NOVA IMS Information Management School

MDSAA

Master Degree Program in

Data Science and Advanced Analytics

ADVANCED ANOMALY DETECTION MODELS FOR UNCOVERING FRAUDULENT HEALTHCARE INSURANCE PROVIDERS

Diogo Cordoeiro Lamy Carneiro

Master Thesis

presented as partial requirement for obtaining a Master's Degree in Data Science and Advanced Analytics

NOVA Information Management School

Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

ADVANCED ANOMALY DETECTION MODELS FOR UNCOVERING FRAUDULENT HEALTHCARE INSURANCE PROVIDERS

by

Diogo Cordoeiro Lamy Carneiro

Master Thesis presented as partial requirement for obtaining the Master's degree in Data Science and Advanced Analytics, with a specialization in Business Analytics

Supervised by

Prof. Dr. Jorge Miguel Ventura Bravo, NOVA IMS & Université Paris-Dauphine PSL

November, 2024

STATEMENT OF INTEGRITY

I hereby declare having conducted this academic work with integrity. I confirm that I have not used plagiarism, any form of undue use of information or falsification of results along the process leading to its elaboration. I further declare that I have fully acknowledged the Rules of Conduct and Code of Honor from the NOVA Information Management School.

Lisboa, 30 November 2024

Diogo Lamy

ACKNOWLEDGEMENTS

This work is a dissertation conducted in collaboration with Multicare as a partial requirement for obtaining a Master's degree in Data Science and Advanced Analytics.

First, I would like to express my gratitude to Multicare's GAF and GAC teams for their collaboration throughout the course of this dissertation. I am especially grateful to Nuno Rua, the head of the GAF team, for placing his trust in the potential of my work and providing me with the opportunity to carry out my dissertation at Multicare. His confidence was crucial to developing the direction of this project.

Within the GAF team, I extend thanks to Sandra Inácio, whose deep business knowledge of Multicare and expertise in fraud detection provided an invaluable foundation for my research. Her insights into the nuances of fraud issues at Multicare enriched the relevance of this study.

A special note of appreciation goes to Marli Ferreira, responsible for the statistical studies and clinical monitoring functional area in GAC team, for her significant contributions to the technical aspects of this work. Her extensive experience in statistical studies, coupled with her thoughtful guidance on data specific challenges, was crucial in complex methodological decisions and achieving meaningful results.

I am also indebted to my academic supervisor, Professor Jorge Bravo, whose expertise, mentorship, and encouragement have been a guiding light throughout this journey. His critical feedback and constructive suggestions have not only enhanced the quality of this dissertation but have also profoundly shaped my academic growth.

Lastly, I wish to thank everyone at Multicare and in my academic circle who, directly or indirectly, supported and inspired me to complete this work. This accomplishment would not have been possible without the support of all these people.

ABSTRACT

Fraud detection in healthcare insurance is a critical area of research due to its implications for financial sustainability and operational efficiency. This dissertation addresses the challenge of detecting anomalous billing behaviours in healthcare providers in the context of Multicare, a leading health insurance entity in Portugal. Although traditional methods depend on standard rule-based systems, they often fail to capture complex fraud patterns in the dynamic healthcare landscape. This study fills a significant gap by applying the Isolation Forest algorithm, an unsupervised machine learning model, to identify potential fraudulent providers. The research evaluates how well this advanced anomaly detection technique aligns with domain-specific knowledge, particularly from Multicare's Actuarial and Technical Control (GAC) and Fraud Investigation (GAF) teams. The study incorporates detailed data preprocessing steps, feature engineering, and the integration of business insights to guarantee that the model results are precise and practical. Key findings demonstrate that the Isolation Forest model successfully detects anomalous billing practices in eight medical specialties, with metrics indicating a high recall rate and robust alignment with GAF fraudulent providers classifications. This alignment underscores the model ability to enhance Multicare's existing fraud detection systems by identifying deviations that traditional approaches might overlook. Furthermore, the results highlight the importance of interdisciplinary collaboration, using technical and clinical expertise to refine detection processes. This research contributes to the broader field of healthcare fraud detection by proposing a scalable, data-driven approach that integrates machine learning with domain-specific insights. It lays the groundwork for future studies to explore inpatient billing behaviours, clustering techniques, and more comprehensive analyses of fraudulent patterns, opening the path for more effective and comprehensive fraud detection frameworks in the insurance sector.

KEYWORDS

Anomaly Detection; Isolation Forest; Fraud; Providers

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS AND ACRONYMS

GAC	Multicare's Actuarial and Technical Control Team
GAF	Multicare's Fraud Investigation Team
FWA	Fraud, Waste and Abuse
kNN	k-Nearest-Neighbour
LSTM	Long Short-Term Memory
MLP	Multi-Layer Perceptron
SOM	Self-Organizing Map
LOF	Local Outlier Factor
LRD	Local Reachability Density
CRISP-DM	Cross-Industry Process for Data Mining
GDP	Gross Domestic Product
KPIs	Key Performance Indicators
SAS	Statistical Analysis System
VAP	Total Expenditure
PCA	Principal Component Analysis
iForest	Isolation Forest
iTrees	Isolation Trees
TP	True Positives
FP	False Positives
TN	True Negatives
FN	False Negatives
IF	Isolation Forest

1. INTRODUCTION

Detecting fraud within the realm of health insurance continues to be a significant challenge and an essential subject of research, especially for entities like Multicare that operate within the complex health insurance system. This dissertation investigates how anomaly detection methods can be applied to uncover fraudulent billing practices, aiming to identify fraudulent healthcare providers within the Multicare network. The main research question addressed is: In what way can advanced anomaly detection methods successfully recognize unusual billing patterns to detect fraudulent healthcare providers?

The importance of this study derives from its capability to tackle the universal and expensive problem of detecting healthcare insurance fraud, a sector vital not only for the stability of insurers like Multicare but also for the wellbeing of society at large. Fraudulent actions in healthcare are often associated to the term FWA (Fraud, Waste and Abuse), which contribute significantly to global healthcare costs, intensify financial burdens, raise premiums, and undermine trust in healthcare institutions. It is estimated that fraud drains the healthcare sector of billions each year (around 10% of total healthcare spending worldwide), affecting the proper allocation of resources and the quality of patient care (Rosenbaum, Lopez, & Stifler, 2009; Timofeyev & Jakovljevic, 2022; Viaene & Dedene, 2004).

This research seeks to identify fraudulent billing behaviours more accurately using an anomaly detection model. Unlike conventional methods, which often lack the precision for dealing with the nuances of healthcare data, these models are designed to uncover delicate patterns of deviations that indicate fraud (Viaene & Dedene, 2004). This is especially relevant to Multicare, where enhancing fraud detection measures can reduce financial losses and boost operational efficiency (Timofeyev & Jakovljevic, 2022). Moreover, incorporating specialized knowledge into the modelling refines the applicability and effectiveness of these techniques by aligning them with the practical aspects of healthcare billing.

This research is compelling not just for its immediate advantages to the organization but also for its wider implications. It demonstrates how innovative data-driven methods can be customized for complicated real-world challenges, providing a scalable solution that can be adopted by other insurers and healthcare systems. By tackling these issues, the research adds to a growing body of work aimed at minimizing fraud while promoting accountability and transparency in global healthcare systems.

The existing literature highlights the considerable promise of anomaly detection techniques in uncovering healthcare fraud. Shan, Murray, & Sutinen (2009) illustrated the effectiveness of density-based outlier detection methods, such as the local outlier factor (LOF), in spotting improper billing patterns by utilizing domain knowledge and compliance records. In a similar approach, Du & Yu (2023) showcased the use of the Isolation Forest algorithm to manage

diverse and high-dimensional healthcare datasets, enhancing the efficiency of anomaly detection while decreasing the necessity for manual audits.

Nonetheless, challenges remain, particularly in customizing models to fit domain-specific billing practices. Shan, Murray, & Sutinen, (2009) stressed the importance of aligning model results with specialized compliance histories, a gap that is frequently overlooked in more general anomaly detection strategies. Du & Yu, (2023) also recognized that although Isolation Forests increase detection accuracy, they are heavily dependent on well pre-processed data and may struggle when domain-specific details are missing. This dissertation tackles these challenges by improving data preprocessing and model setup to more accurately represent the realities of healthcare fraud detection. In particular, the research incorporates domain-specific business rules into the anomaly detection process, ensuring that models not only detect anomalies but do so in a manner that aligns with practical, real-world auditing and compliance standards. This work advances the development of more effective and context-sensitive fraud detection systems in health insurance by closing the gap between theoretical anomaly detection models and their practical utilization.

The novelty of this work is found in its comprehensive approach to adapting advanced anomaly detection models to match organizational needs and domain-specific knowledge. Unlike conventional methods that often function independently of practical business and clinical contexts, this study incorporates insights from Multicare's Actuarial and Technical Control (GAC) and Fraud Investigation (GAF) teams to build a model suited to real-world demands.

By closing the gap between data-driven strategies and domain expertise, the research guarantees that the models are both technically robust and contextually relevant. This work is distinguished by its dual focus on assessing model performance and ensuring practical applicability. Through systematic evaluations across various feature configurations and a detailed assessment of potential anomalous providers, the study optimizes its detection approach to align with healthcare providers detailed billing practices and Multicare's organizational objectives. Furthermore, the continuous feedback from domain experts during the evaluation stage elevates the model's sensitivity and specificity, ensuring it identifies fraudulent activities more precisely and efficiently.

This methodology merges advanced anomaly detection techniques with domain knowledge to fill specific holes in past methods. It boosts the practical effectiveness of fraud detection systems by spotting and removing unrelated variables, such as non-contributory medical acts, and refining model configurations based on expert feedback. This guarantees a more accurate and efficient detection mechanism, ultimately reducing financial losses and improving the accountability of healthcare providers within Multicare's network.

The findings of this research illustrate the considerable effectiveness of anomaly detection models, especially the Isolation Forest model, in pinpointing unusual billing activities across

different medical specialties. Important metrics like accuracy and recall prove the model's strong ability to identify potentially fraudulent providers. This study focuses on its main objective of boosting and refining the current GAC model by incorporating insights from Multicare's GAF team. The results underscore the necessity of collaboration between technical and clinical expertise to fill existing gaps in fraud detection and ensure consistency with real-world billing practices. This research offers significant promise for improving the GAC model's efficiency in spotting abnormal billing patterns, thereby elevating organizational processes and protecting financial assets. Moreover, the study emphasizes the wider impact of furthering anomaly detection models as an essential tool for financial sustainability in the health insurance sector.

The outline of the remainder of the thesis is as follows. Section 2 provides a comprehensive review of the existing literature, exploring the theoretical perspectives of healthcare insurance fraud and anomaly detection methodologies. Section 3 describes the methodology, illustrating the application of the CRISP-DM framework and the specific steps taken to prepare and analyse the data. Section 4 presents the evaluation of the anomaly detection models, highlighting key results and their alignment with the business objectives. Section 5 offers a conclusion, summarizing the findings and their implications, followed by Section 6, which discusses the limitations of this research and offers recommendations for future work.

2. LITERATURE REVIEW

2.1 THEORETICAL VIEW ON HEALTHCARE INSURANCE FRAUD

Fraudulent activities in health insurance represent a significant challenge to the global pursuit of universal health coverage and the improvement of healthcare outcomes. The effectiveness of healthcare systems varies among countries, with corruption and fraud as fundamental factors contributing to government inefficiencies and adverse effects on local healthcare. When healthcare professionals and managers intentionally deviate from ethical practices, it compromises the integrity of health systems. This results in substantial financial losses for organizations and obstructs the fundamental ambition of achieving universal health coverage (Timofeyev & Jakovljevic, 2022).

Within the insurance sector, the relationship between insurers and policyholders is established with the highest level of honesty and the duty to share crucial details. Fraud in insurance emerges when there is a deliberate distortion of facts to mislead and obtain unapproved advantages, frequently arising from an imbalance of information that usually benefits insurers more than policyholders.

Within the realm of health insurance, fraudulent activities may occur when people mislead insurers to receive claim payouts, they are not eligible for or exhibit a careless lack of concern for the accuracy of their claims. This unethical conduct can also affect group sponsors and members, as it frequently involves altering the claims process for personal financial gain. These strategies can encompass bribery, kickbacks, and other unlawful schemes (Rosenbaum, Lopez, & Stifler, 2009).

Insurance fraud encompasses many different types of schemes, such as claim fraud, premium fraud, third-party fraud, and insider/agent fraud. Insurance fraud can be categorized in various ways, including internal versus external, underwriting versus claims, and soft versus hard fraud. Internal fraud involves misconduct by individuals within the insurance industry, like insurers, agents, and managers, whereas external fraud is committed by applicants, policyholders, or claimants. Underwriting fraud takes place during the policy issuance stage, whereas claims fraud generally includes the intentional exaggeration or fabrication of claims. The distinction between soft and hard fraud lies in that soft fraud usually arises from opportunistic actions, whereas hard fraud involves carefully planned schemes (Viaene & Dedene, 2004). Actuaries play a critical role in assisting insurers and regulators in identifying, quantifying, and ultimately reducing the extent to which insurance fraud occurs (Bravo 2020, 2022; Bravo et al. 2021, 2022, 2023; Bravo & El Mekkaoui, 2018). Many regulatory authorities require insurance companies to submit anti-fraud plans.

This dissertation examines the connection between underwriting and claims associated with healthcare insurance fraud. By thoroughly analysing these subjects, the aim is to offer valuable insights for comprehending and preventing healthcare insurance fraud.

2.2 ANOMALY DETECTION

Irregularities can impact the efficacy of a machine learning model due to mistakes in data production or undefined attributes, resulting in false predictions and diminished credibility in data-driven design applications (Yousefpour, Shishehbor, Foumani, & Bostanabad).

2.2.1 SUPERVISED, SEMI-SUPERVISED AND UNSUPERVISED LEARNING

Anomaly detection can be tackled using different learning strategies, including supervised, semi-supervised, and unsupervised methods. In supervised anomaly detection, models are trained on datasets that have both normal and anomalous examples labelled. This learning process enables the model to distinguish between these categories by understanding the features and patterns it has absorbed. After training, the model can be applied to detect anomalies in new, unseen data.

On the other hand, semi-supervised methods focus on training a model exclusively with normal samples. As a result, the model becomes skilled at recognizing common examples, without having been exposed to labelled abnormal cases during its training phase. After being developed, it is employed to detect anomalous occurrences in new data.

Conversely, unsupervised anomaly detection entails utilizing models on datasets that might contain both normal and abnormal samples. The objective is to identify cases that differ from the norm without having previous information about which cases are typical. Unlike the other two approaches, unsupervised detection does not divide its operations into separate training and testing stages and functions without any labelled anomalous samples during the learning period. This research will concentrate on methods related to unsupervised learning (Yousefpour, Shishehbor, Foumani, & Bostanabad).

2.2.2 RELATED WORK

This section of the study provides a summary of existing academic research focused on healthcare insurance fraud detection, establishing the foundation for this research focused on detecting fraudulent healthcare providers by analysing insurance billing practices. The analysis explores anomaly detection methods tailored for the healthcare insurance industry, utilizing existing literature to highlight extensive strategies for detection and their practical applications in revealing insurance fraud detection.

Anomaly detection involves identifying data patterns that diverge from the usual or are distinctive depending on the context. These patterns are often referred to as outliers or by other names based on the specific circumstances. Numerous techniques have been developed over time to identify outliers, with characteristics designed to suit data distributions and goals. Capelleveen, (2013) proposed a method to detect healthcare fraud within the Medicaid system by employing unsupervised learning techniques that specifically target outliers in provider billing activities when compared to their peers' behaviours in the industry. This approach involves developing metrics based on observed patterns in healthcare claim data,

such as the frequency of performed procedures, the nature of services provided, and the amounts claimed. The objective is to identify anomalies, as fraudulent providers often exhibit uncommon practices that differ from typical providers in their field (Capelleveen, 2013).

