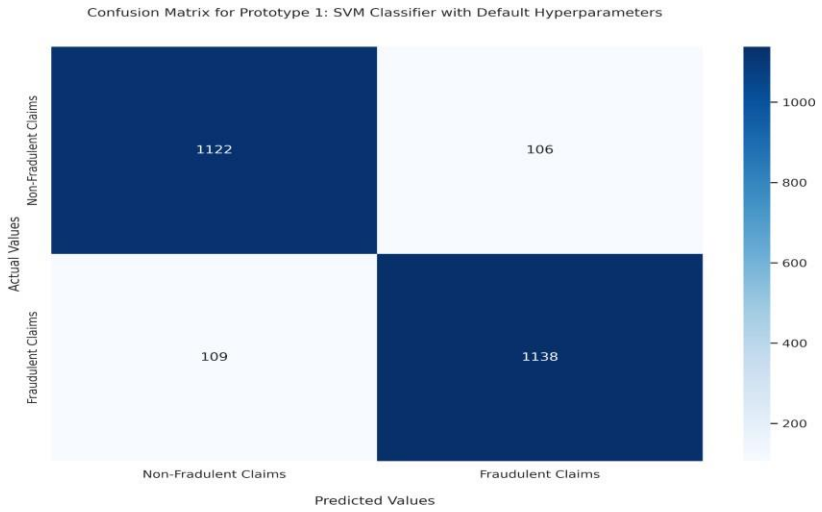

Figure 4.8: Confusion matrix for Model One

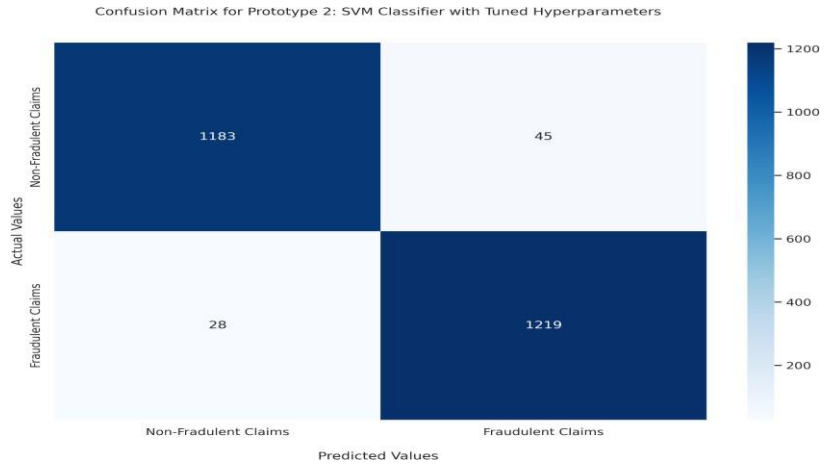

Figure 4.9: Confusion matrix for Model Two

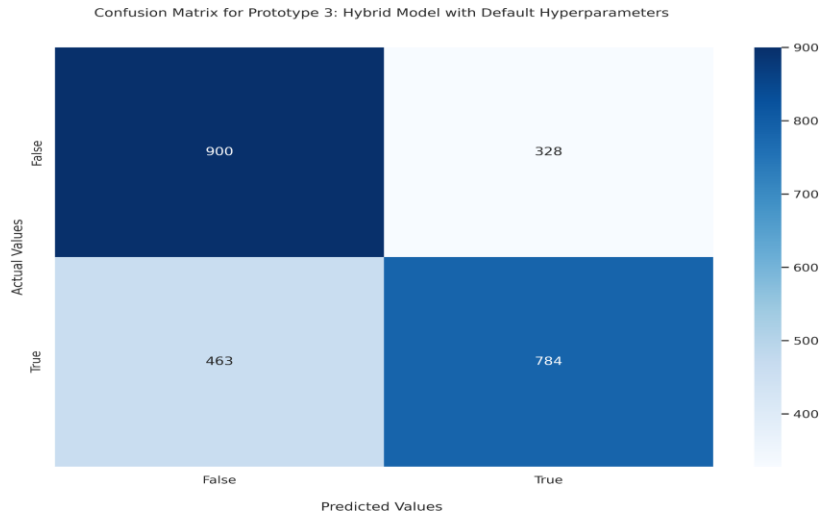

Figure 4.10: Confusion matrix for Model Three

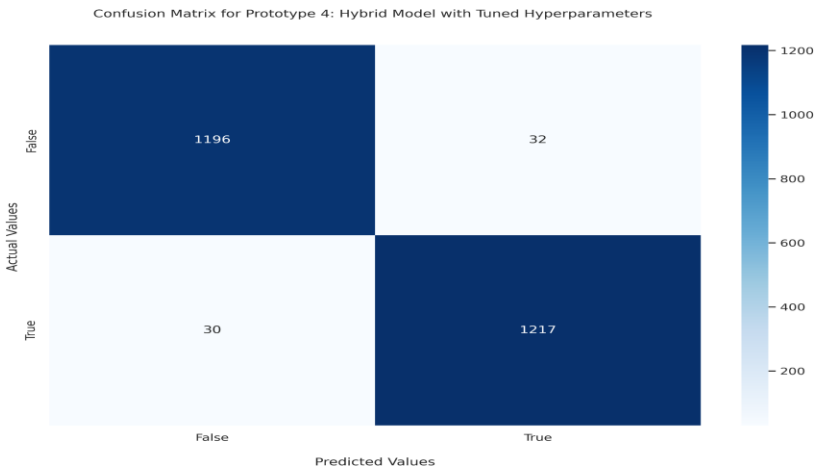

Figure 4.11: Confusion matrix for Model four

4.4.2.4 Classification Report

Further to the classification accuracy and confusion matrix, we examined other performance metrics such as precision, recall, and the F1 score to gauge the improvements of the four iterations of our model. Precision-measured the proportion of positively predicted cases against all predicted positive cases.

$$Precision = \frac{TP}{(TP + FP)} \quad \ldots(1)$$

Recall measured the percentage of True Positives against all actual positive cases.

$$Recall = \frac{TP}{(TP + FN)} \quad \ldots(2)$$

The F1 score evaluated the mean of precision and recall providing a balanced measure of the model's performance.

$$F1\ Score\ = \frac{2 * (Precision * Recall)}{(Precision + Recall)} \quad \ldots(3)$$

The table 4.5 below shows the classification reports of the algorithm.

Table 4.5: Summary of the classification report on the algorithms

	Precision:	**Recall:**	**F-Score:**
Model one	91.48%	91.26%	91.37%
Model two	96.44%	97.75%	97.09%
Model three	70.33%	62.55%	66.21%
Model four	97.44%	97.59%	97.52%

4.5 Discussion

While comparing the performance of the four prototypes, the hybrid model with optimized hyperparameters performed better than the other three prototypes in the classification of fraudulent claim transactions. Prototype 4 recorded the highest accuracy, precision, and F-Score among the four models. However, prototype 2 recorded the highest recall score at 97.75%. Bauder et al. (2017) posit that hybrid machine learning models have the potential to outperform a single algorithm due to their robust nature, adaptability, complementary strengths, and fusion of decisions.

The introduction of hyperparameter tuning in model 1 and model 3 improved the accuracy of the base models. Hyperparameter tuning is selecting optimal parameters for a machine learning model. Hyperparameters control the behavior and influence the model's performance to find the best combination of parameters that will lead to the best performance of the model on a specific dataset. The study adopted the grid search approach, which predefined values for each parameter. Model 2 and Model 4 exhaustively evaluated all possible combinations of values defined in the grid set and returned the best parameter values for selection for each grid entry. Bergstra & Bengio (2012) note that grid search is computationally expensive for large search spaces and grid entries. In our study, the grid search hyperparameter tuning from Model 1 to Model 2 runs for approximately X seconds, while that from Model 3 to Model 4 runs for Y seconds. It is prudent to note that the grid set for the SVM model consists of AB para while that of the hybrid model consists of Z parameters.

The iteration from Model 3 to Model 4 introduced Principal Component Analysis to the pipeline before the k-Means algorithm step aimed at dimensionality reduction. Dimensionality reduction is the feature reduction process while preserving as much relevant information as possible to mitigate overfitting, noise reduction and improve the computational efficiency of the model. While there is no predefined cutoff for the number of components to be used in the PCA, Abdi & Williams (2010) suggest that the number of components selected should explain a high percentage of the total variance in the dataset. The introduction of PCA to Model 4 increased the overall rise in the performance metrics.

4.6 Conclusion

Based on the benchmark results of the SVM and in comparison with the hybrid models, we note that the models with tuned hyperparameters scored better than those with the default parameters. Model 4 has the best accuracy, precision, and F1 scores in this case. Model 2 came in second with the best overall recall but second in accuracy, precision, and F1 scores. Model 1 was ranked third, with all performance measures being the third best. Model 3 was ranked in the fourth position. The hybrid classification model that uses both K-Means and SVM recorded a slight improvement in the classification of fraudulent and genuine claims compared to the classification of a single SVM model. Therefore, we can conclude that the hybrid machine learning model is more effective in detecting medical insurance claims compared to the lone SVM model.