Capelleveen, (2013) work describes several techniques used for outlier detection, including:

- **Peer-group analysis:** This technique examines a provider's actions over a period in comparison to those of similar providers. A significant deviation in a provider's behaviour from their peers triggers an alert.
- **Break-point analysis:** This method monitors individual providers to detect sudden changes in their patterns, such as an unexpected rise in claims or services provided.
- **Cluster analysis:** Providers are grouped based on common behaviours, and those that fall outside these groups are potential outliers.
- **Single anomalies:** This method singles out specific transactions or claims that stand out from the usual pattern, often revealing rare cases that could indicate fraudulent activity.

The framework developed by Capelleveen (2013) was assessed through a case study centred on Medicaid dental claims in Arkansas. This study reviewed millions of claims and found seventeen providers with distinctly different behaviour patterns. In the end, twelve of these providers were flagged for fraudulent activities, resulting in a precision rate of 71,00%. This demonstrates the efficacy of outlier detection techniques in prioritizing suspicious cases and reducing false positives.

Yang & Hwang, (2006) presents a framework aimed at detecting healthcare fraud within the National Health Insurance system of Taiwan. It employs the peer-group analysis methodology, strengthened by unsupervised learning techniques and process mining. This framework evaluates provider behaviours against standards defined by the peer-group through organized sequences of medical events. The process mining component carefully extracts frequently observed behaviour patterns from clinical data, providing a dependable reference for expected actions in various medical scenarios. Clinical cases are illustrated through temporal graphs that illustrate the relationships and sequences of medical activities over time.

A crucial aspect of this analysis is identifying structural patterns and selecting pertinent features. Through process mining, frequently occurring patterns in the data are recognized, transforming them into features that could indicate fraudulent behaviour when they diverge from expected norms. To improve accuracy, the model reduces unnecessary information by focusing on the most relevant features using a filter-based technique. Temporal graphs are created for each clinical instance, and pattern discovery algorithms are used to identify any anomalies, highlighting provider actions that differ from expected conduct. This feature

extraction and model training method streamlines fraud detection, significantly minimizing the need for intensive manual reviews and specialized expertise. An empirical assessment of the National Health Insurance system of Taiwan data demonstrates the model's ability to detect fraudulent patterns that traditional methods might overlook, underscoring its potential as an effective and scalable solution for addressing healthcare fraud.

A different investigation conducted by He H. , Hawkins, Graco, & Yao, (2000) explores how healthcare fraud is detected within Australia's Medicare system. This research merges a genetic algorithm with the k-Nearest-Neighbour (kNN) approach, set within the same peer-group analysis outlier detection technique. By integrating these methods, the study enhances fraud detection accuracy by optimizing how kNN classifies medical practices.

The genetic algorithm is crucial as it adjusts the weights of different features, thereby boosting the model's sensitivity to fraud behaviours. Important features like frequent prescribing and numerous patient visits are highlighted, enabling the algorithm to focus on patterns associated with fraudulent behaviour. The updated kNN model evaluates each provider's actions against a benchmark established by peer practices, identifying outliers with significant deviations. Through successive optimization generations, the genetic algorithm repeatedly adjusts the features, leading to improved classification by concentrating on slight indicators of fraudulent activities. He H. , Hawkins, Graco, & Yao, (2000) validated the model using two datasets: one consisting of General Practitioners, categorizing doctors by risk levels, and another referred to as the "Doctor-Shopper" dataset, focusing on patients seeking excessive prescriptions. The optimized kNN model showed strong performance in both datasets, outperforming traditional models in detecting unusual provider activities and suspicious patient behaviours. This improvement came from using a genetic algorithm to adjust feature weights, creating a sharper distance metric that focused on key behaviours while ignoring less relevant ones. By emphasizing important characteristics, such as the number of providers visited by a patient or specific drug types, the model became more accurate and aligned with domain experts' assessments. This approach not only simplified the detection process but also reduced unnecessary complexity in the data. The study demonstrates that combining peer group analysis with an optimized kNN model is an effective way to identify healthcare fraud. The refined distance metric allowed the model to capture subtle yet meaningful deviations from normal behaviour, offering a scalable and efficient solution for fraud detection in large healthcare datasets, with applications beyond Medicare to other insurance systems.

These research initiatives demonstrate the efficacy of peer-group analysis in detecting healthcare fraud by showcasing provider actions that deviate from group standards. By assessing each provider's actions against peer benchmarks, this approach enhances the accuracy and scalability of detection. It can adjust to changing fraud strategies by employing techniques such as genetic algorithms, kNN, and unsupervised learning to improve fraud detection systems, reduce the necessity for manual interventions, and develop scalability in complex and dynamic data settings.

Both Thornton et al. (2013) and Hansson & Cedervall, (2022) offer models that employ break-point analysis for fraud detection. This approach aims to identify sudden and significant alterations in claims data, which may indicate fraudulent actions. It achieves this by pinpointing notable changes or discrepancies in billing patterns, the nature of provided services, or specific details of claims.

The multidimensional model presented by Thornton et al. (2013) provides a comprehensive approach to identifying Medicaid fraud by analysing claims data across multiple connected dimensions, such as service types, providers, patient demographics, and geographic areas. This multi-layered view facilitates the detection of existing fraud patterns as well as new, previously unnoticed schemes by examining claims thoroughly. Each dimension delivers unique insights that, when combined, reveal complex behavioural patterns among providers and beneficiaries that might go unnoticed if evaluated separately.

By systematically segmenting the data, the framework successfully highlights irregularities in various aspects of provider conduct. It employs a hierarchical system to evaluate the complexity of fraud detection. Initially, obvious fraud patterns are detected through algorithms that identify recurring claims for the same services. As complexity arises, the methodology uses more sophisticated analytics, including time series analyses, to detect unusual increases in billing hours or expensive service instances that may indicate overbilling activities. At the most advanced analytical level, the model can expose collusive fraud networks by identifying links between providers who may work together to exploit the system, such as cases where providers exchange patients to generate unnecessary claims. This approach enables a comprehensive examination of how different factors interact to produce atypical billing patterns, successfully pinpointing both individual occurrences and complex multidimensional schemes.

The outlier technique used in this framework is break-point analysis, which is particularly effective at identifying sudden alterations in claim patterns, making it especially valuable for discovering new fraud techniques. This analysis reveals significant deviations from historical billing patterns, such as unexpected increases in billing amounts, the predominance of expensive services, or changes in the types of services provided by certain providers. These irregularities are marked for further investigation, often uncovering situations where providers quickly change their billing strategies to exploit under-supervised service codes.

The study further demonstrates the model's performance by applying it to real Medicaid cases, successfully uncovering a range of fraud types. It also exposed complex collusive fraud networks by displaying unusual connections between providers who frequently refer patients to each other, indicating coordinated efforts to inflate billing amounts. Through these instances, the research underscores the model's versatility, particularly when coupled with real-time data monitoring, proving its effectiveness in identifying dynamic fraud schemes within Medicaid. The researchers concluded that this model offers a strong and scalable

solution for fraud detection that can adapt to the continuously changing environment of Medicaid fraud schemes.

Hansson & Cedervall (2022) introduced a framework for anomaly detection that successfully uses break-point analysis along with Long Short-Term Memory (LSTM) networks and autoencoders. This system is designed to identify insurance fraud within sequential claims data. They demonstrate the difficulties associated with traditional manual fraud detection methods, mainly due to the extensive and intricate nature of insurance datasets. By integrating LSTM networks within autoencoders, their unsupervised learning model can capture complex temporal patterns present in claim sequences. This allows it to accurately reconstruct standard claim patterns while showing notable reconstruction errors for abnormal or potentially fraudulent claims.

In this work, reconstruction error acts as a key metric for anomalies. When the model struggles to accurately reconstruct atypical sequences, high reconstruction errors occur, indicating possible fraud. A threshold anomaly score is then calculated based on these errors, finding sequences that exceed this threshold for further examination. The model's adaptive nature allows it to detect not only known fraudulent patterns but also new fraud schemes that deviate from historical behaviours, making it highly effective at identifying innovative fraud tactics. Overall, the LSTM auto-encoder model offers substantial support for proactive fraud management within insurance systems by providing a scalable solution for fraud detection.

In summary, the two studies highlight the effectiveness of break-point analysis for identifying healthcare and insurance fraud by recognizing sudden, unusual changes in data patterns. The research by Thornton et al. (2013) illustrates how a multidimensional data model can systematically uncover various Medicaid fraud schemes, ranging from simple duplicate claims to intricate, collaborative fraud networks. Meanwhile, Hansson & Cedervall, (2022) demonstrate the versatility of LSTM-autoencoders in the insurance sector for detecting unusual claim sequences without the need for pre-defined fraud labels. By pinpointing significant deviations from typical behaviours, these models offer powerful and scalable solutions for fraud detection in different environments.

Cluster analysis serves as a fundamental method to uncover patterns of fraud. It does this by categorizing similar behaviours and emphasizing any anomalies found within those groups.

He et al. (1997) employs a distinct method by integrating Multi-Layer Perceptron (MLP) with Kohonen's Self-Organizing Map (SOM) within a structured framework. The MLP, serving as a type of supervised neural network, is designed to initially learn the classification of healthcare provider profiles using pre-established fraud indicators. This method creates an effective baseline for detecting standard fraudulent and nonfraudulent behaviours. In contrast, SOM functions as an unsupervised approach, grouping these profiles by identifying shared behaviours, and offering valuable insight into subtle or new fraud patterns that may not align with defined categories.

The integration of MLP and SOM makes the model proficient at identifying providers with behaviours that are difficult to categorize, allowing for a closer analysis of profiles showing unusual yet potentially fraudulent activities. This layered strategy not only elevates fraud detection but also improves the model's ability to adjust to innovating fraud strategies. The research showed that incorporating SOM led to changes in the classification, moving beyond a binary notion of fraudulent and non-fraudulent to a more detailed framework capable of recognizing potential fraudulent cases. This adaptability is vital in the healthcare sector, where fraudulent schemes are constantly changing to unlock new weaknesses in billing and regulatory frameworks. By recognizing clusters that diverge from expected behaviours, the SOM component of the model acts as a proactive measure, allowing for early detection of fraud before it develops into bigger problems.

Ekina et al. (2013) introduce a Bayesian clustering technique that groups healthcare providers and beneficiaries, recognizing the connection between suspicious activities. This approach uses Markov Chain Monte Carlo methods to assess the likelihood of unusual associations or clusters within the dataset. It is particularly advantageous for detecting patterns associated with conspiracy fraud, where several providers and beneficiaries collaborate to generate fraudulent claims. By utilizing Bayesian inference, this method assesses each provider-beneficiary relationship based on its probability of deviating from standard healthcare practices. This process captures subtle nuances that may be overlooked by other models, which often concentrate on the isolated behaviours of providers or patients. Additionally, the Bayesian clustering method can incorporate knowledge through prior distributions. This detail allows healthcare fraud specialists to adjust their model assumptions according to known fraud patterns or new trends. Such adaptability is essential for identifying healthcare fraud, as schemes evolve in response to changing regulations and oversight. Moreover, the probabilistic nature of the Bayesian model allows it to effectively manage uncertainty in large datasets, improving the dependability of detection in complex, high-dimensional healthcare data. This clustering strategy not only identifies specific instances of suspicious behaviour but also reveals broader patterns of interactions between providers and beneficiaries, illuminating fraud schemes that are frequently overlooked in isolated studies.

Both studies highlight how effective cluster analysis can be in identifying fraud within healthcare by focusing on both individual behaviours and group trends. The hybrid MLP and SOM model presented by He et al. (1997) works in a different number of cases that don't fit into standard fraud categories. Meanwhile, the Bayesian clustering technique developed by Ekina et al. (2013) is successful in uncovering fraudulent activities related to collusion by examining the complex interactions between providers and beneficiaries. These methods showcase how versatile cluster analysis is in tackling the complex issues surrounding healthcare fraud, offering strong tools for quick and adaptive detection in increasingly complicated fraud scenarios.

Shan, Murray, & Sutinen, (2009) and Du & Yu, (2023) explore the single anomaly method for detecting outliers, specifically employing the Local Outlier Factor and Isolation Forest algorithms. These approaches focus on identifying fraudulent providers by examining claims that significantly differ from normal patterns, flagging such transactions as potential signs of fraud due to atypical billing activities.

The LOF method utilizes a density-based strategy to assess each billing by comparing it with its closest neighbours, thereby determining its "local density." Key calculations in this process include:

1. K-distance: measuring the distance from each entry to its k-th nearest neighbour.
2. Local Reachability Density (LRD): evaluates how densely populated a record is in relation to its neighbours.
3. Local Outlier Factor: Calculating the ratio of a record's LRD to that of its neighbours, where higher LOF values indicate potential outliers.

This methodology is utilized on a Medicare billing record dataset, examining aspects such as billing items, total services, and patient information to detect billing patterns that deviate from the standard. By focusing on the most extreme and isolated billing entries, the LOF technique enables precise identification of individual anomalies, providing insights into potential fraud cases.

The Isolation Forest algorithm discussed in the second study employs a distinct method for detecting individual anomalies. In contrast to LOF, the Isolation Forest uses a binary tree partitioning technique. This approach is efficient because outliers, or anomalous points, can be "isolated" more effectively with fewer divisions in the tree structure. When applied to billing practices in healthcare, Isolation Forest analyses factors such as billing amounts, frequency of billed services, and provider-specific patterns to detect claims that deviate significantly from standard billing behaviours. This approach excels at identifying individual billing anomalies without relying on cluster density, making it particularly accurate at uncovering subtle irregularities and potentially fraudulent patterns within large-scale billing datasets.

While utilizing distinct methods, these approaches effectively collaborate to pinpoint individual anomalies by concentrating on specific data points. The LOF method carefully assesses neighbourhood density, making it especially beneficial for structured datasets where anomalies typically appear as deviations with significant costs and frequencies. In contrast, Isolation Forest offers a versatile and density-independent approach ideal for detecting unusual patterns that might arise in complex or infrequent fraud scenarios. LOF and Isolation Forest are potent tools for identifying fraudulent providers by isolating specific billing irregularities within healthcare insurance datasets. By identifying these outliers, these methods aid in focused fraud investigations, enabling the efficient identification of high-risk providers. Incorporating these models with expert domain knowledge and supplementary fraud detection strategies can further enhance their adaptability and effectiveness in tackling

the changing patterns of fraud within healthcare insurance. LOF and Isolation Forest establish a robust foundation for fraud detection, combining accuracy in isolating fraudulent activities within high-dimensional datasets.

When examining various outlier detection strategies to identify healthcare fraud, each approach offers unique benefits, revealing patterns associated with suspicious practices between providers. The analysis of peer-groups and the use of cluster analysis are effective methods to assess the behaviour of the provider in relation to group standards. Peer-group analysis reveals deviations by comparing individual practitioners with their peers, assisting in the identification of providers whose gradual behavioural changes may indicate potential fraudulent intentions, even if they do not show obvious anomalies. On the other hand, cluster analysis, illustrated by He et al. (1997), employs both MLP and SOM to categorize providers exhibiting similar behaviours, thereby highlighting those who deviate from the norm. This method is particularly useful for uncovering collaborative fraud activities, especially in environments where providers might coordinate their efforts to improve fraudulent practices within clustered behavioural patterns.

Break-point analysis provides an important strategy for identifying sudden changes in billing patterns that may indicate potentially fraudulent activity. This technique has been applied in research such as the studies conducted by Thornton et al. (2013) and Hansson & Cedervall, (2022). In these studies, multidimensional data frameworks and LSTM-autoencoders were used to identify significant deviations in claims data. Using this method, it becomes possible to pinpoint instances where providers may quickly alter their billing behaviours, often signalling the emergence of new fraud tactics or attempts to take advantage of updated billing codes.