CHAPTER FIVE

CONCLUSION, LIMITATIONS AND RECOMMENDATIONS

5.1 Introduction

Machine learning and data mining in detecting and preventing fraudulent insurance claims is a field gathering pace and attention from the research community. Researchers have demonstrated the efficacy and applicability of using supervised, unsupervised, and, more recently, a combination of both to detect fraud in insurance claims. This research demonstrated that machine learning algorithms can indeed be used to detect insurance fraud claims. Moreover, the study noted that hybrid machine-learning algorithms could outperform single machine-learning algorithms in classifying insurance fraud. The improved performance, however, requires extra computation to optimize the hyperparameters of the different machine-learning models. The rationale of this chapter is to give a conclusive summary of the study findings and how they have addressed the formulated research hypothesis. The chapter also discusses the limitations identified in the study and recommendations based on the study findings and related literature.

5.2 Conclusion based on the research objectives.

The study addressed the formulated hypothesis by thoroughly investigating the prevalence of insurance fraud, particularly in the medical insurance sector, and recognizing the critical need for robust fraud detection mechanisms. It was established that traditional methods of fraud detection, relying on domain expert analysis and rule-

based systems, have limitations in effectively addressing the evolving landscape of insurance fraud. To bridge this gap, the research proposed a hybrid machine learning approach, integrating the strengths of both supervised (SVM) and unsupervised (K-Means) algorithms. This hybrid model was designed to classify claims into fraudulent or genuine categories, as well as identify new fraud patterns through clustering. The aim was to enhance the overall performance of fraud detection, overcoming the weaknesses inherent in singular use of either approach. The study's objectives were carefully outlined, focusing on understanding the occurrence of medical insurance fraud, exploring existing machine learning techniques, developing, and evaluating the hybrid model's performance.

5.2.1 Investigate how medical insurance fraud occurs.

On the first objective that sought to understand the occurrence of medical insurance fraud, the research uncovered critical insights including fraudulent activities that manifest through various deceptive practices such as misrepresentation of medical services, fictitious claims, and identity theft. Matloob et al. (2020) categorize it into service-availing and service-providing patterns, showcasing deceptive practices by both policyholders and medical professionals. This aligns with our findings, where we observed these fraudulent activities to be prevalent in the landscape of medical insurance fraud. Additionally, Waghade & Karandikar (2018) further classify fraud based on the parties involved, distinguishing between service provider fraud, insurance subscriber fraud, insurance provider fraud, and conspiracy fraud. These classifications shed light on the multifaceted nature of fraud in the healthcare industry, providing a

comprehensive understanding of the deceptive practices that our research aims to combat.

5.2.2 Investigate the current machine learning techniques used in fraud detection.

The study also shed light into the current machine learning techniques used for fraud detection by medical insurance firms. Traditional machine learning approaches, particularly SVM, constitute a cornerstone in the various techniques utilized. However, our research extends beyond conventional methods, advocating for a hybrid model that combines K-Means clustering with SVM for enhanced accuracy and efficacy. The literature review underscores the evolution of fraud detection techniques. It traces the shift from traditional rule-based methods, constrained by the correctness of predefined rules, to the adoption of data mining algorithms, which offer increased efficiency and higher claim analysis throughput (Ai et al., 2018). Our research aligns with this progression, advocating for a hybrid machine learning approach that combines supervised and unsupervised methods. This integration capitalizes on the strengths of both approaches, resulting in improved accuracy and efficiency in detecting fraudulent medical insurance claims. This not only addresses the limitations of previous rule-based systems but also brings the field in line with modern data-driven methodologies.

5.2.3 Develop a hybrid machine learning model to detect fraud in medical insurance claims.

The research journey culminated in the development of a novel hybrid machine learning model. Leveraging a dual approach involving extensive data cleaning and feature engineering, our model achieved superior performance. Notably, the integration of Principal Component Analysis (PCA) for dimensionality reduction in the final iteration yielded a marked improvement in detection capabilities. Moreover, the literature review emphasizes the importance of data mining approaches, including clustering algorithms like K-Means, in healthcare fraud detection. Our research aligns with this by incorporating K-Means clustering as a component of the hybrid model, demonstrating its effectiveness in dimensionality reduction.