To detect fraudulent providers, the method known as single anomaly outlier detection is highly effective. This approach concentrates on highlighting transactions that differ from normal patterns, particularly those showing billing practices indicative of fraud. In studies, such as those by Shan, Murray, & Sutinen, (2009) and Du & Yu, (2023), single anomaly detection is utilized with models like LOF and Isolation Forest. These techniques are particularly valuable for identifying specific fraudulent actions within extensive datasets. LOF relies on a density-based concept, recognizing records with reduced density and flagging those with fewer similar neighbours as potential outliers. This strategy is advantageous for structured datasets where anomalies in billing trends or excessively high-cost services may imply fraud. On other hand, Isolation Forest functions effectively with high-dimensional data without relying on density. It randomly divides the dataset to promptly isolate points, using binary trees to spotlight unusual patterns, such as inflated treatment costs or excessively frequent services differing from usual billing norms.

The Single Anomalies approach focuses on individual claims, enabling precise identification of high-risk providers whose billing practices differ from typical patterns. Unlike methods that

assess groups, single anomalies specifically target isolated actions that may indicate fraudulent behaviour at the level of individual transactions.

While peer-group, cluster, and break-point analysis are essential for recognizing different fraud patterns, spotting individual anomalies is especially effective for targeting specific fraudulent billing practices and accurately identifying high-risk providers. This approach is particularly suitable for the unpredictable and often unique instances of provider fraud, enabling the detection of irregularities without relying on collective standards.

3. METHODOLOGY

This study will employ the CRISP-DM framework as its methodology, offering a systematic strategy for data analysis. The CRISP-DM framework comprises six essential phases: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment. By adopting this framework, structured and informed decision-making driven by data insights will be encouraged.

3.1 BUSINESS UNDERSTANDING

The crucial initial phase of any data mining project involves comprehending the project goals from a business point of view. This comprehension should then be converted into a distinct data mining issue, accompanied by an initial plan to accomplish these goals.

Data mining specialists are required to possess a comprehensive understanding of the business context to determine which data to explore and which methodology to employ. This phase of business understanding involves essential steps including the identification of business objectives, assessment of the current situation, establishment of data mining goals, and formulation of a project plan (Shearer, 2000).

3.1.1 BUSINESS OBJECTIVES

3.1.1.1 BACKGROUND

The Portuguese National Health System, established in 1979, is founded upon the Beveridge model principles, which ensure universal, comprehensive, and cost-free access to healthcare for all citizens. It functions under state supervision, enhancing individual and collective health through the prevention, diagnosis, treatment, and rehabilitation of diseases. The funding predominantly originates from taxes and user fees, with the budget being stipulated by the annual State budget, thereby reflecting the governmental influence on healthcare delivery.

Furthermore, there exists a robust health insurance market in Portugal, which is instrumental in delivering healthcare access. The health insurance sector is dynamic and essential, underscored by metrics such as premiums as a percentage of GDP, claim ratios, and premium distribution. For instance, there was a significant increase in premiums as a percentage of GDP from 1.034 million euros in 2021 to 1.156 million euros in 2022, signifying an 8,90% increase and illustrating the sector's growth and economic impact. The claim ratio, indicative of the efficiency in managing and reimbursing healthcare costs, experienced a slight reduction from 77,20% in 2021 to 76,90% in 2022. The premium distribution is equitable, with 52,00% derived from group policies and 48,00% from individual policies, accommodating diverse healthcare preferences. The market encompasses a substantial population, with 3.423 million insured individuals benefiting from its plans. The average premium per insured person in 2022 amounted to 345,30 euros, with group policies averaging 326,60 euros and individual policies at 365,60 euros, thereby demonstrating the sector's capacity to offer customized and

accessible healthcare options to a large spectrum of individuals. Overall, the Portuguese health insurance sector is integral to fortifying the resilience of healthcare, providing a supplementary path for healthcare utilization beyond the National Health System, and making sure to reach the entire Portuguese population (Multicare - Seguros De Saúde, S.A., 2023).

Multicare, a distinguished entity within the Fidelidade Group's insurance framework, asserts itself as the preeminent leader in Portugal's health insurance sector. In 2022, it reached a market share of 36,10%, reflecting significant advancements in both premium growth and the scale of insured individuals. The organization currently provides coverage for approximately 1.2 million individuals, an increase from 1.14 million in the previous year, yielding approximately 418 million euros in Gross Written Premium. The client portfolio demonstrates a balanced strategic approach, with 42,30% of policies classified as individual and 57,70% as group-related, thereby addressing a diverse spectrum of healthcare needs. The persistent escalation in health insurance premiums averages 8,80%, which accounts for one-third of the population in Portugal. Since 2015, Multicare has sustained a commendable average growth rate of 10% in premiums, thereby solidifying its status as the unrivalled leader within the industry and exemplifying its commitment to offering comprehensive and accessible healthcare services to its clients (Multicare - Seguros De Saúde, S.A., 2023).

The Actuarial and Technical Control Department (GAC) of Multicare performs a pivotal function in overseeing and enhancing the technical performance of the company's insurance offerings. Utilizing sophisticated actuarial and statistical methodologies, the GAC scrutinizes the Multicare Technical Account, assessing profitability and claims patterns, and detecting anomalous trends within the company’s portfolio. Possessing a diverse array of responsibilities, the team collaborates extensively with other departments, contributing to the technical design of products, supervising the Technical Operating Budget, and executing comprehensive data quality evaluations to ensure precise reporting and risk assessment. Through systematic evaluations of critical financial metrics and claim data analysis, the GAC furnishes insights into emerging trends and variances that could affect the financial stability of Multicare. They employ predictive models and conduct scenario analyses to enhance premium pricing strategies, thereby aligning product offerings with market demands while upholding robust underwriting standards. Furthermore, the team assimilates scientific advancements in actuarial science to bolster business growth initiatives and decision-making processes (Multicare - Seguros De Saúde, S.A., 2022).

The Multicare Fraud Investigation Team (GAF) employs a comprehensive strategy for monitoring fraudulent activities within the healthcare sector. They utilize a diverse array of tools and techniques to detect and address fraud effectively. By engaging in various networking initiatives and receiving alerts from distinct divisions within the Fidelidade Fraud Ecosystem, the team remains vigilant against potentially fraudulent activities. Through the application of AI technology to sift through extensive health data, they efficiently identify patterns and irregularities indicative of fraudulent behaviour. Nonetheless, a substantial

portion of their work is conducted manually due to limitations in resources and capacity. Furthermore, the team conducts detailed assessments of both client and provider behaviours, searching for patterns related to pre-existing conditions and any collusion with healthcare providers.

To detect anomalies, they focus on annual reports by specialty in eight key areas: Otolaryngology, Vascular Surgery, Gastroenterology, Stomatology, Ophthalmology, Gynaecology, Dermatology, and Cardiology. Through their analytical and monitoring methodology, they standardize pricing and cluster providers according to the procedures they conduct. They scrutinize similar groups of clients to identify any discrepancies. Their analysis considers various factors, including the frequency and combination of medical services per client, adherence to clinical protocols, cost variations, and past billing behaviours. Based on this analysis, behaviours are classified into three categories: “justified asymmetry”, “asymmetry with potential for correction”, and “asymmetry lacking justification”, ensuring appropriate corrective actions are implemented to mitigate fraudulent practices and defend the integrity of Multicare. Despite their diligent efforts, the team faces challenges due to limited resources, which restrict their capacity to investigate all potential fraud cases (Multicare - Seguros De Saúde, S.A., 2023; Multicare - Seguros De Saúde, S.A., 2024).

3.1.1.2 BUSINESS OBJECTIVES

The issue relates to healthcare insurance fraud and anomalous billing practices specifically associated with healthcare providers in the Multicare network. Insurers in the healthcare sector encounter considerable difficulties in uncovering fraudulent activities and discerning anomalous patterns within provider claims. To address this challenge, it is crucial to answer the following business questions using provider billing data for the eight medical specialties from the years 2022 and 2023:

1. What type of billing behaviour do healthcare providers exhibit?
2. Are abnormal providers detectable?
3. Are there trends or patterns in provider billing practices that indicate potential fraudulent behaviour?
4. What are the most anomalous billing practices for each medical service these providers provide?
5. Are there specific types of medical services that are more prone to billing anomalies and, if so, how can these be identified and addressed?
6. Do specific provider networks exhibit higher rates of suspicious activities than others?
7. How do billing practices vary between different types of providers (eg, general practitioners vs. specialists)?

8. Do providers with a higher number of clients exhibit different patterns compared to those with fewer clients?
9. How are provider's billing practices flagged for fraud compared to those that have not?
10. How effective are current fraud detection systems in identifying fraudulent provider behaviour and where can they be improved?

Within this context, billing practices refer to the comprehensive set of medical acts, episodes, and the total cost associated with each provider. Medical acts include the specific procedures and services executed, episodes refer to the sequences of care related to the specific procedures, and the total cost encompasses all charges invoiced by the provider.

The deployment of anomaly detection techniques to detect fraudulent providers has the potential to provide substantial advantages to Multicare. The effective reduction of healthcare insurance fraud can culminate in considerable financial savings and assist in the mitigation of potential losses. Furthermore, the fulfilment of these objectives will not only safeguard the interests of policyholders and partners but will also result in cost reductions and operational efficiencies through the automation and optimization of fraud detection processes.

3.1.1.3 BUSINESS SUCCESS CRITERIA

The primary goal for achieving business success with the anomaly detection system is to enhance the existing GAC model to achieve greater performance in detecting unusual billing practices and potential fraudulent providers. This refinement will provide the fraud and actuarial departments with advanced insights, improving the effective identification of unusual provider behaviour, which will permit proactive risk management. One key objective is to improve the model's ability to generate actionable insights regarding potentially fraudulent activities. The system should not only detect anomalies but also provide valuable context to help departments differentiate between deviations due to legitimate yet unconventional provider practices and those indicative of possible fraud. By refining the model, the aim is to support proactive decision-making, thereby enhancing stakeholder trust in the insurance system's reliability and integrity.

The evaluation of the success criteria will involve a collaborative assessment by the fraud and actuarial departments, which will monitor the model's influence on operational efficiency and fraud detection outcomes. The Key Performance Indicators (KPIs) will include improvements in the precision of detecting anomalies, a decrease in false positives, and an improvement in the efficiency of the billing practices review process. Additional metrics will assess the model's ability to adapt to evolving fraud patterns and align with broader organizational objectives in risk management and system integrity. Regular reviews will be undertaken to facilitate continuous improvement, and feedback will be solicited from all relevant areas within Multicare. These reviews will ensure the system remains congruent with organizational

priorities and regulatory standards. The GAC and GAF teams will engage in annual evaluations to monitor progress, refine strategies, and respond to emerging trends in healthcare fraud, thereby ensuring the system's effectiveness and relevance over time.

3.1.2 ASSESS SITUATION

Within the framework of this dissertation, Multicare has granted access to a dataset encompassing comprehensive information on healthcare claims, with an emphasis on billing practices across diverse medical specialties. This dataset comprises records of medical acts, episodes, and the total costs allocated to each provider within eight specific specialties: Otolaryngology, Vascular Surgery, Gastroenterology, Stomatology, Ophthalmology, Gynaecology, Dermatology, and Cardiology. The primary aim of this research is to scrutinize these billing practices to detect anomalous patterns, particularly those that may suggest irregular or potentially fraudulent activities among healthcare providers. Through the examination of deviations in billing patterns, this research aims to support Multicare's continuous efforts to maintain transparency and ensure integrity within its network of providers.

The segment under analysis comprises healthcare providers associated with Multicare. The primary objective is to identify outliers and anomalous patterns in billing, predicated on statistical deviations that may indicate potential discrepancies with established medical billing norms. This scrutiny is crucial for ensuring that providers comply with the relevant billing standards, thereby advancing Multicare's goal of reducing unnecessary expenses and preventing fraudulent activities within its healthcare system.

Owing to the sensitive nature of the provided data, this project is governed by stringent confidentiality and privacy constraints. The dataset encompasses personally identifiable information pertaining to healthcare providers and comprehensive billing records, safeguarded by privacy regulations and corporate policies. Compliance with these privacy constraints is imperative, and all data handling and analysis will be executed in alignment with the confidentiality agreement established with Multicare.

The confidentiality declaration articulates that all information disseminated by Multicare, designated as "Confidential Information", shall be utilized solely for this project and disclosed exclusively to authorized persons. This stipulates the limitation on the dissemination of data to third parties. Moreover, if any external legal mandate requires disclosure, Multicare will be notified in advance to ensure that all essential measures are implemented to preserve confidentiality, disclosing only what is mandated by law.

3.1.3 DETERMINE DATA MINING GOALS

The objectives of data mining for this project include a thorough examination of provider behavioural patterns with the intent of identifying anomalies. The research concentrates on identifying abnormal healthcare providers through a meticulous analysis of their billing

practices, employing anomaly detection techniques. By focusing on atypical billing behaviours, such as extraordinary frequencies, costs, and combinations of medical services, the study aims to accurately identify providers whose practices significantly deviate from established norms.

Outlier detection techniques will be applied to billing data, enabling the system to capture subtle irregularities that may signal fraudulent behaviour. By isolating these aberrant billing practices, the research can effectively flag providers engaging in potentially suspicious activities, thereby setting them apart from providers who follow standard billing behaviours. Through this focused analysis of billing practices, the model is intended to provide robust insights into the behavioural patterns of healthcare providers. This ensures that the anomaly detection system highlights providers with atypical billing patterns and proactively enhances Multicare's capability to address potential fraud within the provider network. The objective is to foster a refined and data-driven approach to fraud detection, whereby the identification of abnormal billing practices directly corresponds to the identification of high-risk providers.

Three primary assessments will define the data mining success criteria for this project:

1. The first criterion involves evaluating the model's effectiveness in identifying providers classified as "asymmetry with potential for correction" and "asymmetry lacking justification", according to the GAF team. This assessment will use accuracy, precision, recall, and F1 score metrics to determine how well the model aligns with GAF's classifications.
2. The second criterion examines the model's performance on a sample of providers already flagged as potential deviants by the GAF team. Again, accuracy, precision, recall, and F1 score will be used to evaluate the model's ability to detect these preidentified cases accurately.
3. The final criterion will assess the agreement between different model outputs by comparing the providers classified as deviants in the GAC model, the providers with "asymmetry with potential for correction" and "asymmetry lacking justification" according to the GAF team classification, and the potential anomalous providers detected in this work. This comparison will help evaluate the consistency and effectiveness of the models in identifying anomalous provider behaviours.

Applying these criteria is anticipated to help the model achieve strong performance metrics and demonstrate alignment with expert assessments, enhancing its relevance for making operational decisions within Multicare.

3.1.4 PROJECT PLAN

The project plan will be executed in distinct key phases in accordance with the CRISP-DM framework, to ensure a methodical approach to fulfilling the established business and data mining objectives. Initially, a considerable portion of the project's initial stages will be allocated to data exploration. This phase will acquire a thorough understanding of the

available data pertaining to provider billing practices. This process entails collecting relevant datasets, examining their structure, quality, and completeness, and deriving insights into their underlying patterns and relationships.

Subsequently, the data preparation phase will become the main area, occupying a substantial segment of the project timeline. In this phase, the focus will be on cleaning the data, rectifying inconsistencies, managing missing values, and transforming the data into a format ready for analysis. This phase is crucial as it establishes the foundation for subsequent modelling activities.

Upon the preparation of the data, the project will proceed to the modelling phase. During this step, anomaly detection techniques will be employed to identify and flag atypical instances within the dataset. The development of models will entail experimentation with a range of algorithms and parameter configurations to enhance performance and efficiency. After model development, the evaluation of the results and deployment phases will be prioritized, occupying a lesser yet crucial segment of the project timeline.

During these stages, the principal objective is to harmonize business objectives with data mining goals, ensuring the insights generated from the analytical processes are relevant and practical. This cyclical approach ensures the project stays adaptable and reacts effectively to the continuously changing dynamics of the insurance industry and its inherent difficulties.

3.2 Data Understanding

The data understanding phase begins with gathering data, where the analyst obtains initial datasets for analysis. Following this, efforts are made to improve understanding of the data, identify data quality issues, discover initial insights, and identify interesting subsets that may contain hidden information. This phase includes four specific steps: initial data collection, descriptive analysis, exploratory data review, and validation of data integrity and quality (Shearer, 2000).

3.2.1 Data Collection and Description

The data acquisition process for Multicare's SAS Enterprise Guide system entails the assembly of datasets from multiple sources, including the software's projects and Datamarts. The SAS Enterprise Guide serves as a comprehensive tool for statistical and data analysis, enabling tasks like project-oriented Online Analytical Processing analysis. Within the system, projects consist of data collections, tasks, programs, and associated results, thereby allowing for the construction of numerous processes flows to illustrate various analytical workflows. The valuable datasets stem from SAS projects that compile numerous base tables from sources such as the Health Information System or Datamarts. These datasets generate tables containing essential business data, including policies, individuals, products, claims, and premiums. The data acquisition process involved retrieving data from the years 2022 and 2023, ensuring a comprehensive and longitudinal perspective of Multicare's operations.

Collectively, the gathered datasets offer an exhaustive insight into Multicare's operations and aid in facilitating informed decision-making processes within the organization (Ramos et al. 2022).

The data collected from Multicare's SAS Enterprise Guide system offers comprehensive insights into the billing practices of healthcare providers. In the context of each medical specialty, three distinct datasets are provided: one pertaining to medical acts, one related to episodes, and one concerning the total cost attributable to each provider. This facilitates a detailed analysis of billing practices across diverse healthcare services. Nonetheless, for the specialty of stomatology, only datasets concerning medical acts and total cost are accessible, as it utilizes a divergent analytical approach. These datasets encompass services availed by insurance users in both inpatient and outpatient settings, providing valuable insights into the various types of healthcare services pursued and the corresponding claims submitted by healthcare providers (Ramos et al. 2022).

3.2.2 DATA EXPLORATION AND QUALITY

During the data exploration phase, a systematic analysis of healthcare specialties uncovered significant trends and distinctions that direct the emphasis of this study's fraud detection initiatives. Initially, the dataset highlights diverse specialties possessing unique characteristics that influence the scope of the analysis. For instance, stomatology comprises solely medical act records with distinct variables, gynaecology introduces features differentiating between pregnant and nonpregnant patients, and otolaryngology segregates data into paediatric and non-paediatric cases. These features for specific medical specialties necessitate tailored methodologies to ensure precise evaluations. Specifically, stomatology presents the largest volume of records available for evaluation, where specialties such as vascular surgery and gastroenterology possess fewer records, impacting considerations of data volume and statistical strength. Across all specialties, particular provider networks demonstrate patterns that deviate from the norm, suggesting potential anomalies and necessitating detailed examination owing to their departure from standard record evaluations. Moreover, within a subset of records, the total expenditure (VAP) for inpatient services exceeds that for outpatient services, introducing an additional dimension to expenditure assessment. Finally, hospital groups are often assessed at an aggregated level rather than other establishments, acknowledging the collective impact these providers wield on expenditure and service patterns. These structured observations establish a thorough comprehension of the data and lay the foundation for a nuanced and targeted strategy for fraud detection in subsequent analyses.

Upon examination, the data quality is deemed satisfactory, as no significant issues have been identified. The absence of duplicate values within the datasets ensures data integrity and consistency is maintained. Nevertheless, certain features may lack informativeness due to their limited variability or the insignificant nature of the data, while others exhibit missing values necessitating further attention. It is noteworthy that all features exhibit variation,

thereby contributing to the overall analysis. Addressing the issue of missing values with suitable imputation methods or data-cleaning techniques is imperative to support the integrity of the analysis (Raja, 2020). Furthermore, thorough exploration and validation of the datasets are necessary to identify any potential inconsistencies or inaccuracies that may impact the reliability of the findings (Shearer, 2000). While the datasets appear broadly usable, attention to these data quality aspects is crucial to ensure accurate and reliable analysis results.

3.3 Data Preparation

The data preparation stage involves a wide range of tasks essential for creating the final dataset or the data used by modeling tools, starting from the raw data. Activities encompassed in this phase include the selection of tables, records, and attributes, followed by the refinement of data to make it suitable for modeling purposes. The data preparation process incorporates five principal steps: data selection, data cleaning, data construction, data integration, and data formatting (Shearer, 2000).

3.3.1 Data Selection and Cleaning

In the preparation of the dataset for the analysis of fraud detection, the approach to data selection and cleansing has been customized to preserve outliers, as these are crucial to the study's principal objective of identifying fraudulent providers through outlier and anomaly detection methodologies (Du & Yu, 2023). Contrary to conventional preprocessing practices where outliers might be eliminated to ensure data consistency, this analysis necessitates the retention of all instances of outlier data, as they signify potential billing inconsistencies essential for the identification of providers exhibiting abnormal behaviours. By retaining these outliers, the model is more able to capture a variety of unconventional billing practices, thereby enhancing its capacity to identify patterns that may be indicative of fraud.

The emphasis is placed solely on outpatient data, due to the high variability of inpatient billing practices across various medical specialties, which could potentially undermine the model's efficacy if combined with outpatient data. Inpatient cases typically encompass a wider array of treatment complexities and costs, demonstrating considerable variability among specialties, consequently leading to diverse billing patterns that may obscure the identification of anomalous outpatient practices. Furthermore, certain specialties require more records of inpatient data for a robust statistical analysis. Focusing exclusively on outpatient data not only ensures uniformity in billing practices but also enhances the reliability and efficiency of the anomaly detection model by avoiding the disturbance and inconsistency brought about by inpatient cases. This specialized approach facilitates a more concentrated detection of irregular billing practices within the outpatient context, thereby improving the model's relevance and accuracy in detecting providers who may engage in fraudulent activities.

3.3.2 DATA CONSTRUCTION, INTEGRATION AND FORMATTING

The processes of data construction, integration, and formatting are fundamental in the preparation of data for the purposes of anomaly detection and analysis. Data construction entails the creation of new features that optimize the dataset's practicality for modelling objectives, with a focus on simplifying complex variables to capture the most prominent features. Combining various tables or records to form a consistent dataset, data integration merges different sources of information and performs aggregations to develop a unified view that includes the crucial attributes of each entity. Lastly, data formatting involves the modification of the data's structure and values to ensure it is compatible with modelling tools, which may include the conversion of categorical or symbolic fields into numerical values or the reorganization of data fields to facilitate efficient processing. These steps lay a robust foundation for precise and efficient data analysis, thereby enhancing the efficacy of the anomaly detection process (Shearer, 2000).

Feature scaling represents an essential procedure in standardizing the range and magnitude of features, thereby ensuring equitable contribution to model computations. This procedure is critically important in the field of machine learning because it enables more effective gradient descent and more even weight adjustments, especially in situations where features vary greatly in size. The Standard Scaler method, as utilized in this study, modifies each feature to attain a mean of zero and unit variance, thereby centring values proximally to zero while preserving the original dispersion of data. This scaling methodology is particularly effective in retaining outliers, which are crucial for identifying anomalies within billing practices. By standardizing data without confining, it to a predetermined range, the Standard Scaler permits uniform calculations based on distance across features, thereby enhancing the precision of anomaly detection analyses (Hansson & Cedervall, 2022; Hamid, et al., 2024).

$$x_{scaled} = \frac{x - \mu}{\sigma}, \tag{3.1}$$

where x_{scaled} is the scaled sample point, x is the sample point, μ is the mean and σ is the standard deviation.

In this study, categorical variables are encoded utilizing the one-hot encoding technique, which is highly appropriate for transforming categorical data into numerical features that align with machine learning algorithms. One-hot encoding delineates each category as a distinct binary vector, thereby preserving the unique identity of each category without imposing any artificial hierarchical order. This method proves to be effective for this dataset, given that the categorical variables encompassed possess a limited number of unique categories, making one-hot encoding both efficient. While this technique may become less feasible with extensive category sets due to increased memory and computational requirements, the attributes of this dataset ensure the encoded data remains compact and efficient. Therefore, one-hot encoding is suitable for the representation of categorical features within the context of this

research, facilitating efficient data processing without necessitating more sophisticated encoding techniques (Fursov, et al., 2022).

Principal Component Analysis (PCA) constitutes a technique for dimensionality reduction that converts high-dimensional data into a set of principal components, which capture the most significant variance present in the dataset. Through the projection of the original data onto a lower-dimensional space, PCA preserves essential information while eliminating redundancy, with each principal component indicating a direction of maximal variance within the data. This transformation is predicated upon a matrix of orthonormal eigenvectors, facilitating the identification of primary patterns within the data through the decomposition of its covariance structure (Jolliffe & Cadima, 2016). Within this study, the reduction of the data to three principal components proves particularly efficacious in detecting atypical billing behaviours among healthcare providers, as it simplifies the data structure, making outliers more noticeable. Furthermore, this three-dimensional representation facilitates the clear visualization of clusters and anomalies, thereby aiding in the identification of irregular billing patterns. PCA enhances data analysis and modelling efficiency by mitigating computational complexity, thereby making it highly suited for the detection of potential anomalies in billing practices (Du & Yu, 2023).

New variables were created across all datasets, utilizing the GAF team's business expertise to mimic their approach to identifying potentially fraudulent providers. Each variable was meticulously constructed to encapsulate distinct facets of billing behaviour that may indicate anomalies, with a focus on the unique characteristics of each dataset and specialty. For the medical acts, episodes, and total cost datasets, variables were devised to evaluate the average cost per client for each medical act, episode, or provider. This analysis facilitates the identification of cost deviations that may suggest irregular billing practices.

Furthermore, a variable was incorporated into the medical acts and episodes datasets to capture the frequency at which clients undergo specific acts or episodes for each provider relative to the overall Multicare network. This assists in pinpointing providers whose practices diverge from network norms, thereby identifying potential outliers. The total cost dataset is concentrated on comparing the average cost per client for each provider with the network average, where providers exhibiting significantly higher costs are flagged as potential anomalies. To mitigate distortions in the outlier detection analysis, cases not aligning with the analysis scope or those with fewer than ten clients were marked as null. This methodology was applied universally across all specialties, with additional customizations for specific instances: in gynaecology, supplementary variables were established to differentiate between pregnant and nonpregnant clients; in otolaryngology, the analysis was separated into paediatric and non-paediatric cases; and in stomatology, solely the medical acts and total cost datasets were utilized, as this specialty does not conform to the episode’s dataset. These new variables provide robust groundwork for detecting potentially fraudulent billing behaviours among healthcare providers (see Appendix A).

In the context of unsupervised learning, the process of feature selection predominantly depends on domain expertise to identify the most relevant features, given the absence of labelled data to direct the process. For this study, the feature selection effort was guided by the GAF team's business expertise, which played a pivotal role in identifying features most likely to aid in the detection of anomalous billing practices within healthcare providers.

Two versions of the feature sets (see Appendix B) were created to achieve a balance between comprehensiveness and precision. The initial version is less restrictive, encompassing an extensive array of features potentially instrumental in discerning anomalous billing behaviours, even though their individual effects differ. The second version is more focused, incorporating only the crucial features necessary for identifying irregular billing practices, thus enhancing the model's efficiency and clarity. Moreover, a correlation analysis of all features was performed to evaluate feature relationships and refine the feature selection, thus minimizing redundancy and ensuring the retention of the most informative variables for the anomaly detection model (Dy & Brodley, 2004).

3.4 MODELLING

During the modelling phase, various methodologies are meticulously selected, applied, and refined to achieve optimal alignment with the data mining task at hand. This comprehensive process encompasses the selection of an appropriate modelling technique, the establishment of a test design for model evaluation, the construction of the model, and the subsequent assessment of its performance. The outcomes are rigorously reviewed in conjunction with domain expertise to ascertain their alignment with business objectives (Shearer, 2000).

3.4.1 UNSUPERVISED LEARNING ALGORITHM

These algorithms possess the capacity to autonomously discern essential patterns across diverse domains, including pattern recognition, market analysis, and fraud detection. Despite their extensive applications, the formulation of theoretically robust and evaluable methodologies persists as a significant challenge, prompting the ongoing introduction of new algorithms in recent years (Celebi, 2016).

3.4.1.1 ISOLATION FOREST

The Isolation Forest (iForest) constitutes an anomaly detection technique renowned for its efficiency and scalability. It involves the creation of an ensemble of isolation trees (iTrees) adapted to a specific dataset, in which anomalies are identified based on their abbreviated average path lengths within the trees. Two principal parameters, namely the number of trees and the sub-sampling size, have a substantial impact on the performance of iForest. The method is distinguished by its rapid convergence, achieving high detection accuracy even with a limited number of trees and a proficient subsampling size.

Unlike other methods, iForest capitalizes on tree isolation characteristics (see Figure 3.1). This unique feature allows the construction of partial models and efficient subsampling,

contributing to its computational efficiency. iForest relies on something other than distance or density measures for anomaly detection, reducing the computational overhead associated with these metrics.

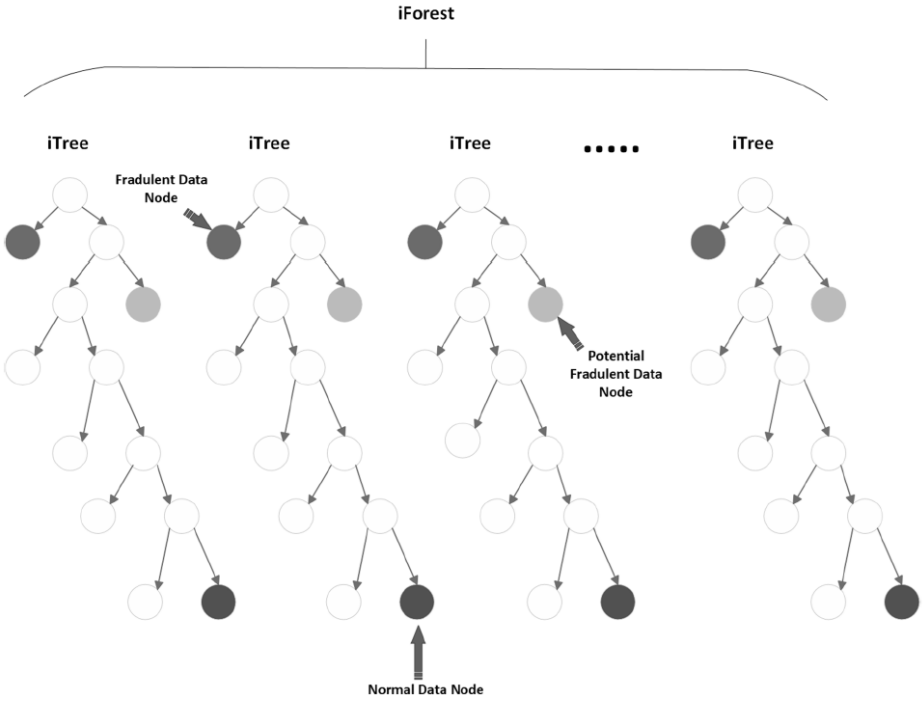

Figure 3.1 - Isolation Forest; Source: (Hamid, et al. 2024)

iForest demonstrates superior performance concerning time complexity, exhibiting linear time complexity characterized by a low constant, while requiring minimal memory resources. Consequently, iForest is especially particularly suitable for processing large datasets and addressing high-dimensional problems, even in the presence of numerous irrelevant attributes (Liu, Ting, & Zhou, 2008).

The primary result generated by this algorithm is the anomaly score assigned to each instance, calculated in the following manner:

$$s(x,n) = 2^{-\frac{E(h(x))}{c(n)}}, \tag{3.2}$$

where $E(h(x))$ denotes the average number of edges to be separated for instance x and $c(n)$ is the normalization constant for a dataset with n instances.

The anomaly score for an instance x is expressed as a function of the average $h(x)$ and the normalization factor $c(n)$. In particular:

- When $E(h(x))$ approaches $c(n)$, s tends towards 0,5.
- When $E(h(x))$ approaches 0, s tends towards 1.
- When $E(h(x))$ Approaches n1, s tends towards 0.