5.2.4 Evaluate the performance of the hybrid model to detect fraud and compare it with a lone SVM model.

On the effectiveness of hybrid learning models in detecting medical insurance claims compared to a lone SVM model, the comparative evaluation of our models revealed compelling outcomes. For starters, the literature review provides valuable insights into the strengths and weaknesses of both supervised (SVM) and unsupervised (clustering) methods. This knowledge base supported our comparative evaluation, where we demonstrated the superior performance of the hybrid machine learning model over the lone SVM model. The hybrid machine learning model, fine-tuned through hyperparameter optimization, emerged as the most effective in detecting fraudulent medical insurance claims. This model exhibited superior accuracy, precision, and F1

score, reaffirming the efficacy of the hybrid approach. Model 2, with optimized SVM hyperparameters, demonstrated exceptional recall rates, emphasizing the significance of parameter optimization. As such, our findings highlighted notable improvements in accuracy and precision, underscoring the significance of combining these approaches. This validation through empirical results solidifies the efficacy of our proposed hybrid model and reinforces the importance of leveraging diverse machine learning techniques in medical insurance fraud detection.

In conclusion, our research underscores the pivotal role of hybrid machine learning models in bolstering fraud detection in the medical insurance domain. By addressing each research question, we have provided a comprehensive framework that not only illuminates the intricacies of medical insurance fraud but also offers a robust solution to combat this prevalent issue. This study represents a significant stride towards a more secure and resilient medical insurance landscape.

5.3 Limitations of the Study

Specificity of Dataset: The study relies on historical claim data from a single health insurance firm in Kenya. This may limit the applicability of the developed hybrid model to other settings with different demographic and healthcare system characteristics.

Resource Intensiveness: Implementing a hybrid machine learning model, while promising, may require significant computational resources and expertise. This could

potentially limit its applicability in resource-constrained environments or for smaller insurance firms with limited technical capabilities.

Model Adaptability and Generalization: While the hybrid model shows effectiveness in the context of this study, its adaptability to new and evolving fraud patterns may require ongoing human intervention and model updates. The study does not address the long-term adaptability of the model to keep pace with emerging fraud tactics.

Availability of Historical Data: The effectiveness of the hybrid model is contingent on the availability of comprehensive and accurate historical data. In cases where such data is limited or incomplete, the model's performance may be adversely affected.

5.4 Recommendation

This section highlights various recommendations for future research and practice in the field of medical insurance fraud detection based on the study findings and limitations.

5.4.1 Recommendations for Future Practice.

Continuous Monitoring and Adaptation: The study's findings can be extended to the existing fraud detection models in the insurance industry with added accuracy by using singular classification algorithms. The study answers the question of the performance of the hybrid model in fraud detection. The study recommends an integrated approach with the model’s prediction capabilities and core applications to

detect real-time fraud. This can be achieved using Application Programmable Interfaces (APIs) to get the classification rating based on the dataset's features.

Collaboration and Information Sharing: Foster collaboration among insurance firms, industry associations, and regulatory bodies to share insights and best practices in fraud detection. This can lead to a collective effort in combating fraud across the industry.

Education and Training: Provide comprehensive training to fraud analysts, investigators, and claims processors on the latest fraud detection techniques and technologies. This empowers them to effectively identify and respond to fraudulent activities.

5.4.2 Recommendations for Future Research

Cross-Industry Insights: While the developed model recorded an accuracy rate of 97.49%, further research needs to be conducted on improving the computational and speed performance of tuning hyperparameters in a hybrid machine-learning model. The study adopted grid search cross validation which exhaustively fits the parameter set. The study can be validated against other medical insurance firms to revalidate the outputs and reinforce learning.

Longitudinal Studies: Conduct longitudinal studies to track the evolution of fraud patterns over time. This can provide valuable insights into emerging trends and the effectiveness of fraud detection measures.

Global Perspectives: Consider conducting comparative studies using hybrid machine learning models across different regions or countries to understand how fraud patterns vary in different healthcare systems and regulatory environments.

References

Abdallah, A., Maarof, M., & Zainal, A. (2016). Fraud Detection System: A Survey. *Journal of Network and Computer Applications*, 90-113.

Abdi, H., & Williams, L. (2010). Principal component analysis. *Wiley Interdisciplinary Reviews: Computational Statistics 2(4)*, 433-459.