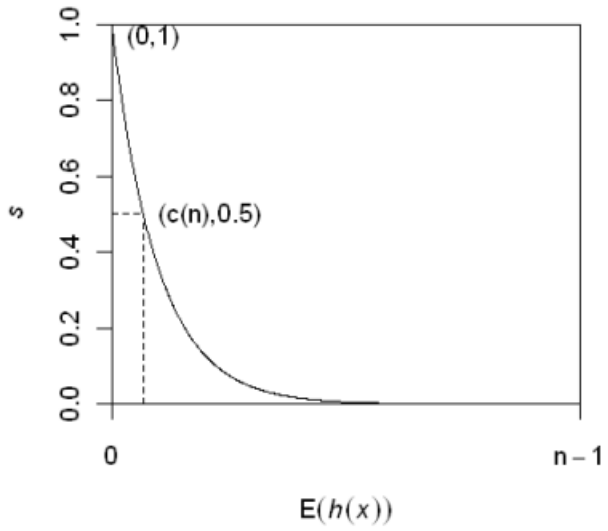

Figure 3.2 - Relationship between s and $h(x)$;
Source: (Liu, Ting, & Zhou, 2008)

The monotonic relationship between s and $h(x)$ is illustrated in Figure 3.2, with specific conditions applied. The assessment using the anomaly score includes the following:

1. Instances returning s close to 1 are definite anomalies.
2. Instances with values much smaller than 0,5 are likely normal.
3. Instances with values around 0,5 indicate the absence of distinct anomalies in the entire sample.

The Isolation Forest algorithm was selected for this analysis based on insights derived from the literature review. It was identified as particularly efficacious in detecting anomalous billing behaviours in healthcare data, utilizing singular anomalous outlier detection techniques. The model was applied across datasets of medical procedures, episodes, and total costs for each medical specialty to effectively capture distinctive billing irregularities. Implemented with the Python library scikit-learn, Isolation Forest was chosen due to its robust alignment with the study's focus on identifying isolated, unusual billing patterns.

The model was configured with critical hyperparameters for optimal performance. The number of estimators, n_estimators, was set to 100, which is the default in scikit-learn, as this configuration provides a robust ensemble of isolation trees for detecting outliers. Regarding max_samples, which determines the number of samples used to construct each tree, the full length of each respective dataset was utilized. This approach ensures that the sample size is optimized, aligning with the dataset split, thereby enhancing the model's capacity to classify transactions accurately. This decision aligns with recommendations from Joshi, Soni, & Jain, (2021), who underscore the significance of larger sample sizes in achieving reliable isolation trees. The contamination parameter, which represents the proportion of expected anomalies, was set to 0,05, following the guidance of Bairy, Muniyal, & Shetty, (2024). However, given that dataset sizes vary by specialty, slight modifications were made to this contamination rate

to account for differences in dataset length and to improve accuracy. This adaptive adjustment enabled the model to capture more precise anomaly patterns tailored to each specialty. The results from the Isolation Forest model were visualized in three dimensions (see Figure 3.3), providing an intuitive perspective on the anomalies detected. This visualization assists in identifying the providers' billing practices that deviate from expected norms, thereby highlighting unusual billing behaviours across specialties. The Isolation Forest model, as implemented through scikit-learn, thus provided an efficient and accurate framework for anomaly detection in healthcare billing data. (Liu, Ting, & Zhou, 2008; Liu, Ting, & Zhou, 2012)

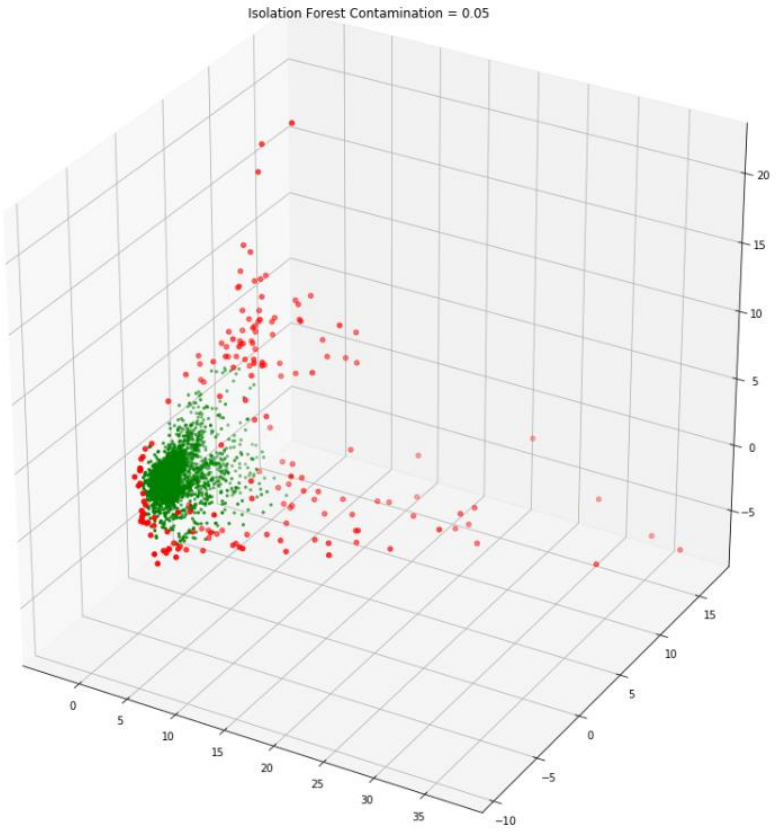

Figure 3.3 - iForest 3D Plot Results

4. EVALUATION AND DEPLOYMENT

The model undergoes a comprehensive review during the evaluation phase to ascertain its alignment with business objectives, address any previously overlooked issues, and evaluate its applicability in real-world scenarios. Essential steps encompass the evaluation of results, review of the process, and determination of subsequent deployment steps. During the implementation phase, the insights derived from the model are systematically integrated into business processes, which may vary from simple reports to intricate systems. This phase entails the planning of deployment, the establishment of monitoring and maintenance strategies, the production of a final report, and a thorough project review. The review process aims to identify successes, failures, and valuable insights for forthcoming projects, which may include participant feedback to enhance future data mining endeavours (Shearer, 2000).

4.1 EVALUATION

A model in machine learning is validated by evaluating its prediction performance. This evaluation should be representative of how the model would perform when deployed in a real-life setting (Raschka, 2018).

Evaluating unsupervised tasks often poses considerable difficulties because there are no true labels, making it hard to establish objective performance metrics. Nevertheless, in this study, despite the primary focus being on an unsupervised approach, the evaluation method incorporates a supervised binary classification task. This involves training a model with observations and their true labels to predict these target labels for new instances. By using this supervised framework for evaluation, the study guarantees a more solid and quantifiable evaluation of how effective the unsupervised model is in identifying irregular billing practices (Provost & Fawcett, 2013).

In the realm of classification, the results depend on the predicted value and the actual value. Consider a classification model geared toward predicting fraud, where fraud is designated as the positive class (1), and the absence of fraud is denoted as the negative class (0). Instances where a predicted fraud is correctly identified as fraudulent are termed true positives (TP). On the other hand, if the prediction is inaccurate, such instances are labelled false positives (FP). Accurate negative predictions fall into the category of true negatives (TN), while their corresponding errors are false negatives (FN). The confusion matrix serves as a valuable analytical instrument for delving into the nuances of an evaluation test, providing a detailed breakdown of the results. It forms the foundation for computing various performance metrics by tabulating the frequencies of each conceivable prediction outcome made by a model, offering a comprehensive view of the model's performance (Kelleher, Mac Namee, & D'Arcy, 2015).

		Prediction	
		Positive	Negative
Target	Positive	*TP*	*FN*
	Negative	*FP*	*TN*

Figure 4.1 - The structure of a confusion matrix

A variety of performance metrics can be directly extracted from the confusion matrix, with precision and recall representing two of the most computed measures. Precision evaluates the level of confidence in accurately identifying instances possessing the positive target level, thereby reflecting the reliability when the model forecasts a positive target level. Conversely, recall measures the model's efficacy in identifying all instances associated with the positive target level. Both precision and recall produce values within the [0, 1] range, wherein higher values signify enhanced model performance.

$$Precision = \frac{TP}{(TP + FP)} \tag{4.1}$$

$$Recall = \frac{TP}{(TP + FN)} \tag{4.2}$$

In the domain of fraud prediction, the target class is often marked by a pronounced imbalance. Typically, a considerable proportion of instances are classified as non-fraudulent, thereby creating an uneven distribution in the target class. This imbalance introduces distinct challenges for machine learning models, which may develop a predisposition towards the majority class. Consequently, accurately identifying instances that belong to the minority class (those associated with fraudulent activities) becomes more challenging. Managing this class imbalance effectively is essential for the development of robust and reliable fraud prediction models. In this context, the true negatives will constitute the majority, and their relevance in fraud prediction will be reduced due to the imbalanced nature of the target classes.

The importance of metrics such as precision and recall are emphasized by the previously mentioned factors. By specifically examining true positives and false negatives, these metrics assume an essential role in assessing the effectiveness of fraud prediction models. In this context, the F1 score, calculated as the harmonic mean of precision and recall, is recognized as a valuable metric. Its characteristic of reduced sensitivity to large outliers is particularly beneficial, enhancing the robustness of model performance evaluation. This attribute is especially critical in fraud detection scenarios, where achieving a balanced trade-off between precision and recall is imperative for an accurate evaluation of the model (Kelleher, Mac Namee, & D'Arcy, 2015).

$$F1\ score = \frac{2 \times Precision \times Recall}{(Precision + Recall)} \quad (4.3)$$

4.1.1 EVALUATE RESULTS

The model was systematically tested using different feature configurations to assess its effectiveness in detecting unusual billing practices. Initially, the Isolation Forest model was deployed without implementing any feature selection, utilizing all accessible features across different specialties and datasets. Subsequently, two additional configurations were explored: one that employed only numerical features, and another that incorporated two distinct versions of feature selection as previously delineated in this study.

These configurations were executed for each specialty, scrutinizing datasets for diverse medical acts, episodes, and total cost metrics to assess the model's adaptability and performance across varied data structures. Across all specialties and datasets, the Isolation Forest model successfully identified anomalous billing practices, highlighting atypical billing behaviours. Nonetheless, given that the primary aim of this study is to detect fraudulent providers as opposed to isolated anomalous billing events, the analysis was expanded to evaluate the providers linked with these anomalies. Each provider identified by the model for anomalous billing practices underwent further evaluation to assess connections between datasets. A consistency evaluation criterion across various datasets was created to increase the rigor of anomaly detection. Specifically, a provider was classified as potentially anomalous if it exhibited anomalous billing practices in at least two datasets: medical acts, episodes, and total costs. This criterion rests on the understanding that a provider's billing pattern need not deviate across all metrics to be deemed suspicious; a consistent deviation in at least two billing dimensions is sufficient to indicate potentially fraudulent behaviour. This approach offers a comprehensive perspective of each provider's billing practices, elevating the model's capability to detect anomalous providers.

First Data Mining Success Criteria:

Following the model's identification of potential anomalous providers, the outcomes were evaluated by comparing them to the GAF team's categorization of providers as exhibiting "asymmetry with potential for correction" and "asymmetry lacking justification." This evaluation employed recall as the primary metric to ascertain which feature configuration most effectively identifies these anomalous providers (see Table 4.1). Analysis of the recall rates across various feature combinations revealed that the second version of feature selection consistently delivered superior results, except for ophthalmology in 2023.

Table 4.1 - Recall Performance of iForest Across Feature Configurations and Specialties (2022 and 2023)

	All Features		Numerical Features Only		Feature Selection All Features V1	
	2022	**2023**	**2022**	**2023**	**2022**	**2023**
Cardiology	0,429	0,250	0,286	0,375	0,286	0,250
Vascular Surgery	0,667	0,556	0,667	0,556	1,000	0,556
Dermatology	0,200	0,500	0,200	0,667	0,600	0,667
Stomatology	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
Gastroenterology	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
Gynaecology	0,444	0,500	0,444	0,417	0,444	0,583
Ophthalmology	0,357	0,500	0,286	0,500	0,500	0,700
Otolaryngology	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.

	Feature Selection All Features V2		Feature Selection Numerical Features V1		Feature Selection Numerical Features V2	
	2022	**2023**	**2022**	**2023**	**2022**	**2023**
Cardiology	**0,857**	0,625	0,286	0,500	0,714	**0,750**
Vascular Surgery	**1,000**	**0,778**	0,889	0,667	1,000	0,778
Dermatology	0,600	**1,000**	0,600	0,667	**0,800**	1,000
Stomatology	**0,961**	**0,922**	n.a.	n.a.	n.a.	n.a.
Gastroenterology	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
Gynaecology	**0,778**	**0,750**	0,444	0,583	0,778	0,667
Ophthalmology	**0,571**	0,667	0,500	**0,733**	0,571	0,700
Otolaryngology	**1,000**	**0,840**	n.a.	n.a.	n.a.	n.a.

This consistent performance highlights the efficacy of the second feature selection version in aligning the model's outputs with anticipated outcomes. For specialties that deviated from the normative analysis, this version of feature selection was assumed to be the optimal configuration. Notably, within the gastroenterology specialty, the GAF team did not identify any providers within the "asymmetry with potential for correction" or "asymmetry lacking justification" classifications; hence, an evaluation of this specialty could not be conducted under this criterion. These results underscore the value of customized feature selection in improving the model's accuracy in detecting anomalous billing practices, ensuring that the elected configurations effectively capture significant anomalies within the dataset.

The assessment of the Isolation Forest model across eight medical specialties using metrics such as precision, recall, and F1 score affords a comprehensive perspective on the model's efficacy in identifying anomalous providers, based on the classifications of the GAF team (see

Table 4.2). The model exhibited consistently high recall rates across all specialties, evidencing its robust capacity to detect potential anomalous providers. This high recall signifies that the model effectively identified most providers categorized as "asymmetry with potential for correction" and "asymmetry lacking justification", which is critical for achieving the primary objective of identifying potentially fraudulent providers. Nevertheless, the metrics indicate variability in precision and accuracy, suggesting that, while the model is proficient in detecting anomalous behaviour, it also tends to classify some non-anomalous providers as suspicious. This tendency underscores the model's sensitivity but reveals a relatively high false positive rate in certain specialties, in which it may overestimate potential anomalies. Specifically, stomatology and vascular surgery yielded more balanced results, characterized by high recall and enhanced precision and accuracy scores, suggesting the model's alignment with these specialties and superior specificity, thereby reducing false positives.

Table 4.2 - Performance Metrics of iForest Based on GAF Classifications for Providers Identified Within “Asymmetry with Potential for Correction” and “Asymmetry Lacking Justification” (2022 and 2023)

	Cardiology		**Vascular Surgery**		**Dermatology**		**Stomatology**	
	2022	**2023**	**2022**	**2023**	**2022**	**2023**	**2022**	**2023**
Accuracy	0,103	0,098	0,409	0,333	0,115	0,194	0,480	0,427
Precision	0,105	0,102	0,409	0,368	0,125	0,194	0,490	0,443
Recall	0,857	0,750	1,000	0,778	0,600	1000	0,960	0,922
F1 Score	0,188	0,179	0,581	0,500	0,207	0,324	0,649	0,599

	Gastroenterology		**Gynaecology**		**Ophthalmology**		**Otolaryngology**	
	2022	**2023**	**2022**	**2023**	**2022**	**2023**	**2022**	**2023**
Accuracy	n.a.	n.a.	0,125	0,102	0,069	0,203	0,253	0,276
Precision	n.a.	n.a.	0,130	0,106	0,072	0,220	0,253	0,292
Recall	n.a.	n.a.	0,780	0,750	0,571	0,733	1000	0,840
F1 Score	n.a.	n.a.	0,220	0,186	0,129	0,338	0,404	0,433

In conclusion, the Isolation Forest model with optimized feature selection exhibited a high level of recall, signifying its competence in identifying anomalous providers in accordance with expert evaluations. The model's tendency towards higher recall, particularly with certain feature configurations, indicates its potential efficacy in fraud detection. Nevertheless, achieving a harmonious improvement in precision and accuracy across different specialties remains an area requiring further refinement to optimize the model's applicability in real-world fraud detection contexts. This evaluation highlights the model's capability in detecting a substantial proportion of anomalous providers, while also stressing the necessity for modifications tailored to specific specialties to effectively manage false positives.