Ai, J., Lieberthal, R., Skyla, S., & Wojciechowski, R. (2018). Examining Predictive Modeling–Based Approaches to Characterizing Health Care Fraud. *Society Of Actuaries*.

Altman, N. S., & Krzywinski, M. (2017). Points of Significance: Classification Evaluation. *Nature Methods, 14*(8), 755-756.

Association of Certified Fraud Examiners. (2019). *Insurance Fraud Handbook.* Association of Certified Fraud Examiners, Inc.

Association of Kenya Insurers. (2020). *2020 Insurance Industry Report.* Nairobi: Association of Kenya Insurers.

Association of Kenya Insurers. (2021). *Information Paper on Insurance Fraud.* Nairobi: Association of Kenya Insurers.

Bauder, R., Khoshgoftaar, T., & Seliya, N. (2017). A Survey on the state of healthcare upcoding fraud analysis and detection. *Health Services & Outcomes Research*, 31-55.

Bergstra, J., & Bengio, Y. (2012). Random Search for Hyper-Parameter Optimization. *Journal of Machine Learning Research*, 281-305.

Carcillo, F., Le Borgne, Y.-A., Caelen, O., Kessaci, Y., Obleb, F., & Bontempi, G. (2021, May). Combining Unsupervised and Supervised Learning in Credit Card Fraud. *Business Analytics Emerging Trends and Challenges, 557*, 317-331.

Dou, Y., & Xiong, H. (2017). Research on Recognition of Medical Insurance Fraud Based on Modified Support Research on Recognition of Medical Insurance Fraud Based on Modified Support. *International Conference on Computer Technology, Electronics and Communication*, (pp. 1021-1025).

Enders, C. K. (2010). *Applied Missing Data Analysis.* Guilford Press.

Gupta, R. Y., Mudigonda, S. S., & Baruah, P. K. (2021, March). A Comparative Study of Using Various Machine Learning and Deep Learning-Based Fraud

Detection Models For Universal Health Coverage. *International Journal of Engineering Trends and Technology*, 96-102.

Hanafy, M., & Ming, R. (2021). Using Machine Learning Models To Compare Various Resampling Methods In Predicting Insurance Fraud. *Journal of Theoretical and Applied Information Technology, 99*(12), 2819-2833.

Haraty, R., Dimishkieh, M., & Masud, M. (2015). An Enhanced k-Means Clustering Algorithm for Pattern Discovery in Healthcare Data. *International Journal of Distributed Sensor Networks, 2015*, 1-11. doi:http://dx.doi.org/10.1155/2015/615740

Joudaki, H., Rashidian, A., Minaei-Bidgoli, B., Mahmoodi, M., Geraili, B., Nasiri, M., & Arab, M. (2015). Using Data Mining to Detect Health Care Fraud and Abuse: A Review of Literature. *Global Journal of Health Science*, 194-202.

Kose, I., Gokturk, M., & Kilic, K. (2015). An interactive machine-learning-based electronic fraud and abuse detection system in healthcare insurance. *Applied Soft Computing Journal, 36*, 283–299. doi:https://doi.org/10.1016/j.asoc.2015.07.018

Larnyo, E., Dai, B., Udimal, T. B., & Chen, W. (2018). Detecting and Combating Fraudulent Health Insurance Claims Using ANN. *Journal of Health, Medicine and Nursing, 56*.

Lawand, S., & Kulkarni, U. (2019). Survey on Fraud Prediction for an Application Using Data Mining. *International Journal of Emerging Technologies and Innovative Research, 6*(6), 209-212. doi:http://doi.one/10.1729/Journal.22988

Matloob, I., & Khan, S. (2019). A Framework For Fraud Detection in Government Supported National Healthcare Programs. *Electronics, Computers and Artificial Intelligence, ECAI 2019.* Romania.

Matloob, I., Khan, S., ur Rahman, H., & Hussain, F. (2020). Medical Health Benefit Management System for Real-Time Notification of Fraud using Historical Medical Records. *Applied Sciences, 10(15)*. doi:https://doi.org/10.3390/app10155144

Naik, J., & Laxminarayana, A. (2017). Designing Hybrid Model for Fraud Detection in Insurance. *National Conference On Advances In Computational Biology, Communication, And Data Analytics* , 24-30.

Ogbuabor, G., & Ugwoke, F. (2018). Clustering Algorithm for a Healthcare Dataset Using Silhouette Score Value. *International Journal of Computer Science & Information Technology, 10*(2), 27-37.

Raj, P., & Lin, J. (2020). The Digital Twin Paradigm for Smarter Systems and Environments: The Industry Use Cases, Volume 117 1st Edition. *In Advances in Computers*, 1-34.

Rawte, V., & Anuradha, G. (2015). Fraud detection in health insurance using data mining techniques. *2015 International Conference on Communication, Information & Computer Technology(ICCICT).*

Schröer, C., Kruse, F., & Gómez, J. (2021). A Systematic Literature Review on Applying CRISP-DM Process Model. *Procedia Computer Science*, 526-534.

Segal, S. Y. (2016). Accounting frauds - review of advanced technologies to detect and. *Economics and Business Review*, 45-64.

Waghade, S. S., & Karandikar, A. (2018). *A Comprehensive Study of Healthcare Fraud Detection based on Machine Learning.* Nagpur: International Journal of Applied Engineering Research. Retrieved from https://www.ripublication.com/ijaer18/ijaerv13n6_140.pdf

Wakoli, L., Orto, A., & Mageto, S. (2014). Application of The K-Means Clustering Algorithm In Medical Claims Fraud / Abuse Algorithm In Medical Claims Fraud / Abuse Detection. *International Journal of Application or Innovation in Engineering & Management, 3*(7), 142-151.

Yuan. (2014). Handling missing values in data mining: a survey. Computer Science and Information Systems. *11(1)*, 103-127.

Zhang, C., Xiao, X., & Wu, C. (2020). Medical Fraud and Abuse Detection System Based on Machine Learning. *International Journal of Environmental Research and Public Health, 17*(7265), 1-11.

Zhang, Y., & Ma, S. (2020). *Ensemble Machine Learning: Methods and Applications.* Springer.

Zhou, S., & Zhang, R. (2020). A Novel Method for Mining Abnormal Expenses in Social Medical Insurance. *International IOT, Electronics and Mechatronics Conference, Proceedings.* Institute of Electrical and Electronics Engineers Inc. doi:https://doi.org/10.1109/IEMTRONICS51293.2020.9216354

Zhou, S., He, J., Yang, H., Chen, D., & Zhang, R. (2020). Big Data-Driven Abnormal Behavior Detection in Healthcare Based on Association Rules. *IEEE Access*, 129002–129011. doi:https://doi.org/10.1109/ACCESS.2020.3009006

APPENDICES

APPENDIX I: RESEARCH AUTHORIZATION

KENYATTA UNIVERSITY
GRADUATE SCHOOL

E-mail: dean-graduate@ku.ac.ke
Website: www.ku.ac.ke

P.O. Box 43844, 00100
NAIROBI, KENYA
Tel. 8710901 Ext. 57530

Our Ref: J57/26173/2019

DATE: 29th August, 2022

Director General,
National Commission for Science, Technology
and Innovation
P.O. Box 30623-00100
NAIROBI

Dear Sir/Madam,

RE: RESEARCH AUTHORIZATION FOR BRIAN NDIRANGU MUTHURA – REG. NO. J57/26173/2019.

I write to introduce **Brian Ndirangu Muthura** who is a Postgraduate Student of this University. The student is registered for M.Sc degree programme in the Department of Computing and Information Technology.

Brian intends to conduct research for a M.Sc Project Proposal entitled, "A Hybrid Model for Detecting Insurance Fraud Using K Means and Support Vector Machine Algorithms".

Any assistance given will be highly appreciated.

Yours faithfully,

PROF. ELISHIBA KIMANI
DEAN, GRADUATE SCHOOL

EM/ben

APPENDIX II: RESEARCH PERMIT

REPUBLIC OF KENYA

NATIONAL COMMISSION FOR SCIENCE,TECHNOLOGY & INNOVATION

Ref No: 892696

Date of Issue: 22/October/2024

RESEARCH LICENSE

This is to Certify that Mr.. Brian Mdrunge Muthura of Kenyatta University, has been licensed to conduct research as per the provision of the Science, Technology and Innovation Act, 2013 (Rev 2014) in Nairobi on the topic: A HYBRID MODEL FOR DETECTING INSURANCE FRAUD USING K MEANS AND SUPPORT VECTOR MACHINE ALGORITHMS for the period ending : 22/October/2025.

License No: NACOSTI/P/24/41001

892696

Applicant Identification Number

Director General

NATIONAL COMMISSION FOR SCIENCE,TECHNOLOGY & INNOVATION

Verification QR Code

NOTE: This is a computer generated License. To verify the authenticity of this document, Scan the QR Code using QR scanner application.

See overleaf for conditions