Second Data Mining Success Criteria:

In the assessment of the Isolation Forest model according to the second data mining success criterion, notable variations in performance are observed relative to the first criterion (see Table 4.3). Both evaluations employ the same optimal feature configuration identified from the first criterion, which emphasizes high recall, yet the outcomes underscore distinctions in the model's performance concerning the objectives of each criterion. Recall predominantly remains high across various specialties, congruent with the first criterion; however, there is an enhancement in precision within several specialties, such as Gynaecology, Stomatology, and Otolaryngology. This improvement in precision indicates that the model is more precise in detecting potentially fraudulent providers, denoting fewer false positives when correlated with the group of providers previously identified by the GAF team as potentially anomalous.

Table 4.3 - Performance Metrics of iForest Based on Providers Pre-Identified as Potential Deviants by the GAF Team (2022 and 2023)

	Cardiology		**Vascular Surgery**		**Dermatology**		**Stomatology**	
	2022	**2023**	**2022**	**2023**	**2022**	**2023**	**2022**	**2023**
Accuracy	0,453	0,283	0,723	0,632	0,513	0,300	0,741	0,533
Precision	0,105	0,441	0,409	0,368	0,150	0,387	0,490	0,689
Recall	0,857	0,441	1,000	0,778	0,600	0,571	0,960	0,702
F1 Score	0,188	0,441	0,581	0,500	0,240	0,462	0,658	0,695

	Gastroenterology		**Gynaecology**		**Ophthalmology**		**Otolaryngology**	
	2022	**2023**	**2022**	**2023**	**2022**	**2023**	**2022**	**2023**
Accuracy	0,320	0,320	0,366	0,306	0,425	0,397	0,495	0,365
Precision	0,421	0,444	0,685	0,435	0,591	0,580	0,671	0,583
Recall	0,571	0,533	0,440	0,507	0,602	0,558	0,654	0,494
F1 Score	0,485	0,485	0,536	0,468	0,596	0,569	0,663	0,535

Within the domains of stomatology and vascular surgery, the model consistently exhibits elevated levels of accuracy and recall across both evaluative criteria. Nevertheless, there is a marked enhancement in F1 scores, notably in stomatology for the year 2023 and vascular surgery for the year 2022, indicating an improved equilibrium between precision and recall in the second evaluative criterion. This enhancement may suggest that the model's feature selection is optimally aligned with the requirements of these specialties, proficiently identifying both recognized and previously undetected anomalies.

Conversely, the specialties of dermatology and gastroenterology exhibit greater variability in performance compared to the initial criterion. Dermatology demonstrates a reduced precision and F1 score in the second criterion, which may imply that the model encounters difficulties in discerning the subtleties within dermatology's flagged cases, particularly when compared with GAF's records. This outcome could be attributed to inherent complexities or the lower

incidence of flagged cases in these datasets, implying that specific specialties may necessitate additional features or refined configurations to improve specificity.

The transition from a high recall emphasis in the initial criterion to enhanced precision in the subsequent criterion demonstrates the model's adaptability and capability to respond to potential anomalous providers. These findings suggest that, although the model preserves a strong aptitude for identifying high-risk providers across both criteria, it attains an improved precision and recall equilibrium in the second criterion by efficiently concentrating on the subset of providers flagged by the GAF team.

This comparative analysis highlights that the optimal configuration of features facilitates a versatile detection strategy capable of identifying prospective anomalies and corroborating previously flagged instances. The flexibility of the Isolation Forest model enhances its effectiveness as a scalable solution for fraud detection across various healthcare domains, as it exhibits differential responses contingent upon the specific criteria utilized for assessing anomalous behaviour.

Third Data Mining Success Criteria:

The third criterion for success in data mining evaluates the correspondence between the outputs of the Isolation Forest (IF) model and the GAC model, and, where applicable, the classifications by the GAF team of providers designated as having “asymmetry with potential for correction” or “asymmetry lacking justification”, with the results in the Table 4.4 and Table 4.5. Notably, the GAF team's evaluations are limited to providers identified as deviant by the GAC model. Consequently, the GAF classifications are dependent on detections by the GAC model, indicating that any providers not highlighted by the GAC are not subsequently assessed by the GAF. As a result, the concordance column in this analysis represents the alignment solely between the IF and GAC models.

Table 4.4 - Comparison of Model Classifications and Concordance Rates Across Specialties (2022)

	GAC - 0 IF - 0 GAF - 0	GAC - 0 IF - 1 GAF - ?	GAC - 1 IF - 0 GAF - 0	GAC - 1 IF - 0 GAF - 1	GAC - 1 IF - 1 GAF - 0	GAC - 1 IF - 1 GAF - 1	Concordance of the models
Cardiology	78,74%	5,84%	7,71%	0,23%	6,07%	1,40%	**86,21%**
Vascular Surgery	80,20%	4,46%	8,91%	0,00%	1,98%	4,46%	**86,63%**
Dermatology	80,43%	6,38%	8,51%	0,85%	2,55%	1,28%	**84,26%**
Stomatology	45,92%	0,00%	40,79%	0,26%	6,58%	6,45%	**58,95%**
Gastroenterology	86,96%	5,31%	3,86%	0,00%	3,86%	0,00%	**90,82%**
Gynaecology	78,59%	4,10%	8,66%	0,46%	6,61%	1,37%	**86,56%**
Ophthalmology	70,53%	9,96%	5,89%	1,22%	10,77%	1,63%	**82,93%**
Otolaryngology	78,17%	9,15%	3,29%	0,00%	4,69%	4,46%	**87,32%**

Table 4.5 - Comparison of Model Classifications and Concordance Rates Across Specialties (2023)

	GAC - 0 IF - 0 GAF - 0	GAC - 0 IF - 1 GAF - ?	GAC - 1 IF - 0 GAF - 0	GAC - 1 IF - 0 GAF - 1	GAC - 1 IF - 1 GAF - 0	GAC - 1 IF - 1 GAF - 1	Concordance of the models
Cardiology	78,65%	7,42%	7,19%	0,46%	4,87%	1,39%	**84,92%**
Vascular Surgery	84,24%	2,46%	5,42%	0,99%	3,45%	3,45%	**91,13%**
Dermatology	80,54%	8,14%	5,43%	0,00%	3,17%	2,71%	**86,43%**
Stomatology	54,14%	0,00%	31,41%	0,53%	7,75%	6,18%	**68,07%**
Gastroenterology	86,08%	5,67%	4,64%	0,00%	3,61%	0,00%	**89,69%**
Gynaecology	71,46%	11,99%	7,43%	0,72%	6,24%	2,16%	**79,86%**
Ophthalmology	72,18%	9,62%	5,23%	1,67%	6,69%	4,60%	**83,47%**
Otolaryngology	76,74%	9,35%	5,04%	0,48%	2,88%	5,52%	**85,13%**

For both the years 2022 and 2023, the majority of providers are categorized under (GAC - 0 / IF - 0 / GAF - 0), suggesting a consistent classification by both models as non-anomalous. Notably, the field of gastroenterology exhibits the highest percentage of non-anomalous providers in both years, with values of 86,96% in 2022 and 86,08% in 2023. This observation underscores the considerable concordance between the IF and GAC models in categorizing typical provider behaviours.

The categorization (GAC - 0 / IF - 1 / GAF - ?) represents providers that the Isolation Forest model identified as anomalous, but which were not flagged by GAC and therefore were not initially evaluated by the GAF team. This variation is especially evident in medical specialties that have unique billing practices, such as ophthalmology and gynaecology. In the field of ophthalmology, the IF model distinctly identifies 9,96% of providers in the year 2022 and 9,62% in 2023 as anomalous, while in gynaecology, these percentages increase from 4,10% in 2022 to 11,99% in 2023.

These results underscore the IF model's increased sensitivity to detecting potential anomalies that may not be flagged by the GAC model, thereby providing a complementary perspective in the identification of outliers. Nonetheless, subsequent evaluations by the GAF team may result in shifts in classification contingent upon the evaluation outcomes. Providers exhibiting "asymmetry with potential for correction" or "asymmetry lacking justification" are reclassified under (GAC - 0 / IF - 1 / GAF - 1), aligning with the GAF's determinations. Conversely, providers identified as having a "justified asymmetry" are reclassified under (GAC - 0 / IF - 1 / GAF - 0). This dynamic reassessment underscores the iterative nature of anomaly detection and highlights the ability of models such as the IF to uncover providers deserving further examination, even when initially unnoticed by the GAC.

Among the providers identified as anomalous by both models, there is a significant concordance, as evidenced in categories (GAC - 1 / IF - 1 / GAF - 0) and (GAC - 1 / IF - 1 / GAF -

1). For instance, in cardiology, the proportion of providers deemed anomalous by both models within the category (GAC - 1 / IF - 1 / GAF - 0) diminishes slightly from 6,07% in 2022 to 4,87% in 2023, indicating stable performance over the years. The category (GAC - 1 / IF - 1 / GAF - 1), in which the agreement extends across all three evaluations, exemplifies full concordance for high-risk providers. Although the percentages remain relatively low across all specialties, they underscore the model's accuracy in detecting the most critical anomalies.

Conversely, providers deemed anomalous by the GAC model, yet not by the IF model, as exemplified in (GAC - 1 / IF - 0 / GAF - 1), illustrate situations where the GAC model discerns deviations based on its own criteria that are not identified by the IF model. Such instances are expected to be more prevalent in specialties such as Dermatology and Cardiology in the year 2022.

The column concerning Model Concordance indicates a strong alignment between the IF and GAC models across most specialties. Notably, the domain of Vascular Surgery exhibits the highest level of concordance in both years, attaining 86,63% in 2022 and rising to 91,13% in 2023, subsequently followed by Gastroenterology and Dermatology. These elevated rates emphasize the stability of both models in the consistent classification of typical and atypical providers. On the other hand, the specialty of Stomatology displays noticeable changes in concordance, attributed to the different assessment approaches impacted by the diverse datasets used for evaluation. The concordance for Stomatology increases from 58,95% in 2022 to 68,07% in 2023. However, these values remain comparatively low relative to other specialties, evidencing the disparities in assessment methodology. The observable shifts predominantly depend on the (GAC - 1 / IF - 0 / GAF - 0) category, since the GAC model identifies numerous deviant providers in this specialty. This implies that the expansive classification of deviant providers in the GAC model significantly influences model alignment, thereby highlighting the critical role of dataset variability in the interpretation of results.

The assessment of the third criterion for data mining success delineates the strengths and complementary aspects of the IF and GAC models in the detection of anomalous providers. While both models agree in classifying the majority of providers as non-anomalous, the IF model is effective in identifying outliers within specialties that exhibit complex billing patterns, such as Gynaecology and Ophthalmology. The high concordance rates observed in specialties like vascular surgery and gastroenterology underscore the robustness of both models in identifying both typical and atypical behaviours. The integration of GAF classifications provides an additional layer of validation, highlighting the necessity for comprehensive integrated model assessments to enhance the detection of healthcare fraud.

2022

	Positive	Negative
Positive	102	13
Negative	362	2712

Accuracy = 0,882
Precision = 0,220
Recall = 0,887
F1 Score = 0,352

2023

	Positive	Negative
Positive	120	21
Negative	372	2609

Accuracy = 0,874
Precision = 0,244
Recall = 0,851
F1 Score = 0,379

Figure 4.2 - Confusion Matrix and Performance Metrics for iForest

The analysis of the confusion matrix for the years 2022 and 2023 accentuates fundamental aspects of the Isolation Forest model's efficacy in identifying anomalous providers when assessed in relation to the aggregated results of GAC and GAF classifications.

The accuracy of the model remains consistently elevated, registering values of 0,822 in 2022 and 0,874 in 2023. It exhibits a strong concordance with GAF and GAC classifications in the identification of anomalous and non-anomalous providers. This high level of accuracy underscores the model's capacity to correctly classify most providers.

Recall serves as an essential metric for anomaly detection, demonstrating robustness with rates of 0,887 in 2022 and 0,851 in 2023. These results underscore the model's efficacy in identifying the most anomalous providers as recognized by GAC and GAF. Nevertheless, the observed marginal decline in 2023 implies potential for enhancement in minimizing false negatives, especially among providers categorized as (GAC - 1 / IF - 0 / GAF - 1).

The enhancement in precision from 0,220 in 2022 to 0,244 in 2023 underscores the challenge of maintaining a balance with respect to false positives. Providers identified under the categories (GAC - 0 / IF - 1 / GAF - ?) and (GAC - 1 / IF - 1 / GAF - 0) markedly contribute to the prevalence of false positives. This phenomenon reflects instances where the IF model detects potential anomalies that are either validated asymmetries or lack corroboration from GAF assessments.

The F1 score exhibits a progressive improvement, rising from 0,352 in 2022 to 0,379 in 2023. This enhancement signifies the model's augmented capability to align its detections with actual anomalies while maintaining precision at an optimal level.

Confusion matrix observations:

- True Positives: The high recall indicates the model's ability to identify anomalous providers accurately, particularly those falling under (GAC - 1 / IF - 1 / GAF - 1)

and (GAC - 0 / IF - 1 / GAF - 1). It underscores the utility of the IF model to complement the GAC model, particularly in capturing subtle anomalies.

- False Positives: The category (GAC - 0 / IF - 1 / GAF - ?) reflects providers more likely to be classified as "justified asymmetry" by GAF. Although these cases highlight the sensitivity of the IF model, they also reveal potential overclassification, which could be addressed with refined feature selection or additional domain-specific evaluation criteria.
- True Negatives: The substantial proportion of (GAC - 0 / IF - 0 / GAF - 0) cases confirms the strong alignment of IF and GAC in identifying non-anomalous providers, ensuring the model's reliability in filtering out typical provider behaviour.
- False Negatives: The instances of (GAC - 1 / IF - 0 / GAF - 1) represent a critical area for improvement, emphasizing the need for the IF model to capture better anomalies classified as deviant by GAC and GAF.

Overall, the Isolation Forest model exhibits high recall and accuracy, thereby affirming its potential as an effective instrument for the detection of anomalous providers. Nevertheless, the observed moderate precision underscores the necessity for continuous refinements to mitigate overclassification and enhance compatibility with GAC and GAF evaluations.

4.1.2 Review Process and Determine the Next Steps

The review process of the evaluation focused on the thorough assessment of the results of vascular surgery specialty for the years 2022 and 2023, conducted by the GAF team. This assessment specifically addressed two categories of concern: (GAC - 0 / IF - 1 / GAF - ?) and (GAC - 1 / IF - 1 / GAF - 0).

In the category denoted as (GAC - 0 / IF - 1 / GAF - ?), the Isolation Forest model identified nine providers in 2022 and five providers in 2023 as anomalous. The GAC model did not detect these providers; therefore, the GAF team did not initially evaluate them. Upon further examination, most of these instances were deemed to exhibit "justified asymmetry" for several reasons. Firstly, numerous providers had an insufficient client base to justify classification as anomalous. Secondly, if a provider had previously been classified under "justified asymmetry," this antecedent classification influenced subsequent year assessments, provided the current deviations were minimal. Lastly, deviations within particular medical procedures are generally not prioritized for further analysis by the GAF team, as they may not align with the team's criteria for identifying significant anomalies. However, in 2022, a provider within this category was identified as demonstrating "asymmetry with potential for correction" owing to significant billing practice deviations detected by the IF model. This instance emphasizes the necessity of reviewing this category and underscores the likelihood of the GAF team overlooking cases exclusively identified by the IF model.

Within the classification denoted as (GAC - 1 / IF - 1 / GAF - 0), the analysis concentrated on a group of four providers in 2022 and seven providers in 2023. These providers were identified

as deviant by both the GAC and IF models but were subsequently classified by the GAF team as exhibiting "justified asymmetry." A meticulous re-evaluation of these cases was conducted to validate their classifications. Following the re-assessment, all providers within this category were verified as manifesting "justified asymmetry," consistent with the underlying reasons reported in the preceding category. Notably, in numerous scenarios, the billing deviations highlighted by the IF model were related to medical act appointments, which the GAF team generally does not regard as anomalous.

To ensure a thorough and comprehensive evaluation of the findings, subsequent steps should extend the in-depth analysis employed in vascular surgery to encompass the remaining medical specialties. This procedure should prioritize providers identified by the Isolation Forest model within the (GAC - 0 / IF - 1 / GAF - ?) classification, as the GAF team has not previously scrutinized these providers, necessitating the validation of their classifications. Special attention should be directed towards identifying cases of "asymmetry with potential for correction" or "asymmetry lacking justification" that may have been disregarded by the GAC model. Furthermore, the (GAC - 1 / IF - 1 / GAF - 0) classification requires additional examination to ensure that all instances denoted as "justified asymmetry" receive adequate evaluation, particularly concerning flagged deviations related to medical acts not presently prioritized by the GAF team.

By employing the rigorous review process used in vascular surgery across all specialties, the study seeks to validate the efficacy of the Isolation Forest model in identifying anomalous providers and to refine the methodology for detecting fraudulent practices. This comprehensive approach is anticipated to enhance the robustness and reliability of the findings, thereby contributing to the development of more effective fraud detection mechanisms for Multicare.

4.2 DEPLOYMENT

Following an exhaustive assessment of the outcomes and the anomalous providers identified by the Isolation Forest model, the subsequent phase entails the GAC team evaluating how the insights from this project can improve their current fraud detection framework. This evaluation encompasses an analysis of patterns, the significance of features, and distinct billing behaviours identified by the Isolation Forest model to facilitate the more effective detection of anomalous practices.

The insights derived from this project underscore providers that may not have been identified under the GAC model yet demonstrate potentially fraudulent billing behaviours. By assimilating these findings, the GAC team can enhance the model's sensitivity to subtle deviations and augment its capacity to accurately classify anomalous practices. This integration may necessitate recalibrating thresholds, incorporating additional feature sets, or adopting new algorithms to detect outliers that correspond to real-world fraudulent behaviours.

Furthermore, the project underscores the significance of maintaining continuous feedback mechanisms between the GAC and GAF teams. The providers flagged by the Isolation Forest model present an opportunity to reevaluate the standards for classifications such as “justified asymmetry”, “asymmetry with potential for correction”, and “asymmetry lacking justification”. The findings indicate that a collaborative methodology can assist in validating detected anomalies, refining classification standards, and identifying previously unrecognized fraudulent activities.

This project emphasizes the necessity for consistent supervision of the models to accommodate the changing billing practices. Given the dynamic nature of the healthcare environment, the inclusion of findings from the Isolation Forest model into the GAC framework will maintain its applicability and strength over time. Ultimately, the integration of these methodologies seeks to enhance the ability of the GAF team to identify fraudulent providers and safeguard the integrity of Multicare.

5. CONCLUSION

This dissertation finding highlights its dual contribution: the advancement of methodologies for healthcare fraud detection, and the provision of actionable insights specifically tailored to the operational requirements of Multicare. Employing the Isolation Forest model, this research effectively demonstrated its capability to identify anomalous providers across various medical specialties, thereby addressing the critical issue of fraudulent billing practices. The findings underscore the model's potential to detect significant outliers, thus offering Multicare an extremely valuable resource for the enhancement of its current fraud detection processes.

The practical contributions of this research to Multicare are primarily situated in its capacity to augment the GAC model's accuracy in detecting abnormal billing practices. The comprehensive evaluation and analysis undertaken throughout this study have provided Multicare's GAF team with a structured framework for more effectively reviewing and assessing providers, consequently optimizing their operational processes. Additionally, the methodology formulated in this research presents the possibility of smooth integration into Multicare's existing workflows, thereby facilitating a more efficient allocation of organizational resources towards cases deemed high-risk.

From a comprehensive standpoint, the results of this study hold significance for the healthcare insurance sector in its entirety. The presented methodology offers a scalable and adaptable solution that integrates sophisticated machine learning techniques with domain expertise, highlighting how targeted anomaly detection can enhance fraud detection outcomes. The prospective adoption of these insights by Multicare positions it as a leader in utilizing data analytics to address healthcare fraud.

In conclusion, this thesis makes a significant contribution to the domain of healthcare fraud detection while simultaneously providing Multicare with a strategic advantage by furnishing it with a robust and practical tool for defending against fraudulent billing practices. The integration of advanced data-driven techniques tailored to Multicare's specific requirements highlights the significance of this research, offering a framework for both immediate implementation and future innovation.

6. LIMITATIONS AND RECOMMENDATIONS FOR FUTURE WORKS

This research provides substantial insights into the identification of fraudulent providers through anomaly detection techniques; however, specific limitations emerged during the investigative process, meriting consideration in subsequent studies.

A significant limitation lies in the absence of detailed clinical information that delineates which medical acts and episodes are indicative of fraudulent billing practices. Certain medical acts are not deemed anomalous based on the established experience of abnormal billing detection. On the other hand, some acts exhibit deviations that require evaluation through clinical knowledge to determine if they justify further analysis. This gap in understanding obstructs the capability to accurately identify anomalous billing behaviours, highlighting the necessity for a more comprehensive mapping of clinical and billing anomalies. Additionally, a limitation arises from the GAC model's lack of full alignment with real-world fraud detection criteria. While the methodology was grounded in the expertise of the GAF team, the main authority in identifying fraudulent providers, the datasets used were based on the GAC framework. This misalignment may result in certain anomalous billing practices identified by the GAF team not being effectively captured by the GAC model, potentially leading to under-detection of fraudulent providers and missed billing anomalies. Addressing this issue requires continued collaboration between the GAC and GAF teams to ensure that the model is better aligned with practical fraud detection criteria, thereby enhancing the accuracy of detection and mitigating financial losses for Multicare.

Future research efforts like this project should emphasize enhancing the integration of clinical and business knowledge within fraud detection frameworks. A more profound dependence on the business expertise of domain experts is imperative for the refinement of the model's inputs and outputs. Their proficiency in identifying fraudulent activities guarantees that the model closely aligns with practical fraud detection requirements. In addition, future projects should focus on removing or filtering out unnecessary data, like medical act appointments, as they often add noise and increase the rate of false positives. Enhanced collaboration between teams ensures that models are effectively aligned with the pragmatic demands of fraud detection and are capable of more accurately identifying anomalous billing practices.

Grounded in the findings and methodologies of this project, subsequent research endeavours might investigate novel pathways to enhance the understanding of fraudulent behaviours. Expanding the analytical scope to include inpatient billing practices offers a promising opportunity. This expansion could ascertain whether inpatient behaviours are anomalous and warrant further scrutiny. Utilization of clustering techniques on providers identified as anomalous may reveal patterns or similarities among fraudulent practitioners, thereby offering enhanced insights into their billing activities. Another significant research direction involves examining the interactions of each fraudulent provider to discern recurring patterns, justifications, or rationales underlying their fraudulent billing acts. Such profound contextual

analysis would augment the understanding of fraud mechanisms and strengthen the development of more robust and focused fraud mitigation strategies.

By addressing the identified limitations and pursuing the recommended research directions, upcoming studies have the potential to refine existing methodologies, enhance inter-team collaboration, and construct models with increased efficacy in detecting fraudulent billing practices. Such advancements will significantly augment Multicare's sustained efforts to mitigate financial losses and fortify its fraud detection framework.

BIBLIOGRAPHICAL REFERENCES

Aggarwal, C. C. (2015). *Data Mining: The Textbook* (Vol. 1). New York: Springer.

Agrawal, S., & Agrawal, J. (2015). Survey on Anomaly Detection using Data Mining Techniques. *Procedia Computer Science, 60*, 708-713. doi: https://doi.org/10.1016/j.procs.2015.08.220

Bairy, M., Muniyal, B., & Shetty, N. P. (2024). Enhancing healthcare data integrity: fraud detection using unsupervised learning techniques. *International Journal of Computers and Applications*, 1006–1019. doi:10.1080/1206212X.2024.2408262

Capelleveen, G. C. (2013). Retrieved from http://essay.utwente.nl/64417/

Celebi, M. E. (2016). *Unsupervised learning algorithms* (Vol. 9). Springer.

Chandore, P., & Chatur, P. (n.d.). Outlier detection techniques over streaming data in data mining: A research perspective. *International Journal of Recent Technology and Engineering (IJRTE), 2*(1), 157-162.

Du, J., & Yu, B. (2023). Application of Isolation Forest algorithm in fraud detection of medical insurance big data. *2023 8th International Conference on Intelligent Informatics and Biomedical Sciences (ICIIBMS). 8*, pp. 504-509. IEEE.

Dy, J. G., & Brodley, C. E. (2004, Aug). Feature selection for unsupervised learning. *Journal of Machine Learning Research, 5*, 845-889.

Ekina, T., Leva, F., Ruggeri, F., & Soyer, R. (2013). Application of Bayesian methods in detection of healthcare fraud. *Chemical Engineering Transaction, 33*.

Fursov, I., Kovtun, E., Rivera-Castro, R., Zaytsev, A., Khasyanov, R., Spindler, M., & Burnaev, E. (2022). Sequence embeddings help detect insurance fraud. *IEEE Access, 10*, 32060-32074.

Gao, Y., Sun, C., Li, R., Li, Q., Cui, L., & Gong, B. (2018). An efficient fraud identification method combining manifold learning and outliers detection in mobile healthcare services. *IEEE Access, 6*, 60059-60068. doi:10.1109/ACCESS.2018.2875516

Hamid, Z., Khalique, F., Mahmood, S., Daud, A., Bukhari, A., & Alshemaimri, B. (2024). Healthcare insurance fraud detection using data mining. *BMC Medical Informatics and Decision Making, 24*(1), 112.

Han, J., Kamber, M., & & Pei, J. (2012). *Data Mining: Concepts and Techniques.* doi: https://doi.org/10.1016/C2009-0-61819-5

Hansson, A., & Cedervall, H. (2022). *Insurance fraud detection using unsupervised sequential anomaly detection.*

He, H., Hawkins, S., Graco, W., & Yao, X. (2000). Application of genetic algorithm and k-nearest neighbour method in real world medical fraud detection problem. *Journal of Advanced Computational Intelligence and Intelligent Informatics, 4*(2), 130-137.

He, H., Wang, J., Graco, W., & Hawkins, S. (1997). Application of neural networks to detection of medical fraud. *Expert Systems with Applications, 13*(4), 329-336. doi:https://doi.org/10.1016/S0957-4174(97)00045-6

Jain, A. K., Murty, M. N., & Flynn, P. J. (1999). Data clustering: a review. *ACM Comput. Surv., 31*(3), 264–323. doi:10.1145/331499.331504

Jolliffe, I. T., & Cadima, J. (2016). Principal component analysis: a review and recent developments. *Philosophical Transactions. Series A, Mathematical, Physical, and Engineering Sciences, 374*(2065), 20150202. doi:https://doi.org/10.1098/rsta.2015.0202

Joshi, A., Soni, S., & Jain, V. (2021). An Experimental Study using Unsupervised Machine Learning Techniques for Credit Card Fraud Detection. *GIS SCIENCE JOURNAL, 8*(5), 1869-9391.

Kelleher, J. D., Mac Namee, B., & D'Arcy, A. (2015). *Fundamentals of Machine Learning for Predictive Data Analytics.* London, England: MIT Press books.

Liu, F. T., Ting, K. M., & Zhou, Z.-H. (2008). Isolation Forest. *Eighth IEEE International Conference on Data Mining*, (pp. 413-422). Pisa, Italy. doi:10.1109/ICDM.2008.17

Liu, F. T., Ting, K. M., & Zhou, Z.-H. (2012, March). Isolation-Based Anomaly Detection. *ACM Transactions on Knowledge Discovery from Data, 6*(1), 3. doi:https://doi.org/10.1145/2133360.2133363

Liu, Q., & Vasarhelyi, M. (2013). Healthcare fraud detection: A survey and a clustering model incorporating geo-location information. *29th world continuous auditing and reporting symposium (29WCARS).* Brisbane, Australia.

Multicare - Seguros De Saúde, S.A. (2022). *Gabinete de Atuariado e Controlo Técnico.* Multicare - Seguros De Saúde, S.A.

Multicare - Seguros De Saúde, S.A. (2023). Multicare - Seguros De Saúde, S.A.

Multicare - Seguros De Saúde, S.A. (2024). *Definição Funcional do Gabinete AntiFraude (GAF).* Multicare - Seguros De Saúde, S.A.

Nassif, A. B. (2021). Machine Learning for Anomaly Detection: A Systematic Review. *IEEE Access*, 78658-78700.

Portela, F. G. (2019). *2019 IEEE International Conference on Applied Science and Advanced Technology (iCASAT).*

Provost, F., & Fawcett, T. (2013). *Data Science for Business: What You Need to Know about Data Mining and Data-Analytic Thinking.* O'Reilly Media, Inc.

Raja, P. T. (2020). Missing value imputation using unsupervised machine learning techniques. *Soft Computing, 24*, 4361–4392. doi:10.1007/s00500-019-04199-6

Ramos, A., Pontes, B., Cordeiro, M., Barceló, M., & Coelho, P. (2022). *EMPLOYEE HANDBOOK - Gabinete de Atuariado e Controlo.* Multicare - Seguros De Saúde, S.A.

Raschka, S. (2018). *Model evaluation, model selection and algorithm selection in machine learning.*

Rayan, N. (2019). Framework for Analysis and Detection of Fraud in Health Insurance. *2019 IEEE 6th International Conference on Cloud Computing and Intelligence Systems (CCIS)*, (pp. pp. 47-56.). Singapore. doi:doi: 10.1109/CCIS48116.2019.9073700

Rosenbaum, S., Lopez, N., & Stifler, S. (2009). *Health insurance fraud: An overview.* Washington, D.C.

Sathya Narayana, M., S. Prasad, B. V., Srividhya, A., & Pandu Rang Reddy, K. (2011). Data Mining Machine Learning Techniques - A Study on Abnormal Anomaly Detection System. *International Journal of Computer Science and Telecomunications, 2*(6).

Shan, Y., Jeacocke, D., Murray, D. W., & Sutinen, A. (2008). Mining medical specialist billing patterns for health service management. *Proceedings of the 7th Australasian Data Mining Conference*, *87*, pp. 105-110.

Shan, Y., Murray, D. W., & Sutinen, A. (2009). Discovering inappropriate billings with local density based outlier detection method. *Proceedings of the Eighth Australasian Data Mining Conference*, (pp. 93-98).

Shearer, C. (2000). The CRISP-DM model: the new blueprint for data mining. *Journal of data warehousing, 5*, 13-22.

Shin, H., Park, H., Lee, J., & Jhee, W. C. (2012). A scoring model to detect abusive billing patterns in health insurance claims. *Expert Systems with Applications, 39*, 7441-7450. doi:https://doi.org/10.1016/j.eswa.2012.01.105

Sumalatha, M. R., & Prabha, M. (2019). Mediclaim Fraud Detection and Management Using Predictive Analytics. *2019 International Conference on Computational Intelligence*

and Knowledge Economy (ICCIKE), (pp. pp. 517-522). Dubai, United Arab Emirates. doi:doi: 10.1109/ICCIKE47802.2019.9004241

Thornton, D., Mueller, R. M., Schoutsen, P., & van Hillegersberg, J. (2013). Predicting healthcare fraud in Medicaid: A multidimensional data model and analysis techniques for fraud detection. *Procedia Technology, 9*, 1252-1264. doi:10.1016/j.protcy.2013.12.140

Timofeyev, Y., & Jakovljevic, M. (2022). Editorial: Fraud and Corruption in Healthcare. *Frontiers in Public Health, 10*. doi:doi: 10.3389/fpubh.2022.921254

Viaene, S., & Dedene, G. (n.d.). Insurance Fraud: Issues and Challenges. *The Geneva Papers on Risk and Insurance - Issues and Practice, 29*(2), 313-333. doi:doi: 10.1111/j.1468-0440.2004.00290.x.

Yang, W.-S., & Hwang, S.-Y. (2006). A process-mining framework for the detection of healthcare fraud and abuse. *Expert Systems with Applications, 31*(1), 56-68. doi:https://doi.org/10.1016/j.eswa.2005.09.003

Yousefpour, A., Shishehbor, M., Foumani, Z. Z., & Bostanabad, R. (n.d.). Unsupervised Anomaly Detection via Nonlinear Manifold Learning. Retrieved from https://arxiv.org/abs/2306.09441

Zhou, Z.-H. (2016). *Machine Learning.* Springer Nature.

APPENDIX A

Medical Acts Dataset Feature Description:

COD_PRESTADOR	Set of numbers that identifies the Provider.
REDE	Name of the healthcare network the insured person used.
CLUSTER_PRESTADOR	Cluster predefined by Multicare describes the type of medical procedure associated with the client provider (outpatient and inpatient).
COD_NOME_DESPESA_HOMOG	Code of simplified expenditure that allows equivalent acts to be grouped together. (Correspondence provided by GAF).
AMB_SUM_VAP	Total amount of medical act expenditure presented to the provider (outpatient sample).
AMB_SUM_VAP_STD	Total amount shown for medical act expenditure at the provider but standardized by network (outpatient sample).
AMB_SUM_SPE	Total amount paid for medical act expenses at the provider (outpatient sample).
AMB_SUM_QTD_ATOS	Total number of acts at the provider (outpatient sample).
AMB_N_CL_PREST_ATO	Total number of clients who incurred the medical act expense at the provider under study (outpatient sample).
AMB_ATOS_CLT_AMB	Average number of medical acts performed per client for a specific provider within an outpatient sample.
AMB_N_CL_PREST	Total internal customers at the provider under study (outpatient sample).
AMB_FREQM_CL_ATO_PREST	Proportion of clients incurring medical act expenses at a specific provider relative to the provider's total internal outpatient clientele, but only if the provider has ten or more clients incurring these expenses.

	$\begin{cases} 0, & \text{AMB\_N\_CL\_PREST\_ATO} < 10 \\ \dfrac{\text{AMB\_N\_CL\_PREST\_ATO}}{\text{AMB\_N\_CL\_PREST}}, & \text{AMB\_N\_CL\_PREST\_ATO} \geq 10 \end{cases}$
AMB_CMC_ATO_PREST	Average expenditure on medical acts per client for a given provider in an outpatient sample. $\dfrac{\text{AMB\_SUM\_VAP}}{\text{AMB\_N\_CL\_PREST\_ATO}}$
AMB_CMC_ATO_PREST_IMP_N	Adjusted average expenditure per client on medical acts for a provider within an outpatient sample. It applies conditions to ensure data reliability and relevance to the type of medical procedure. Specifically, if fewer than ten clients incurred expenses, the value is set to zero. If there are ten or more clients, it further checks if the provider's average expenditure per client falls below the median expenditure for similar medical procedures within the same predefined cluster designated by Multicare. If below the median, the value is also set to zero; otherwise, it returns the average expenditure. This variable helps in identifying providers whose client expenditures align or exceed typical expenditure patterns within a defined procedural category. $\begin{cases} 0, & \text{AMB\_N\_CL\_PREST\_ATO} < 10 \\ 0, & \text{AMB\_CMC\_ATO\_PREST} < \text{MEDIAN}(\text{AMB\_CMC\_ATO\_PREST} \mid \text{CLUSTER\_PRESTADOR} = \text{n}) \\ \text{AMB\_CMC\_ATO\_PREST}, & \end{cases}$
AMB_N_CL_REDE_ATO	Total number of customers who made a given medical act expenditure on the network (outpatient sample).
AMB_N_CL_REDE	Total internal network customers (outpatient sample).
AMB_FREQ_CL_ATO_PREST	Indicates the frequency of clients who have incurred that medical act expense out of the total number of internal clients (outpatient sample).
AMB_FREQ_CL_ATO_REDE	Indicates the frequency of clients who have incurred that medical act expense out of the total number of internal clients in the network (outpatient sample).
DIFF_AMB_FREQ_CL_ATO_PREST_REDE	Adjusted difference in frequency of clients incurring medical act expenses at a specific provider compared

	to the broader network, but only when the provider has ten or more clients incurring such expenses. $\begin{cases} 0, & \text{AMB\_FREQM\_CL\_ATO\_PREST} = 0 \\ \max(0, \text{AMB\_FREQ\_CL\_ATO\_PREST} - \text{AMB\_FREQ\_CL\_ATO\_REDE}), \end{cases}$
AMB_Significativo	Indicates whether the difference in the frequency of use of medical act expenditure at the provider under study compared to the use of the same medical act expenditure in the network is statistically significant, according to GAC team (outpatient sample).

Episodes Dataset Feature Description:

COD_PRESTADOR	Set of numbers that identifies the Provider.
REDE	Name of the healthcare network the insured person used.
CLUSTER_PRESTADOR	Cluster predefined by Multicare describes the type of medical procedure associated with the client provider (outpatient and inpatient).
CONCATENA_DESP_HOMG_EPISODIO	Profile/Set of simplified expenses, which allows equivalent acts to be grouped together, carried out at this provider, in outpatient coverage.
AMB_SUM_VAP	Total of the value presented for the episodes expense profile at the provider (outpatient sample).
AMB_SUM_VAP_STD	Total amount shown for episodes expenditure at the provider but standardized by network (outpatient sample).
AMB_SUM_SPE	Total amount paid for episodes expenses at the provider (outpatient sample).
AMB_SUM_QTD_ATOS	Total number of acts at the provider (outpatient sample).
AMB_N_CL_PERFIL_PREST	Total number of clients who underwent the episodes expenditure profile at the provider under study (outpatient sample).

AMB_ATOS_CLT_AMB	Average number of medical acts performed per client for a specific provider within an outpatient sample.
AMB_N_CL_PREST	Total internal customers at the provider under study (outpatient sample).
AMB_FREQ_CL_PERFIL_PREST	Proportion of clients incurring episodes expenses at a specific provider relative to the provider's total internal outpatient clientele, but only if the provider has ten or more clients incurring these expenses. $\begin{cases} 0, & \text{AMB\_N\_CL\_PERFIL\_PREST} < 10 \\ \frac{\text{AMB\_N\_CL\_PERFIL\_PREST}}{\text{AMB\_N\_CL\_PREST}}, & \text{AMB\_N\_CL\_PERFIL\_PREST} \geq 10 \end{cases}$
AMB_CMC_PERFIL_PREST	Average expenditure on episodes per client for a given provider in an outpatient sample. $\frac{\text{AMB\_SUM\_VAP}}{\text{AMB\_N\_CL\_PERFIL\_PREST}}$
AMB_CMC_PERFIL_PREST_IMP_N	Adjusted average expenditure per client on episodes for a provider within an outpatient sample. It applies conditions to ensure data reliability and relevance to the type of medical procedure. Specifically, if fewer than ten clients incurred expenses, the value is set to zero. If there are ten or more clients, it further checks if the provider's average expenditure per client falls below the median expenditure for similar medical procedures within the same predefined cluster designated by Multicare. If below the median, the value is also set to zero; otherwise, it returns the average expenditure. This variable helps in identifying providers whose client expenditures align or exceed typical expenditure patterns within a defined procedural category. $\begin{cases} 0, & \text{AMB\_N\_CL\_PERFIL\_PREST} < 10 \\ 0, & \text{AMB\_CMC\_PERFIL\_PREST} < \\ \text{MEDIAN}(\text{AMB\_CMC\_PERFIL\_PREST} \mid \text{CLUSTER\_PRESTADOR} = \text{n}) & \\ \text{AMB\_CMC\_PERFIL\_PREST}, & \end{cases}$
AMB_N_EP_PERFIL_PREST	Total number of episodes of a given episode expenditure profile at the provider (outpatient sample).
AMB_N_EP_PREST	Total episodes carried out at the provider (outpatient sample).

AMB_CMEP_PERFIL_PREST	Average expense per episode for a specific expenditure profile at a given provider within an outpatient sample. $$\frac{\text{AMB\_SUM\_VAP}}{\text{AMB\_N\_EP\_PERFIL\_PREST}}$$
AMB_N_EP_PERFIL_REDE	Total number of episodes of a given episode expenditure profile in the network (outpatient sample).
AMB_N_EP_REDE	Total episodes carried out by internal clients in the network (outpatient sample).
AMB_FREQ_EP_PERFIL_PREST_REDE	Proportion of episodes of a specific expenditure profile that occur at a particular provider relative to the total episodes of the same profile within the entire network in an outpatient sample. $$\frac{\text{AMB\_N\_EP\_PERFIL\_PREST}}{\text{AMB\_N\_EP\_PERFIL\_REDE}}$$
AMB_FREQ_CL_PERFIL_PREST	Indicates the frequency of clients who have incurred that episode expense out of the total number of internal clients (outpatient sample).
AMB_FREQ_CL_PERFIL_REDE	Indicates the frequency of clients who have incurred that episode expense out of the total number of internal clients in the network (outpatient sample).
DIFF_AMB_FREQ_CL_PERFIL_PREST_REDE	Adjusted difference in frequency of clients incurring episode expenses at a specific provider compared to the broader network, but only when the provider has ten or more clients incurring such expenses. $$\begin{cases} 0, & \text{AMB\_FREQM\_CL\_PERFIL\_PREST} = 0 \\ \max(0, \text{AMB\_FREQ\_CL\_PERFIL\_PREST} - \text{AMB\_FREQ\_CL\_PERFIL\_REDE}), \end{cases}$$
AMB_Significativo	Indicates whether the difference in the frequency of use of episode expenditure at the provider under study compared to the use of the same episode expenditure in the network is statistically significant, according to GAC team (outpatient sample).

Total Cost Dataset Feature Description:

COD_PRESTADOR	Set of numbers that identifies the Provider.
REDE	Name of the healthcare network the insured person used.
CLUSTER_PRESTADOR	Cluster predefined by Multicare describes the type of medical procedure associated with the client provider (outpatient and inpatient).
CLIENTES	Total number of clients of the provider (outpatient sample).
AMB_SUM_VAP	Total of the value presented for the total expense profile at the provider (outpatient sample).
AMB_SUM_VAP_STD	Total amount shown for total expenditure at the provider but standardized by network (outpatient sample).
AMB_CMC	Average expenditure per client for a given provider in an outpatient sample.
AMB_CMC_PREST_IMP_N	Adjusted average expenditure per client for a provider within an outpatient sample. It applies conditions to ensure data reliability and relevance to the type of medical procedure. Specifically, if fewer than ten clients incurred expenses, the value is set to zero. If there are ten or more clients, it further checks if the provider's average expenditure per client falls below the median expenditure for similar medical procedures within the same predefined cluster designated by Multicare. If below the median, the value is also set to zero; otherwise, it returns the average expenditure. This variable helps in identifying providers whose client expenditures align or exceed typical expenditure patterns within a defined procedural category. $\begin{cases} 0, & \text{CLIENTES} < 10 \\ 0, & \text{AMB\_CMC} < \text{MEDIAN}(\text{AMB\_CMC} \mid \text{CLUSTER\_PRESTADOR} = \text{n}) \\ \text{AMB\_CMC}, & \end{cases}$
AMB_CMC_REDE	Average expenditure per client in the network (outpatient sample).

FREQ_CMC_REAL_PREST_REDE	Relative expenditure per client at a given provider compared to the network average for an outpatient sample, but only if the provider has at least ten clients incurring medical act expenses. $\begin{cases} 0, & \text{CLIENTES} < 10 \\ 0, & \frac{\text{AMB\_CMC}}{\text{AMB\_CMC\_REDE}} < 1 \\ \frac{\text{AMB\_CMC}}{\text{AMB\_CMC\_REDE}}, & \end{cases}$
AMB_CM_PADRAO	Standardized average cost adjusted with a 95% confidence interval. It provides an estimate of the average cost while accounting for the variability in the data, ensuring that the value falls within a range with a 95% probability.
FREQ_CMC_PREST_PADRAO	Relative expenditure per client at a specific provider compared to a standardized average cost adjusted with a 95% confidence interval for the outpatient sample, providing a robust comparison by incorporating data variability. $\begin{cases} 0, & \text{CLIENTES} < 10 \\ 0, & \frac{\text{AMB\_CMC}}{\text{AMB\_CM\_PADRAO}} < 1 \\ \frac{\text{AMB\_CMC}}{\text{AMB\_CM\_PADRAO}}, & \end{cases}$
AMB_Significativo	Indicates whether the difference in the total cost of expenditures at the provider under study, compared to the total cost of the expenditures in the network, is statistically significant, according to the GAC team (outpatient sample).

APPENDIX B

Episodes Dataset Feature Selection Versions

Version 1	Version 2
CLUSTER_PRESTADOR	AMB_FREQ_CL_PERFIL_PREST
AMB_N_CL_PERFIL_PREST	AMB_CMC_PERFIL_PREST_IMP_N
AMB_CMC_PERFIL_PREST_IMP_N	AMB_CMEP_PERFIL_PREST
AMB_FREQ_CL_PERFIL_PREST	DIFF_AMB_FREQ_PREST_REDE
AMB_N_EP_PERFIL_PREST	
AMB_CMEP_PERFIL_PREST	
AMB_N_EP_PERFIL_REDE	
AMB_FREQ_EP_PERFIL_PREST_REDE	
AMB_CMEP_PERFIL_REDE	
AMB_FREQ_EP_PERFIL_REDE	
DIFF_AMB_FREQ_PREST_REDE	

Medical Acts Dataset Feature Selection Versions

Version 1	Version 2
CLUSTER_PRESTADOR	AMB_ATOS_CLT_AMB
AMB_ATOS_CLT_AMB	AMB_CMC_PERFIL_PREST_IMP_N
AMB_N_CL_PREST_ATO	AMB_FREQM_CL_ATO_PREST
AMB_CMC_ATO_PREST_IMP_N	DIFF_AMB_FREQ_CL_ATO_PREST_REDE
AMB_FREQM_CL_ATO_PREST	
AMB_CMC_ATO_PREST	
AMB_FREQ_CL_ATO_PREST	
DIFF_AMB_FREQ_CL_ATO_PREST_REDE	

Total Cost Dataset Feature Selection Versions

Version 1	Version 2
CLUSTER_PRESTADOR	AMB_CMC_PREST_IMP_N
AMB_CMC_PREST_IMP_N	FREQ_CMC_REAL_PREST_REDE
FREQ_CMC_REAL_PREST_REDE	FREQ_CMC_PREST_PADRAO
FREQ_CMC_PREST_PADRAO	

APPENDIX C

Dear Diogo Lamy,

Dear Professor Jorge Bravo,

Thank you for filling in the Research Ethics Checklist. After reviewing your request, you can proceed with the study as we do not foresee any major ethical concerns with the project.

Project No.: **DSCI2024-7-224833**

Project Title: **Advanced Anomaly Detection Models for Uncovering Fraudulent Healthcare Insurance Providers**

Principal Researcher: **Diogo Lamy**

according to the regulations of the Ethics Committee of NOVA IMS and MagIC Research Center this project was considered to meet the requirements of the NOVA IMS Internal Review Board, being considered **APPROVED** on 29/11/2024.

It is the Principal Researcher's responsibility to ensure that all researchers and stakeholders associated with this project are aware of the conditions of approval and which documents have been approved.

The Principal Researcher is required to notify the Ethics Committee, via amendment or progress report, of

- Any significant change to the project and the reason for that change;

- Any unforeseen events or unexpected developments that merit notification;

- The inability of the Principal Researcher to continue in that role or any other change in research personnel involved in the project.

Lisbon, 29/11/2024

NOVA IMS Ethics Committee

ethicscommittee@novaims.unl.pt

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa