T.C.

ISTANBUL AYDIN UNIVERSITY

INSTITUTE OF GRADUATE STUDIES

MACHINE LEARNING BASED FOR INSURANCE CLAIMS FRAUD DETECTION TECHNIQUE FOR SAFEGUARDING THE HEALTH INDUSTRY

MASTER'S THESIS

Ghina ÖZDEMİR

Department of Computer Engineering

Computer Engineering Program

AUGUST, 2024

T.C.
ISTANBUL AYDIN UNIVERSITY
INSTITUTE OF GRADUATE STUDIES

MACHINE LEARNING BASED FOR INSURANCE CLAIMS FRAUD DETECTION TECHNIQUE FOR SAFEGUARDING THE HEALTH INDUSTRY

MASTER'S THESIS

Ghina ÖZDEMİR
(Y2213.011007)

Department of Computer Engineering
Computer Engineering Program

Thesis Advisor: Prof. Dr. Rafet AKDENİZ
Co-Advisor: Teaching Asst. Dr. Mhd Wasim RAED

AUGUST, 2024

THESIS EXAM REPORT

Istanbul Aydın University Institute of Graduate Studies Board of Directors 22.07.2024 date and 2024/12 The thesis of Machine Learning Based For Insurance Claims Fraud Detection Technique For Safeguarding The Health Industry, whose thesis defense exam was held on 9/8/2024 before the jury members formed at the meeting no. Unanimity* and Acceptance** decision was made.

JURY

1st Member (Thesis Advisor)	: Prof. Dr. Rafet AKDENIZ
2nd Member (Co-Advisor)	: Instructor (Ph.D.) Mhd Wasim RAED
3rd Member	: Assoc. Prof. (Ph.D.) Ilham HUSEYINOV
4th Member	: Dr. Osman SELVI

APPROVAL

Istanbul Aydin University Institute of Graduate Studies Board of Directors

…….. date and ……. decision no.

(*) Unanimity/Majority vote will be written in writing.

(**) Acceptance decision will be written in writing.

DECLARATION

I hereby declare with respect that the study “Machine Learning Based for Insurance Claims Fraud Detection Technique for Safeguarding the Health Industry”, which I submitted as a Master thesis, is written without any assistance in violation of scientific ethics and traditions in all the processes from the Project phase to the conclusion of the thesis and that the works I have benefited are from those shown in the Bibliography. (…./08/2024)

Ghina ÖZDEMIR

FOREWORD

At the outset of this academic journey, I am obliged to express my deep gratitude to those whose unwavering support and valuable contributions made this thesis project possible. It would not have been possible without the guidance and direction of Dr. Rafet Akdeniz and Dr. Mhd Wasim Raed, whose expertise and encouragement have shaped the course of this research he has helped me a lot at this period. And I am grateful for the understanding, patience and encouragement of my family and friends during my search the intense knowledge. and how they supported me to complete my thesis.

August, 2024 Ghina ÖZDEMIR

MACHINE LEARNING BASED FOR INSURANCE CLAIMS FRAUD DETECTION TECHNIQUE FOR SAFEGUARDING THE HEALTH INDUSTRY

ABSTRACT

Insurance claims fraud occurs when individuals or groups knowingly mislead insurance companies to get benefits or payments to which they are not legally entitled. Regarding various insurance products, like life, health, vehicle, and property insurance, this type of fraud can take many different forms. The integrity of healthcare systems around the world is seriously threatened by healthcare insurance claims fraud, which can make a lot of losses in the financial sector and subpar patient treatment. In this research the aim is to use Machine Learning and Explainable AI (XAI) to detect and prevent fraud in healthcare insurance claims. Also, to ensure transparency and confidence, provide detailed explanations of how decisions are made, which will aid in understanding and validating the detection process. By using Medicare dataset from Kaggle with four main files namely, inpatient, outpatient, beneficiary, and target data, an integrated dataset is created This dataset is basis for supervised classification problems, where the objective is to classify health care providers as fraudulent or not. Several machine learning algorithms have been used for training, including logistic regression (LR), random forest (RF), decision tree (DT), support vector machine (SVM), and XGBoost. In addition, hard and soft voting ensemble learning techniques are used to combine all the models into more powerful model. Also, the feature selection approach is used to enhance model performance and to compare the results. Additionally, the study investigates the effectiveness of artificial-neural networks, particularly deep learning, in detecting fraud in health insurance claims. Notably, logistic regression appears as a high-performance algorithm, exhibiting high accuracy, F1 scores, and AUC scores. In this research is further illustrated using the SHapley Additive explanation (SHAP), which is a model interpretation technique that enhances the understanding of logistic

regression prediction.

Keywords: Machine Learning, Insurance Claims, Fraud Detection, Explainable AI, SHAP.

SAĞLIK SEKTÖRÜNÜ KORUMA AMAÇLI MAKINE EĞITIMI TABANLI SIGORTA TALEP DOLANDIRICILIĞI TESPIT TEKNIĞI

ÖZET

Sigorta talepleri sahtekarlığı, bireyler veya gruplar, sigorta şirketlerini yasal olarak hak sahibi olmadıkları menfaatleri veya ödemeleri almaları için bilerek yanılttığında ortaya çıkar. Hayat, sağlık, araç ve mülk sigortası gibi çeşitli sigorta ürünleriyle ilgili olarak, bu tür dolandırıcılık birçok farklı biçimde olabilir. Dünyadaki sağlık sistemlerinin bütünlüğü, sağlık sigortası talep sahtekarlığı nedeniyle ciddi şekilde tehdit altındadır ve bu da finans sektöründe çok fazla zarara neden olabilir ve hasta tedavisini aşabilir. Bu araştırmada amaç, sağlık sigortası taleplerinde sahtekarlığı tespit etmek ve önlemek için Makine Öğrenimini ve Açıklanabilir Yapay Zekayı (XAI) kullanmaktır. Ayrıca, şeffaflığı ve güveni sağlamak için, tespit sürecinin anlaşılmasına ve doğrulanmasına yardımcı olacak kararların nasıl alındığına dair ayrıntılı açıklamalar sağlayın. Kaggle'ın Medicare veri kümesini yatan hasta, ayakta tedavi, yararlanıcı ve hedef veriler olmak üzere dört ana dosyayla kullanarak entegre bir veri kümesi oluşturulur. Bu veri kümesi, amacın sağlık hizmeti sağlayıcılarını hileli olarak sınıflandırmak olduğu denetimli sınıflandırma sorunlarının temelidir. Eğitim için lojistik regresyon (LR), rastgele orman (RF), karar ağacı (DT), destek vektör makinesi (SVM) ve XGBoost dahil olmak üzere çeşitli makine öğrenimi algoritmaları kullanılmıştır. Ek olarak, tüm modelleri daha güçlü bir modelde birleştirmek için sert ve yumuşak oylama topluluğu öğrenme teknikleri kullanılır. Ayrıca, özellik seçimi yaklaşımı, modelin performansını iyileştirmek ve sonuçları karşılaştırmak için kullanılır. Ek olarak, çalışma, yapay sinir ağlarının, özellikle derin öğrenmenin, sağlık sigortası taleplerinde sahtekarlığı tespit etmedeki etkinliğini araştırmaktadır. Özellikle, lojistik regresyon, yüksek doğruluk, F1 puanları ve AUC puanları sergileyen yüksek performanslı bir algoritma olarak görünür. Bu araştırmada, lojistik regresyon

tahmininin anlaşılmasını geliştiren bir model yorumlama tekniği olan SHapley Katkı açıklaması (SHAP) kullanılarak daha da gösterilmiştir.

Anahtar Kelimeler: Makine Öğrenimi, Sigorta Talepleri, Dolandırıcılık Tespiti, Açıklanabilir Yapay Zeka, SHAP.

TABLE OF CONTENTS

LIST OF ABBREVIATIONS

DT : Decision Tree

SVM : Support Vector Machine

LR : Logistic Regression

RF : Random Forest

XGB : eXtreme Gradient Boosting

EDA : Exploratory Data Analysis

PCA : Principal Component Analysis

ANN : Artificial Neural Network

SMOTE: Synthetic Minority Oversampling Technique

TP : True Positives

FP : False Positives

TN : True Negatives

FN : False Negatives

SHAP : SHapley Additive exPlanations

ROC : Receiver Operating Characteristic

AUC : Area under the ROC Curve

CSV : Comma Separated Values

WSGI : Web Server Gateway Interface

UI : User Interface

LIST OF TABLES

LIST OF FIGURES

I. INTRODUCTION

A. Background

Fraud is defined as deliberate and deceptive activity intended to deprive a victim of their right or to profit the offender with illegal gain.

Fraud, present in various forms, poses a significant global threat to numerous industries. Credit card fraud, identity theft, insurance claims fraud, all these malicious activities entail severe consequences on financial stability, reputation, and personal well-being. Traditional rule-based systems designed for detecting fraudulent behavior have proven to be inadequate in adapting to the ever-evolving tactics employed by fraudsters. This is where machine learning emerges as a transformative force capable of revolutionizing the field.

Machine learning algorithms, powered by extensive datasets and computational prowess, excel in the identification of patterns, anomalies, and deviations from the norm. By assimilating historical data and swiftly adapting to emerging threats in real-time, they possess the capability to identify fraudulent behaviors and transactions. This dynamic approach not only enhances detection accuracy but also mitigates false positives, thus minimizing any adverse impact on legitimate users. Since black-box models are the foundation of most fraud detection systems, it is even more important to investigate how to make the logic and outputs of these systems comprehensible and reliable to non-AI professionals.

Explainable AI (XAI) is the branch of artificial intelligence that focuses on creating understandable representations of complicated "black-box" AI algorithms and system outputs. As an alternative, it makes use of so-called "white-box" AI algorithms, which are predicated on easily understood and naturally explicable models (Ji, 2021).

One of the most popular techniques for deciphering certain predictions produced by black-box classifiers is SHAP. The goal of this program is to understand the reasoning behind each prediction. By altering a single data point and seeing the

effect on the classifier's output, it clarifies the relative contributions of various qualities to a prediction.

B. Problem Definition

Fraud in claims, one of the most pressing concerns confronting the insurance industry which results in the substantial losses it incurs. The cost of identifying or anticipating potential fraud in any claim submitted is very high because, if done poorly, it could annoy real consumers and cause claims adjudication to be delayed. Healthcare fraud is an organized crime that involves peers making compelled claims. including beneficiaries, physicians, and providers. An approved individual or organization that healthcare providers are individuals who give medical care or treatment. Physicians, radiologists, nurse practitioners, medical supply companies, urgent care centers, hospitals, laboratories, and others, establishments, and companies offering these services are examples of providers. One of the most pressing concerns facing Medicare today is provider fraud. For instance, an uninsured individual may fraudulently gain insurance benefits by passing for an insured person, or a provider could incorrectly bill a member for an excessive amount that would subsequently be settled with the insurance company.

The following categories broadly describe health insurance fraud claims:

• Upcoding of items: Charging an insurance company for medical equipment that cost more than the supplies themselves. Giving the patient a manual wheelchair while paying for a motorized wheelchair is one example.

• Upcoding of services: billing the insurance company for additional services that exceed the cost of the actual therapy. A 45 min session may be invoiced as a 60 min session, for example.

• Charging insurance companies for services that were not performed. Example: signing checks with someone else's signature.

• Unnecessary services: Making claims for services that are unrelated to a patient's health. An example would be a patient who made a claim for daily insulin injections while having no symptoms of diabetes.

• Duplicate claims: Submitting bills that are almost similar but differ in small

details, such as the date, in order to charge the insurance company twice for the same service. For example, to receive twice the benefit as the first claim, a portion, such as a date, is changed when the claim is submitted a second time.

• Abuse of the codes: Insure explains the policy with a code, but the claimant doesn't know about the code, which can be changed later to give fewer services (Rawte & Anuradha, 2015: 1-5).

C. Purpose of Study

"Machine learning based for insurance claims fraud detection technique for safeguarding the health industry" critical to current and future research scenario in health insurance industry. Insurance fraud has become a common issue, causing huge financial losses to insurers, and negative impact on health system efficiency. Using machine learning techniques to identify fraudulent activities in insurance claims, this study aims to address the following important areas: minimizing financial loss, health care a enhancement, and fraud prevention and deterrence.

By looking at the main machine learning traits that each method used to identify fraudulent claims in the research in order to better understand the differences in the fraud patterns picked up by the machine learning techniques. by utilizing SHAP (Lundberg & Lee, 2017), an explainable AI tool, to make the mechanics of the processes employed by these methods transparent that is, the much-discussed "black box" of machine learning. The degree to which a particular characteristic influences a particular prediction made by a machine learning model is shown by SHAP values.

D. Research Objective

This study aims to create and apply machine learning methods for the detection of healthcare insurance claims fraud. By leveraging advanced analytical techniques and using SHAP to understand the ML models in fraud detection processes, thereby mitigating financial losses and preserving the integrity of insurance systems. the contributions of this study are listed below:

- Use ML and XAI techniques in the field of insurance claims fraud detection to get background information and understand how the model works and why

the model is making certain predictions.

- Implement SHAP to the best ML method to help detect fraud.
- Use SHAP plots in insurance claims fraud detection which is advantageous as they provide a clear understanding of the factors influencing a model's predictions. By visualizing the impact of each feature on individual predictions.
- Use Ensemble Learning: combines multiple individual models to create a more powerful model that can make better predictions by using hard and soft voting.
- Use grid search cv to get the best hyperparameters to each ML model to attain optimal performance.

The study explores the fraud patterns detected by machine learning techniques by examining key characteristics used. Using the explainable artificial intelligence tool SHAP, the study reveals the mechanics of these approaches, highlighting the impact of a feature on a machine learning model's prediction.

E. Thesis Outline and Workflow

The outline of thesis is shown below:

Chapter 2: Literature Review: has a critical analysis of existing literature on the thesis topic. It involves summarizing, synthesizing, and evaluating the current state of knowledge.

Chapter 3: Methodology: includes overview of the proposed approach and contains Data Collection, EDA and data visualization, Data preprocessing, ML models and feature engineering, feature selection, deep learning architectures, PCA, and Ensample learning.

Chapter 4: Experimental Results and Discussion: performance of machine learning models, performance of deep learning model and comparison and analysis of results.

Chapter 5: Explainable AI: define SHAP and its plots.

Chapter 6: Deployment: create user interface using flask.

Chapter 7: Conclusion and Future Work: summary of the research and Future Directions for Research.

References: the cite all sources are used.

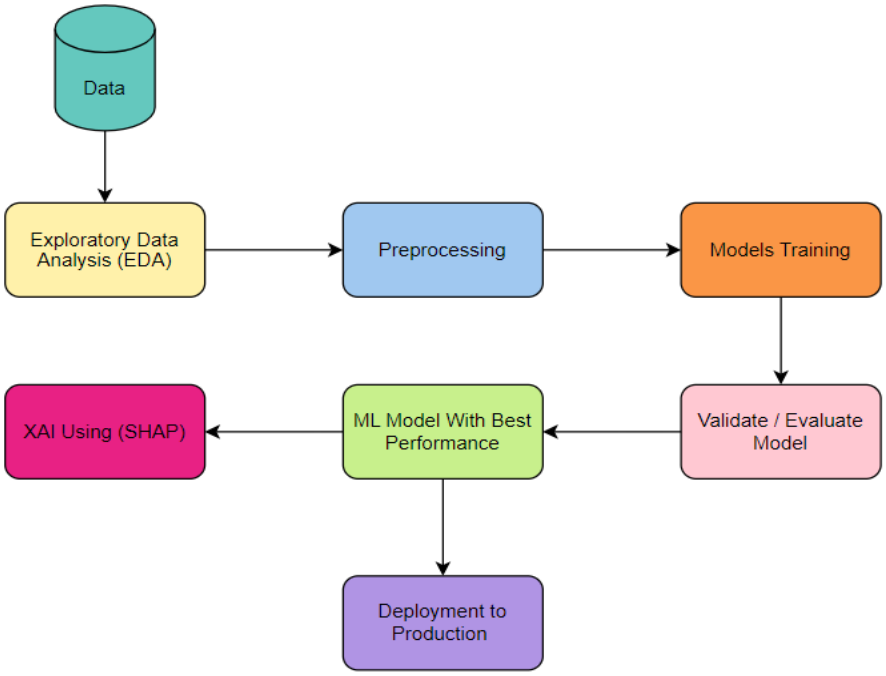

Figure 1 Workflow Diagram

The workflow in Figure 1 for insurance claims fraud detection encompasses a systematic process from data ingestion to model deployment, incorporating various stages such as data collection, exploratory data analysis (EDA), preprocessing, model training, validation, explanation using SHAP, and deployment with a user interface.

II. LITERATURE REVIEW

A. Insurance Claims Fraud Detection

In (Bauder, et al., 2016: 784-790), Bauder, et al. worked on a machine learning algorithm to distinguish when physicians submit medical insurance claims with suspicious behavior. Their new study may be able to provide some light on whether and when physicians deviate from the norm in their field, it could be a sign of misuse, ignorance, or fraud, or of billing practices. The dataset they use is provided by the United States Medicare system. Because of the size of the dataset, they only sampled physicians who only practiced in one state. The model is assessed using a 5-fold cross-validation calculation of F1-score, recall, and precision for the multinomial Naive Bayes model. With an F-1score of over 0.9, the model could accurately predict many types of physicians. The results show that it is possible to classify physicians into their various fields using nothing but the treatments they charge for and machine learning in a unique way. Their study offers a methodology that might help identify physicians who might be abusing insurance programs so that additional research can be done on them.

An innovative method is suggested by Lavanya Settipalli and G.R. Gangadharan in (Settipalli & Gangadharan, 2023) to use unsupervised multivariate analysis for provider-submitted healthcare claims. To track the actions of providers, they suggested a model analyzes continuous data and multivariate categorical data in two steps. In the first step, Weighted Multi Tree (WMT) is designed to compare provider profiles using categorical data, relationships between rendered services and profiles, to spot fake services. The Weighted multi-tree is an unsupervised method that doesn't need labeled data or the participation of medical experts. The second phase uses multiple Density-Based Clustering (DBC) methods to create a univariate fraud detection model that identifies fake claims based on continuous data of claims, including service counts and service charges. By performing experiments on the CMS part B program's claims from different medical specialties or provider types, their projected WMTDBC's performance is evaluated. Their empirical findings

demonstrate that, in comparison to state-of-the-art models, the WMTDBC technique improves detection performance.

The study in (Peng, et al., 2006: 116-120) was conducted by Yi Peng et al. to comprehend and identify possible healthcare frauds from massive databases the dataset that was used is given by a US insurance carrier and includes health claims data. using the clustering technique. They specifically employ two clustering techniques, CLUTO, EM and SAS, to a sizable real-world health insurance dataset and contrast how well these techniques perform. CLUTO is software designed by the University of Minnesota's for clustering datasets and examining the traits of the various clusters. SAS EM A selection of DM functions is supported by the SAS EM commercial software package. Many businesses employ SAS EM. They show that SAS EM offers more relevant clusters than CLUTO while SAS EM is faster. Clustering often has two uses. to comprehend and cluster insurance claim data, the project uses clustering as a stand-alone method. After this investigation is over, they will have some records of labeled insurance claims. They can use these labeled data to apply further algorithms. In the future, they advise utilizing classification algorithms because the ultimate objective is to forecast insurance claims with a trustworthy level of precision.

By using 2 criteria: anomalies based on illness and period-based claims the trial in (Verma, et al., 2017: 1-7) by Aayushi Verma, Anu Taneja, and Anuja Arora aims to identify fraudulent trends in the data. Results from rule-based mining are analyzed using both criteria. Period-based claim anomalies outliers are detected using statistical decision rules and k-means clustering, and disease-based anomalies outliers are discovered using Gaussian distribution-based mining using association rules. These anomalies represent insurance claims for fraud in the data. The suggested method has been tested on a real-world dataset from a health insurance company. Association rule mining is used to find reliable fraud detection rules and often occurring fraud patterns for various organizations. Using association rules mining to separate data, K-means clustering is used to improve performance and decrease time complexity. Outlier detection: To reveal insurance claim fraud, outliers are found in clustered data. Finding observations that differ so much from other observations that it raises the possibility that they were produced by a distinct cause requires the use of outlier detection techniques. The contributions in (Verma, et al.,

2017: 1-7) are a proposal for a system for detecting insurance fraud in claim data, identification of frequently occurring fraud patterns, and fraud detection. Their research focuses to maximize the value of medical claim coverage while also supporting the expenditure and payment for treatments that are required. They demonstrate the effectiveness of their suggested strategy in extracting fraudulent claims from the available data utilizing data mining methods.

A unique hybrid technique, developed by Vipula Rawte and G. Anuradha, is used in (Rawte & Anuradha, 2015: 1-5) to identify fraudulent claims in the health insurance sector. To identifying health insurance fraud and flagging it for additional investigation, the proposed approach uses a hybrid model. Because of the dynamic nature of the data and the constant generation of new data, they selected the Support Vector Machine (SVM) for classification, and Evolving Clustering Method (ECM) for clustering. This method groups insurance claims by kind of ailment first, after which they are categorized to look for any duplicate claims. The steps of their approach are: The doctor invoices the patient for the equipment and services served to them throughout their treatment, The patient submits insurance claims. To identify fraudulent claims, claims are presented to the hybrid framework, which uses classification (SVM) and clustering (ECM). The false claims are flagged by an expert for further examination by the insurance provider. The insurance provider is then notified of the valid claims, and the patients are reimbursed.

To maintain track of a customer's medical history and the insurance information that has been approved, healthcare insurance firms still rely on an antiquated system. This opens the door for several health insurance frauds to take place. Blockchain technology aids in data security, integrity, and fraud detection. is used by G. Saldamli et al in their search (Saldamli, et al., 2020: 145-152). They used the following technologies to create the frontend, backend, and database for their prototype: Flask and BigchainDB for backend programming, React.js, and Redux for front-end application development. REST API calls for client-server communication and an EDI Validator to verify the accuracy of insurance records before uploading them to the blockchain. Placing the program on a blockchain makes it constantly accessible to users. In the digital world, security is the primary concern. They are making sure that the system's security is not jeopardized. Theft of personal information is one of the most pressing modern concerns. Neo4j is a graph database

used for analytics, graph-based search, and fraud detection. Neo4j. The research shows that, rather than using a centralized approach, the usage of blockchain to store health information may be effectively secured by distributing data among several machines that are overseen and approved by a distributed community.

Using a novel multistage methodology, the paper (Johnson & Nagarur, 2016: 249-260) by Marina Evrim Johnson and Nagen Nagarur suggests a new way for insurance companies to find provider and patient fraud. The six stages of their methodology are as follows: To identify how a provider differs from their peers in terms of practice, the first stage of provider profiling groups providers based on resource utilization variables. The likelihood that a variable on a claim form does not belong to a group of population parameters is computed in the second step of demographic screening. The likelihood values that the claim amounts are inflated are determined in the third stage of claim amount screening. The risk values of claims being fraudulent are then calculated at stage four of the fraud risk quantification process utilizing distance, density, and likelihood values. The claim amount, error rates, and avoidable cost are used to calculate the risk threshold value for each claim in stage five of the fraud detection process. After that, claims are evaluated for fraud by comparing them to the risk threshold values in the sixth stage. The first three steps are designed to find anomalies in claim amounts, services, and provider relationships. The data gathered in the first three phases are then combined in stage four to provide an overall risk measure. The risk threshold values are then computed in step five using a decision tree-based algorithm. The risk threshold value from stage 5 and the risk value obtained in stage four are compared to determine whether the claim is false. Real-world insurance data show good performance from the research technique. They used the methodology on four distinct specialties, and they found that it worked as well for each of them, with 86% accuracy. Additionally, they contrasted the approach with semi-supervised and unstructured neural network approaches. The outcome showed that, in terms of accuracy rates, their multi-stage algorithm performed better than neural network approaches.

Using actual health insurance data, Lu, Fletcher, and J. Efrim Boritz in (Lu & Boritz, 2005: 633-640) apply the Benford's Law approach to the identification of abuse and fraud in claim of health insurance. and by employing a digital analysis strategy that takes an unsupervised learning way to manage missing data. They show

increased accuracy in detecting aberrant data suggestive of fraud over the conventional Benford approach and highlight some of the difficulties in the investigation of healthcare claims fraud. They evaluate their new approach using real health insurance claims data given by Manulife Financial, and they show better precision for identifying potentially fraudulent insurance claims.

Dallas Thornton et al. organize their framework for design science contributions and tackle a pertinent issue in healthcare fraud detection. The work (Thornton, et al., 2014: 684-694) provides an artifact, a description of how to apply outlier detection to healthcare fraud, and an assessment of this model's performance when applied to a state-wide collection of actual healthcare claims from more than 500 providers. The model is assessed by using it in practice with real healthcare data and having professionals review the analysis' findings. The work adds to the body of knowledge by laying out a plan for future uses of outlier detection in healthcare and perhaps other related fields. They used the Medicaid domain context to discuss considerations for applying it to various data contexts. They explained the model to the relevant parties, explaining how to really use the general procedure and the various scoring techniques. They gained a great deal of knowledge on antifraud initiatives from their research. It takes considerable effort to design analysis tools and understand the results of those approaches. subject matter expertise in healthcare. A successful outcome can be defined as the identification out of 360 main dentists, 17 offer extra inquiry, of whom 12 out of 17 have been assessed and judged appropriate for official investigation.

Overestimated claims, post-dated policies, and other types of client-side insurance fraud are all common. However, insurance vendor fraud is sometimes seen in the form of policies from fictitious companies, failure to pay premiums, and other things. Various classification algorithms, including Multi-Layer Perceptron (MLP), Random-Forest (RF), Naive Bayes (NB), K-Nearest Neighbor (KNN), Support Vector Machine (SVM), Linear Regression (LR), Adaboost, and Decision-Tree (DT). are compared in paper (Rukhsar, et al., 2022: 33-40) by Laiqa Rukhsar et al. to identify insurance fraud. F1-Score, Recall, and Precision, performance indicators are used to evaluate how well the algorithms work. The results of the classification algorithms show that, when compared to the other strategies, DT provides the highest accuracy (79%). Additionally, Adaboost displays an accuracy of 78%, which is

nearer to DT.

In order to eliminate the interference caused by camouflage in fraud detection, C. Sun et al. suggest in (Sun, et al., 2018: 14162-14170) the PCDHIFD which refers to healthcare insurance fraudster detection via patient cluster divergence. The following is a list of the paper's main important points in their work: 1) To assess the density of each item, improve the density peak clustering DPC and use the betweenness centrality of the nodes in the graph, which can avoid the difficult cutoff distance decision. 2) extract each cluster's semantic interpretation. They locate a dense core area around the density peak for each cluster and then derive a semantic interpretation for each cluster. 3) Patients' long-term health-seeking activities are considered to lessen the effect of camouflage because camouflage can only be used temporarily. and Three steps make up their work: 1) Calculate the degree of similarity between graph G and patient-level hospital admission representing patient hospital admission. 2) Using a graph-based dense peak clustering technique such as GDPC in G, cluster each cluster and extract its semantic interpretation. 3) Determine each patient's likelihood of fraud by computing the Patient Cluster Divergence for the full period. According to experimental findings, their strategy can considerably increase the accuracy of fraud detection in the presence of camouflage. To be more precise, their PCDHIFD performs over 15% better than the comparable.

The fraud detection problem is formulated by Haque and Tozal in (Haque & Tozal, 2021: 2356-2367) using limited, conclusive claim data made up of medical diagnosis and procedure codes. They address the issue of fraudulent claim detection by employing a novel representation learning algorithm that turns diagnosis and procedure codes into Mixtures of Clinical Codes (MCC). Using robust Principal Component Analysis and LSTM networks. They also look towards MCC extensions. They presumptively believe that each claim represents latent or obvious combinations of clinical concepts, which are combinations of codes for procedures and diagnoses. They expand the MCC model by utilizing Robust Principal Component Analysis (RPCA + MCC) and (LSTM + MCC) to separate out important concepts from claims and categorize them to non-fraudulent or fraudulent. Their findings show an expansion in the ability to quickly identify fraudulent medical claims. When generating non positive claims, MCC+ RPCA and MCC behave

consistently across a range of idea sizes and replacement probabilities. On the inpatient dataset, MCC + LSTM achieves 50% recall, 61% precision, and 59% accuracy. Additionally, it displays 72% recall, 83% precision and 78% accuracy. Since both MCC+ RPCA and MCC use an SVM, they find similarities in the findings. They are certain that the proposed issue formulation, showing learning, and solution will launch fresh investigation into the identification of bogus insurance claims utilizing sparse but reliable data.

In (Johnson, et al., 2019: 1-35) study they use CMS Medicare data and the label the data using LEIE data. They assess the effectiveness of 6 deep learning algorithms for resolving class imbalance. They also consider a variety of distributions of the class and investigate the connection between the ideal decision threshold and the minority class size. On the supplied LEIE / CMS datasets, they attain the greatest ROC AUC scores to yet by combining deep learning with techniques for resolving class imbalance. With average ROC AUC scores of 0.8509 and 0.8505, ROS-RUS or ROS beats all algorithm-level techniques and baseline models in removing the imbalance between classes in the training set. They conclude that because deep learning with ROS-RUS has training times four times faster than baseline models, it is the most effective method for identifying fraud in the CMS Medicare data sets. Using a class distribution of 99:1, RUS performs significantly better than baseline techniques and algorithm-level, but performance suffers as imbalance levels are further reduced. Based on ROC AUC scores, algorithm-level approaches, and baseline methods both perform statistically similarly. However, study of G-Mean scores and decision threshold intervals indicates that algorithm-level methods produce compared to baseline models, more stable decision borders. It is usually necessary to modify classification decision criteria using a validation set when training neural networks with unbalanced input, according to an observation that shows a good judgment threshold and the size of the minority class have a strong linear relationship.

In (Dhieb, et al., 2020: 58546-58558), by Najmeddine Dhieb et al. create an automated and safe insurance system architecture that minimizes warns, secures insurance operations, human involvement, and educates about high-risk clients, uncovers lowers financial loss and false claims for the insurance industry. Following the presentation of the blockchain infrastructure to provide safe data sharing and

transactions across several agents interacting inside the insurance network, the XGBoost ML method is recommended for usage with the insurance services. They also compare the outcome to that of other cutting-edge algorithms. The acquired results show when tested on an automobile insurance dataset, xgboost beats other existing learning algorithms. When identifying fraudulent claims, for example, it achieves 7% greater accuracy compared to decision tree models. Additionally, they suggest an online learning method to automatically handle changes from the insurance network in real-time, and they demonstrate that it performs better than another cutting-edge online algorithm. Lastly, they emulate artificial intelligence and blockchain-based architecture by combining the created machine learning modules with the hyper ledger fabric composer.

An actual analysis comparing various deep and machine learning models on various datasets for the detection of fraudulent transactions is provided in the paper of Pradheepan Raghavan and Neamat El Gayar in (Raghavan & El Gayar, 2019: 334-339) their study's goal is to identify which methodologies are best suited to different sorts of datasets. Their study may aid practitioners and businesses. Since many organizations currently invest in new technologies, it's important to understand how different approaches work with different types of datasets to advance their operations. In their research, SVMs are the most effective approaches for detecting fraud with big datasets, and CNNs may be used for even better results. and SVM, RF, and KNN ensemble methods can offer good enhancement for smaller datasets. In most circumstances, Convolutional Neural Networks (CNN) outperform other deep learning approaches such as DBN, RBM, and autoencoders.

The first study on utilizing LightGBM and CatBoost to encode categorical data in order to detect Medicare fraud was published in (Hancock & Khoshgoftaar, 2021: 268) by Taghi M. Khoshgoftaar and John T. Hancock. They demonstrate that CatBoost outperforms competing algorithms having a mean AUC value of 0.77452 in the duties of identifying Medicare fraud. Their analysis reveals that this outcome is noticeably superior to LightGBM's mean AUC value of 0.76132 at a 99% confidence level (with a p value of 0). They further demonstrate that CatBoost produces a mean AUC value of 0.88 when a further category variable is added (Healthcare provider state), which is much more than the average AUC value of 0.85137 obtained by LightGBM. Their experiential data shows categorical features

that make CatBoost a superior classifier for detecting Medicare fraud compared to other classifiers.

Several machine learning techniques are compared in (Bauder & Khoshgoftaar, 2017: 858-865) by Taghi M. Khoshgoftaar and Richard A. Bauder to identify Medicare fraud. They conducted a comparison study with guided techniques to machine learning that are unsupervised and hybrid utilizing class imbalance and four performance metrics reduction by 80-20 under sampling and oversampling. They compile the Medicare data from 2015 Using fraud labels from the List of Excluded Persons Entities database for various provider types. Their findings indicate that the effective identification of dishonest providers is feasible, and the 80 / 20 sampling approach showing the most efficient performance among the students. Additionally, supervised techniques delivered better results. compared to unsupervised or mixed techniques, although these outcomes vary according to the class imbalance sampling method and type of provider.

In (Gupta, et al., 2021: 96-102), Rohan Yashraj Gupta et al. conducted a study of several deep and machine learning models for developing in the healthcare system a fraud detection model. To address the challenges of data imbalance and classification model selection, they used six classification models and three different data imbalance techniques, including SMOTE, ADASYN, and TGANs. They have also employed six different neural network model variations. They did this by using information from Ayushman Bharat (PM-JAY India), the biggest universal health care program in the world. In this investigation, a total of 26 models were put to the test. These models' effectiveness was evaluated using a variety of measures, including F1-score, specificity, accuracy, and sensitivity. In their study, they found that a NN model trained on sparse data outperformed other models. This model produced the highest F1-score of 0.95.

T. M. Khoshgoftaar and M. Johnson investigate the use of healthcare provider summary data for fraud detection in (Johnson & Khoshgoftaar, 2022: 236-242). They use the most recent CMS Part. Utilizing large datasets to curate two new collections of labeled data for supervised learning. The new two data values are contrasted with a well-known baseline data set from associated works with two well-liked ensemble learners and six cross validation runs, various complementing performance statistics tests and metrics. Results of classification indicate that the aspects of the suggested

provider summary are reliable indications. of medical fraud. a test for two-way analysis of variance and the new features' 95% confidence intervals gives dramatically improved fraud detection performance when utilized to improve current data sets.

A detailed study using machine learning techniques to identify dishonest Medicare providers is provided in (Smita, et al., 2023) by K. Smita et al. To create and evaluate three distinct learners, they leverage provider exclusions for fraud labeling and publicly available Medicare data. Given that there are so few genuine fraud labels, they use logistic regression to create two class distributions to decrease the impact of class imbalance. The results demonstrate that, in comparison to Logistic Regression, other methods perform poorly. The performance of learners in detecting fraud is the best, especially for the 80/20 class distributions with low false negative rates and average AUC scores. They effectively show how machine learning models can be used to identify Medicare fraud. They also use SVM, which have been deemed superior to other classification techniques for several reasons. It has a nonlinear dividing hyper plane, which triumphs over the discrimination inside the dataset, among other important advantages.

Herland, M. et al. (Herland, et al., 2018: 1-21) focuses on Medicare fraud detection using Part B, Part D, and DMEPOS CMS databases. They additionally combine the three main datasets to generate a fourth dataset. They talk about how each of the four dataset’s data was processed the List of Excluded Persons and Entities (LEIE) maintained by the Inspector General's Office is used to map real-world provider fraud labels. Three learners are constructed and evaluated for each dataset as part of their exploratory investigation on Medicare fraud detection. According to the results, the the best overall score 0.816 was obtained by combining the dataset with the LR . based (ROC) Curve efficiency metric. The Part B dataset, with LR of 0.80, came close. All learners combined, the Part B and Combined datasets yielded a good fraud detection result overall, between these datasets, there isn't a statistically significant difference. Thus, in order to identify criminal activity when a physician is receiving payments through all or any of the Medicare components they examined for their analysis, they advise using the Combined dataset, based on their results and the supposition that it is hard to determine which particular Medicare component a physician will defraud. within.

The goal of the (Nabrawi & Alanazi, 2023: 160) study by Nabrawi, E. and Alanazi, A. is to create a health model for Saudi Arabia that can automatically identify fraud from health insurance claims. With maximum accuracy, the model identifies the main cause of fraud. 3 supervised deep and machine learning techniques were applied on the labeled imbalanced dataset. Three Saudi Arabian healthcare organizations provided the dataset. Artificial neural networks, random forest and logistic regression, were the models used. The dataset was balanced using the SMOT method. To omit unimportant features, Boruta object feature selection was used. AUC, F1-score, recall, precision, specificity, and accuracy were used as validation metrics. According to the random forest classifier, system type, education, and age were available, with accuracy of 98.21%. Through logistic regression, 80.36% accuracy, were obtained ANN 94.64%. Three effective models were used in this predictive analytics investigation, and each of them produced respectable accuracy and validation metrics.

B. Explainable AI

Hancock, John T., et al. in (Hancock, et al., 2023:154) have demonstrated how, without necessarily sacrificing classification performance, a new feature selection strategy may be used to produce more explainable models and drastically reduce the amount of data of an imbalanced Data dataset. To construct models that produce satisfactory performance and determine the lowest number of characteristics that can be removed from a dataset, statistical analysis is necessary. In their work, two Big Data datasets features are ranked using an ensemble supervised feature selection technique. The datasets are labeled with information from the LEIE and obtained from Medicare Part B and Medicare Part D insurance claims data. They meet the definition of big data and exhibit extreme imbalance. As a result, their findings show that their feature selection method is a practical method that can be used to other highly unbalanced Big Data sets to reduce their dimensionality for supervised machine learning applications. Additionally, they employ an additional outcome of their feature selection process is explainable models. One can argue that the decreased number of features provides all the information needed for the algorithm to effectively complete the machine learning task when it can be demonstrated that models constructed with a lower number of features produce

results equivalent to employing all features. The type of information that a model employs is simpler to comprehend and articulate when fewer features are needed.

In (Dangers, 2022) by Lennart Dangers delves into the fusion of unsupervised methodologies with a singular supervised approach, unveiling promising applications in anomaly detection within healthcare encounters and insurance trends. Notably, an anomaly detection model demonstrates robust performance, poised for real-world deployment, while predictive analytics offer insights into potential insurance cancellations. However, their study underscores the imperative of defining additional data parameters and business strategies to enhance efficacy further. Within the healthcare fraud domain, the utilization of unsupervised learning techniques, particularly the Isolation Forest algorithm, showcases commendable precision and recall rates, facilitating the identification of novel fraudulent patterns. Moreover, a visual method illuminates suspicious provider networks, augmenting the fraud detection arsenal. Nevertheless, challenges persist, including the computational demands of encoding high-cardinality features and the scarcity of labeled data for rigorous model evaluation. Proposals for future research advocate for the exploration of alternative anomaly detection models and the integration of supervised learning pipelines, alongside a call for enhanced data infrastructure and cloud-based solutions to navigate the intricacies of healthcare data confidentiality.

Yingchao Ji's goal in (Ji, 2021) was to research and assess XAI techniques for credit card fraud detection. Because real credit card transaction datasets are confidential, A fictitious dataset was employed to train the model. The amount of memory and processing power that was available resulted in the dataset being trimmed. DNN and RF were the selected machine learning techniques because of their strong results in earlier studies. The imbalance issue with the selected dataset was resolved by applying the oversampling technique. His study employed the measures of recall, accuracy, precision, and F1-core. The results show that when it comes to credit card fraud detection, the DNN model performs marginally better than the RF model. While the RF accuracy was 96.42\%, the DNN accuracy was 96.84\%.They assessed the explainability of their predictions using two different types of explanations (LIME and SHAP) and selected DNN because of its superior performance. Lastly, a quantitative survey was used to assess the XAI outcomes. The survey's findings showed that the XAI explanations can raise users' perceptions

of the system's reasoning skills somewhat, while LIME outperformed SHAP somewhat in terms of explainability. To provide users with thorough explanations, more research into visualizing data mining and the training data is advised.

Using exclusive insurance claim data, a study (Debener, et al., 2023: 743-768) by Jörn Debener et al. assesses supervised and unsupervised learning. Additionally, in collaboration with an insurance provider, they carry out a field test to examine how well each strategy detects fresh fraudulent claims. They arrive at several significant conclusions. Insurance fraud can be effectively detected by unsupervised learning, particularly in isolation forests. Despite the small number of labeled fraud incidents, supervised learning likewise works well. It's interesting to note that, depending on the input data, both supervised and unsupervised learning can identify new false claims. Consequently, they advise seeing supervised and unsupervised techniques as complements rather than alternatives when it comes to implementation. All things considered, the SHAP analysis verifies that distinct features are prioritized by supervised and unsupervised learning techniques in the identification of claim fraud. This is consistent with the results of the field trial, which showed that the two methods served as complementing fraud detection techniques rather than identifying the same bogus claims.

This research proposes an original method for improving the effectiveness of fraud detection in insurance claims using SHAP, a model explainability technique. The aim is to use SHAP for this purpose, making a significant contribution to the field. By using SHAP, it provides a new way to understand how complex machine learning models detect fraud, shedding light on their decision-making process. Furthermore, using SHAP plots are useful for identifying insurance claim fraud because they offer an understanding of the factors affecting a model's predictions. through displaying how each feature affects distinct predictions.

Using in-depth analysis and clear visualizations, show how SHAP helps us understand which features are most important in predicting fraud. This makes the model's results easier to understand, allowing insurance professionals to make better decisions about evaluating claims and preventing fraud. This research not only demonstrates how SHAP can be used practically in this important area but also highlights its potential to change how detect insurance fraud by bringing clarity, responsibility, and confidence to predictive models.

III. METHODOLOGY

A. Data Collection

The Medicare dataset is from Kaggle, and it consists of 4 CSV files for training and testing that consider each provider's beneficiary information, outpatient claims, and inpatient claims the files are:

- **Inpatient data**

This data reveals information regarding the claims made for people who are admitted to hospitals. Additionally, it includes information about their admission and discharge dates as well as their admittance diagnostic code.

- **Outpatient data**

A data gives specifics on the claims made for patients who visit hospitals but are not admitted.

- **Beneficiary details data**

A data includes beneficiary KYC information, such as health problems and region of residence.

- **Target data(train/test)**

Provider: Each healthcare provider has a different identification number.

Target: The Yes and No values in this column indicate whether the particular provider has been recognized as potentially fraudulent.

The columns of data on inpatients and outpatients:

ClaimID: Each submitted claim has a specific identification number.

Bene_ID: Identifier for each beneficiary registered for an insurance program,

Attending Physician: Attending The physicians who treated the patient are listed by their IDs.

OperatingPhysician: column lists the id of the physicians who performed the

patient's surgery.

ClmDiagnosisCode: includes Diagnosis Codes for procedures that clinicians execute on patients.

ClmProcedureCode: Lists the codes for the procedures that patients go through.

Provider: The healthcare professionals' individual ID.

InscClaimAmtReimbursed: The overall sum that was reimbursed to the claimant following settlement.

ClaimStartDt / ClaimEndDt: These columns include the dates when the claims were initially filed and when they were ultimately resolved.

AdmissionDt / DischargeDt: The date the patient was admitted to the hospital and the date the patient was released.

The columns in Beneficiary Data:

BeneID: It is the beneficiary's ID.

DOD: death date of beneficiary.

DOB: birth date of beneficiary.

Country, State, Race, and Gender: It includes the beneficiary's country, state, race, and gender.

ChronicCond: when the columns that begin with "ChronicCond_" show whether the patient currently has that specific disease. It also shows the patient's risk score.

RenalDiseaseIndicator: Indicates whether the patient is currently suffering from kidney disease.

IPAnnualDeductibleAmt: This is the annual hospitalization premium that the patient must pay.

IPAnnualReimbursementAmt: This variable contains the annual maximum reimbursement for hospitalization.

OPAnnualDeductibleAmt: This amount is the annual outpatient visit premium that the patient must pay.

OPAnnualReimbursementAmt: This number represents the annual maximum reimbursement for outpatient appointments.

All these columns have impact in fraudulent ratio.

B. EDA and Data Visualization

Data analysts utilize EDA, which typically employs data visualization tools, to investigate, understand, and summarize the essential elements of data sets. It helps data analysts detect patterns, spot anomalies, test hypotheses, and check assumptions by advising them on how to change data sources to get the info they require. (ibm.com, 2024). Conducting a proper EDA, with the help of visualization in s histograms, distplot, box plots, scatter plots, heatmaps, etc., can reveal significant relationships among variables, identify potentially insightful factors mixed, and gain a deeper understanding of the characteristics of an embedded dataset.

A histogram is a graphical depiction of a dataset's distribution. This type of bar chart shows the frequencies of the data inside various bins or intervals. The range of values (split into bins) is shown by the x-axis, while the frequency or count of observations falling inside each bin is represented by the y-axis. Histograms are very helpful in deciphering the distribution's form, seeing trends, and finding outliers in a dataset.

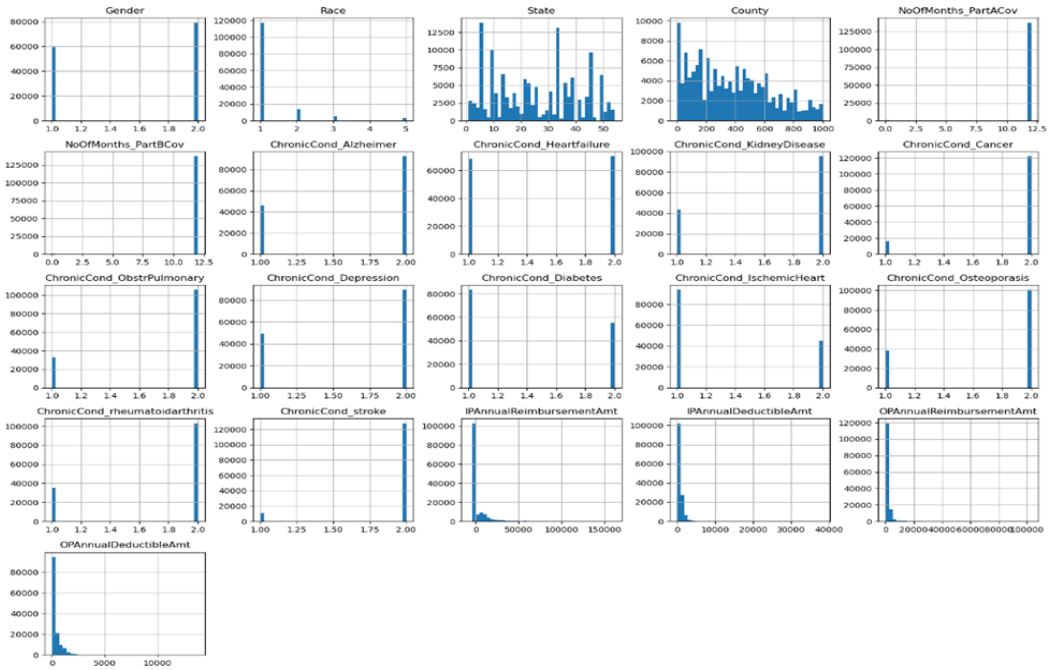

Figure 2 Histogram Graphs of Beneficiary Data

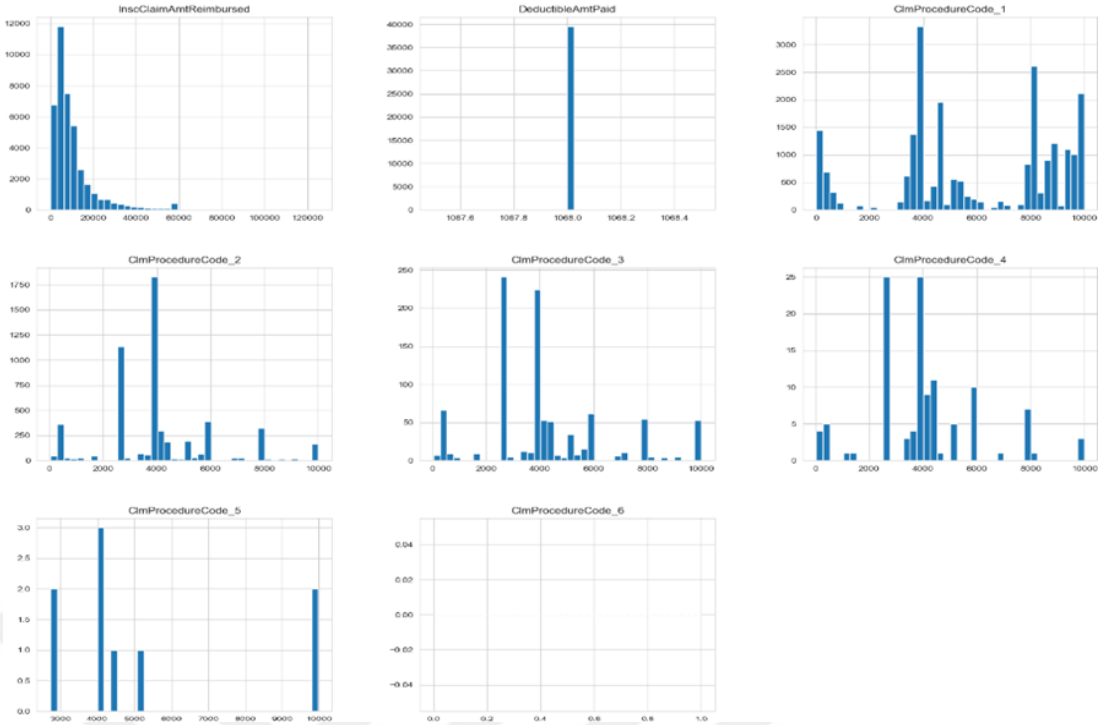

Figure 3 Histogram Graphs of Inpatient Data

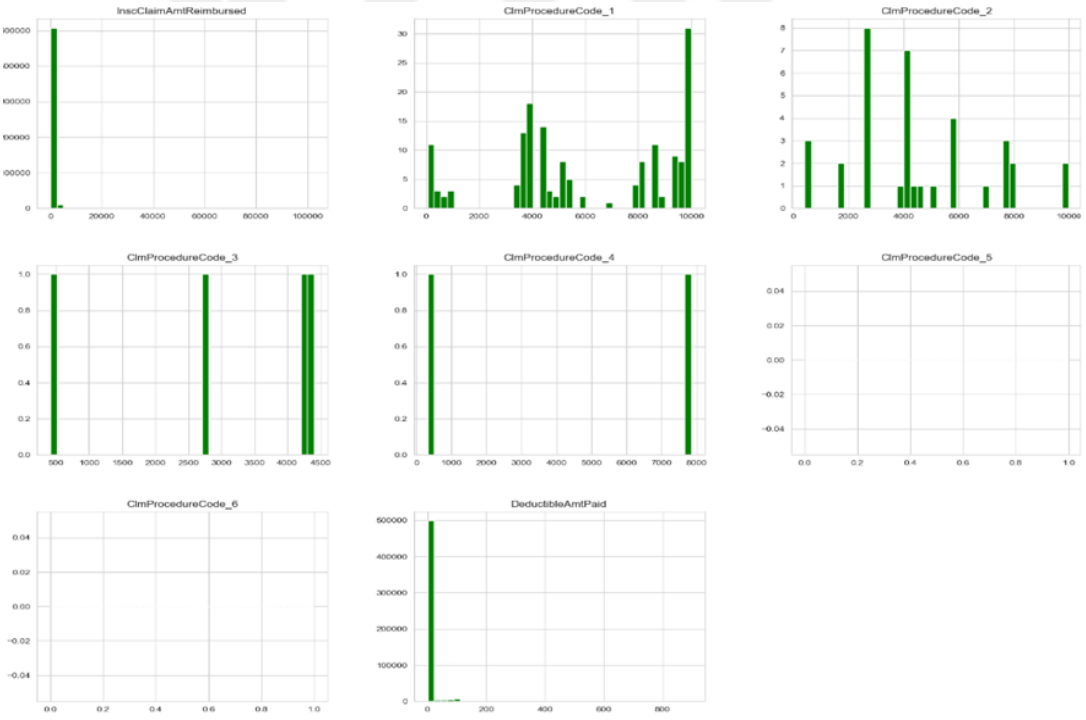

Figure 4 Histogram Graphs of Outpatient Data

Figures 2,3, and 4 show the histogram of data which show frequency distributions of attributes.

1. EDA Finding

The exploratory data analysis (EDA) revealed several findings, including

class imbalance, which refers to an unbalanced distribution of cases. To maintain predictive accuracy for both classes, the model did not favor the majority class. The imbalance between non-fraud and fraud cases in the target variable, with over 90% being non-fraud providers and less than 10% being fraud providers, caused detection issues. This skewed distribution may cause the model to outperform the prediction of the majority group, resulting in poor fraud detection performance. To overcome this challenge, synthetic minority oversampling technique (SMOTE) was used. SMOTE works by artificially generating a small population of samples, thereby balancing the class distribution and reducing the effect of class imbalances on the model performance by representing fraud cases in training data developed, SMOTE helped to improve the model's ability to accurately detect fraud.

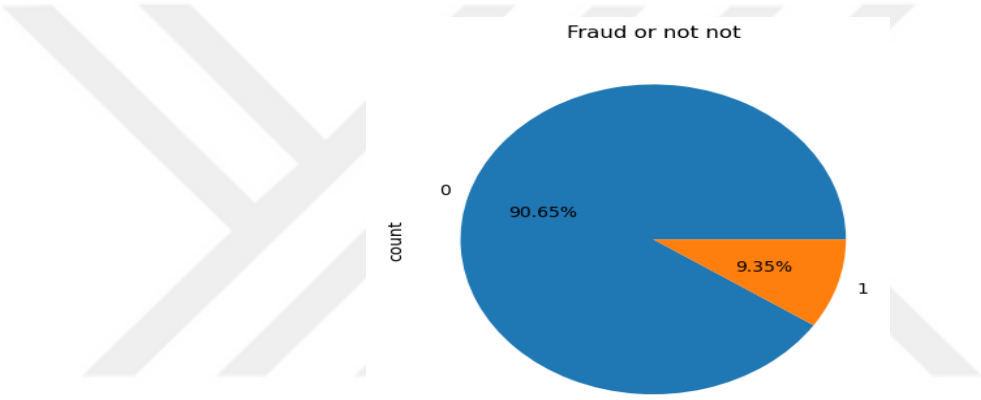

Figure 5 Target Variable (Fraudulent Providers)

Figure 5 shows the target variable Potential Fraud which shows a highly imbalanced ratio between fraud and noun fraud cases.

C. Data Preprocessing

Is a technique in data mining to make data more efficient to use it later in analysis, training, etc. in this research data preprocessing was an important step to make the data useful to work on.

1. Dataset Merging

So, in healthcare datasets there were 4 datasets so by merging them to get final dataset which can work on as shown in figure 6 first merged inpatient and outpatient data by same columns then merge this dataset with beneficiary data by BeneID. Finally, merge this dataset with provider data by provider id to get the final

dataset.

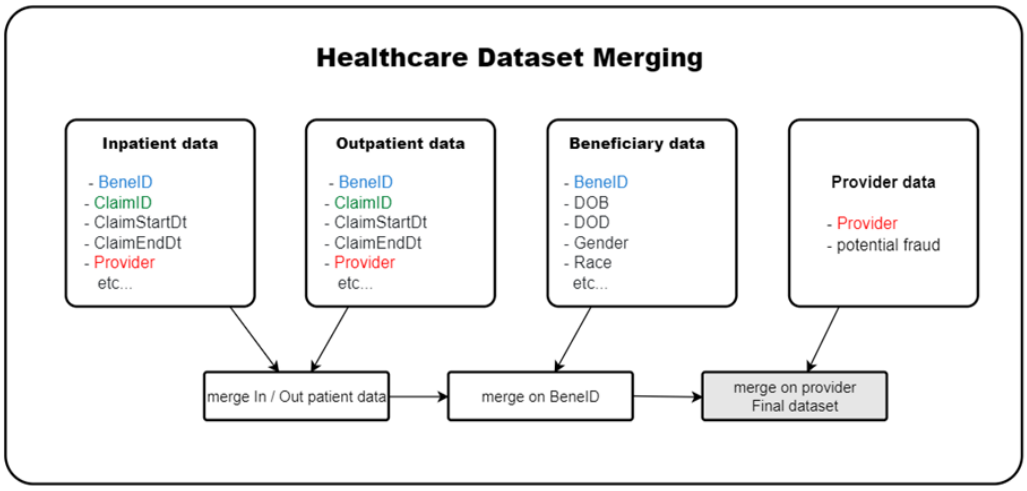

Figure 6 Datasets Merging

2. Convert Categorical Data to Numeric Data

Patterns, trends, and correlations are among the valuable information that categorical data may offer machine learning models. However, because most ML algorithms are made to operate with numbers, they are unable to directly analyze categorical input. For instance, we are unable to compute distances or carry out arithmetic operations on categorical data. Consequently, before feeding categorical input into an ML model, it must be converted to numerical data. One of convert technique is label encoding which converts categorical data to numeric by assigning a unique number to each class starting from 0.

By applying label encoding it converts Potential Fraud target feature from [‘yes’, ‘no’] to [0, 1] which 0 means no which it is not fraud and 1 means yes which is fraud case in the dataset.

3. Handle Missing Value

Some of features in the dataset have missing values so by using fillna (0) it filled them with zeros like:

The physicians’ columns have NAN values, so filled them with zeros.

Missing procedure and diagnostic codes filled them with zeros.

The admission date and period do not apply to outpatient data filled them with zeros.

Beneficiaries that are still alive are not covered by DOD filled them with zeros.

4. Feature Engineering

A critical part in the ML process is feature engineering, which involves transforming or manipulating raw data to produce additional features that improve a model's performance. A machine learning model that receives more relevant and instructive information via effective feature engineering would have a far higher predictive capacity. One of techniques that use it in this research is groupby function that related to feature engineering.

Using pandas groupby, you can apply a function to the categories and group the data according to those categories. It is particularly useful when working with tabular data. Efficient data aggregation is another benefit. With numerous permutations, the Pandas groupby method is incredibly powerful. It simplifies and speeds up dividing the Data frame according to certain parameters (geeksforgeeks.org, 2023).

For the dataset it must make new features using groupby with taking aggregate or mean

Providers, beneficiaries, and physicians are connected to fraudulent activities because. For providers when they complete and submit the claim, so group by them and took the mean of reimbursed, subtracted. It is suspicious if a provider has a high average claim amount or claim duration. For beneficiaries' group by them and took the mean. It is suspicious if a beneficiary's average claim amount is large. For physicians' group by Operating, Attending, and Other Physicians, and took the mean. It is suspicious if physician has high amount.

For procedure and diagnosis codes combine the patients who underwent the same operation and take the average cost by grouping by them.

In addition to physicians and beneficiaries, providers are occasionally linked to diagnoses and procedures. Thus, group by another feature that has a provider ID. After that, count.

Predicting healthcare provider fraud is the goal. To come up with a feature that corresponds to each provider, group by provider then took the sum.

5. Normalization

StandardScaler is a popular preprocessing method in machine learning for standardizing a dataset's characteristics is called StandardScaler. Rescaling the features to have the characteristics of a conventional normal distribution with a $\mu = 0$ and a $\sigma = 1$ is the process of standardization. And this method was applied in the final dataset.

The formula for standardization is:

$$z = \frac{x - \mu}{\sigma}$$

Where:

- z is the standardized value
- x is the original value
- μ is the mean.
- σ is the standard deviation.

6. Synthetic Minority Oversampling Technique (SMOTE)

SMOTE is a machine learning method used to overcome imbalances in class distribution in datasets, especially in classification tasks. When a class has much fewer samples than another, it is said to be imbalanced.

To balance the class distribution, SMOTE creates synthetic instance for the minority class, or the class with less samples. To achieve this, artificial examples are made along the line segments that connect the minority class's existing examples.

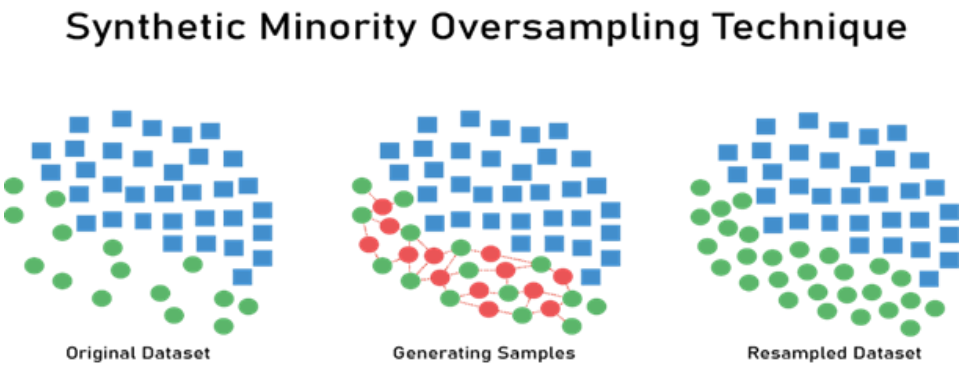

Figure 7 SMOTE Technique

Figure 7 (medium.com, 2023) shows SMOTE Technique

This is an explanation of how SMOTE function works:

- Choosing a Minority Instance: A subset of the closest neighbors of each instance in the minority class are determined.
- Creating Synthetic Instances: The process of creating synthetic instances involves interpolating between the chosen instance and its closest neighbors. To accomplish this, take a random number between 0 and 1, multiply it by the difference between the instance's feature values and those of its neighbors, and add it to the instance's feature values.
- Adding Synthetic instances: By including the synthetic examples in the dataset, the quantity of minority class samples is essentially increased.

When training on unbalanced datasets, the model may become biased toward the majority class. SMOTE helps avoid this from happening. SMOTE seeks to improve model performance on the minority class by producing synthetic examples that better reflect the classes in the training data.

Due to the unbalancing in the dataset for target variable “Potential Fraud” which non fraud cases are 4904 and fraud cases are 506 so by applying SMOTE method both fraud and non-fraud became equal for training set.

D. Machine Learning Models

After preprocessing the data and making it ready for training. The next step is to split the data to train and test which 70% for training models and 30% for testing the performance.

1. Hyperparameter Tuning

Hyperparameter tuning is the process of figuring out which set of hyperparameters is best for an ML model in order to achieve optimal performance. Hyperparameters are model configuration settings that need to be set before training even though they are not learned from the data. To enhance the model's ability to generalize to new, unobserved data, tuning entails determining the optimal values for these hyperparameters. And the technique that is used to find the best hyperparameters in algorithms is grid search.

To have the optimal performance in machine learning models it is better to use hyperparameters tuning. So, by using grid search it shows the best parameters to each model.

GridSearchCV

Grid search can set a list of hyperparameters and a performance metric, the algorithm will try every possible combination to find the best fit. Grid search is a useful technique, but it can be time-consuming and computationally costly, especially when there are many hyperparameters. (aws.amazon.com, 2024).

2. Logistic Regression

A supervised machine learning approach is used for classification problems in which the objective is to estimate the probability that a given instance will belong to a specific class. Logistic regression refers to classification techniques that employ it. Regression is the name given to the process that estimates the probability for a particular class by utilizing a sigmoid function and the output of the linear regression function as input.(geeksforgeeks.org, 2024).

The advantages of Linear Regression: it's simple and Interpretable, efficient and can handle large datasets, Logistic Regression can be regularized to prevent overfitting, provides probabilistic outputs, tends to have low variance, making it less prone to overfitting.

By using grid search it shows that the best parameters for logistic regression are for regularization parameter: c=0.1, and for Regularization type: penalty='l2'

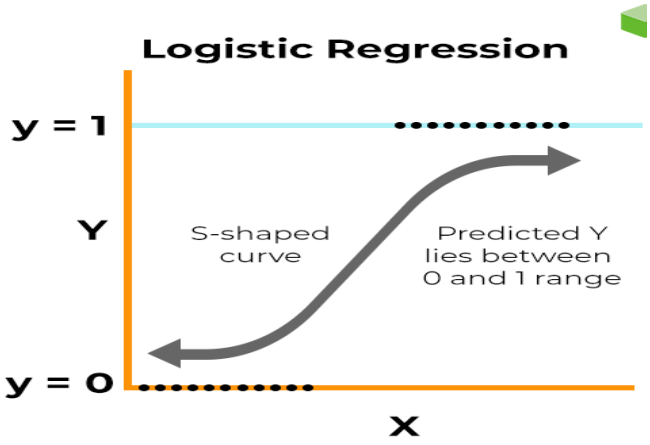

Figure 8 LR Model

Figure 8 (spiceworks.com, 2022) Predictions and their probability are mapped

using logistic regression using a logistic function known as the sigmoid function. A curve with a S shaped form that transforms any actual value in a range of 0 to 1 is known as the sigmoid function.

3. Random Forest

A Supervised learning technique well-liked ensemble learning method used in classification and regression in machine learning, Random Forest constructs several decision trees during training and combines them to produce a forecast that is more reliable and accurate. It is combining several classifiers to solve a complicated problem and to enhance the performance. It's one of good classifiers because it is known for its high accuracy due to the combination of multiple models, it has robustness. and is easy to use, which requires minimal parameter tuning. Also, can handle large datasets with high dimensionality.

By using grid search it shows that the best parameters for random forest are n_estimators=200, min_samples_split=2, max_features= 3, max_depth=15, random_state=42

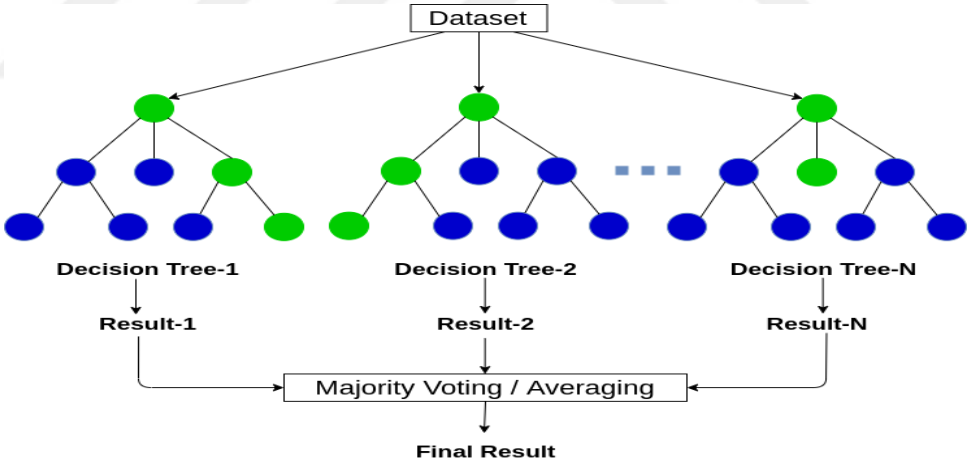

Figure 9 RF Model

Figure 9 (researchgate.net, 2023) shows random forest-based prediction process.

4. Decision Tree

A decision tree is a form of tree structure that looks like a flowchart and has internal nodes representing features, branches representing rules, and leaf nodes representing the algorithm's outcome. It divides the dataset into subsets depending on the most significant attribute at each node, producing a tree-like structure of

decisions. It is a versatile supervised machine-learning approach that may be used for regression and classification problems equally. It is one of the most potent algorithms. Random Forest also uses it to train on different subsets of training data. (geeksforgeeks.org, 2023).

The advantages of decision tree: it doesn't assume a linear relationship between features and the target variable, Implicit Feature Selection which elect features that are most informative for making decisions, Robust to outliers, it can be combined into powerful ensemble methods and it easy to implement. And it's highly interpretable.

By using grid search it shows that the best parameters for decision tree are criterion: gini, max_depth: 300, min_samples_split: 150.

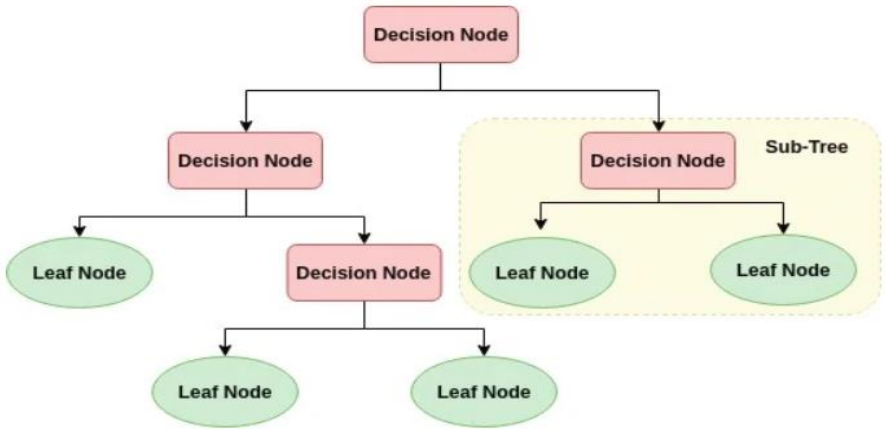

Figure 10 DT Model

Figure 10 (researchgate.net, 2021) shows decision tree -based prediction process.

5. Support Vector Machine

A supervised machine learning method that is used for regression as well as classification. The SVM method's primary goal is to find the best hyperplane in an N-dimensional space for partitioning data points into different feature space classes. The hyperplane attempts to maintain as large a buffer as feasible between the closest points of various classes. The hyperplane's dimension is defined by the number of features. (geeksforgeeks.org, 2023).

The Advantages of SVM: useful in situations with high dimensions because it

makes use of support vectors as a part of the decision function's training points its memory is effective. For the decision functions various kernel functions can be supplied, as well as customized kernels.

By using grid search it shows that the best parameter for support vector machine is kernel="linear".

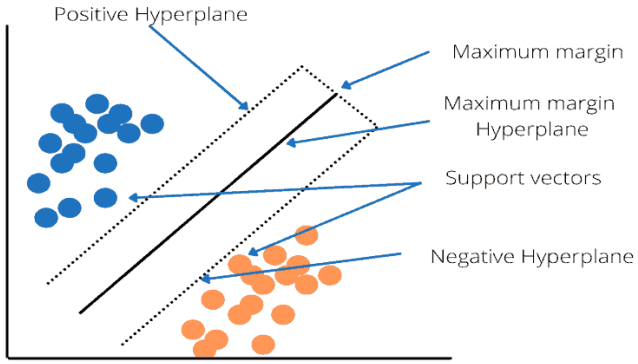

Figure 11 SVM Model

Figure 11 (golinuxcloud.com, 2021) shows how SVM works.

6. XGBoost

A distributed gradient boosting library optimized for maximum efficiency, versatility, and portability is called XGBoost. It belongs to the family of ensemble learning methods. It uses the Gradient Boosting framework to implement machine learning algorithms or classification and regression tasks. A parallel tree boosting method known as GBDT or GBM is provided by XGBoost, which effectively and swiftly resolves a variety of data science problems.

This technique generates decision trees consecutively. Weight plays a vital role in XGBoost. Each variable that is independent is weighted before being introduced into the decision tree, which predicts the outcome. The feature that the decision tree mistakenly forecasted are given extra weight and included in the other decision tree. After that, these individual classifiers are combined to create a strong and accurate model. The tasks it can perform include ranking, classification, and regression problems. (geeksforgeeks.org, 2023).

The advantages of XGBoost: it has a good performance which produces a high-quality result. Scalability which is suitable for large datasets. Interpretability

which provides feature importance. Customizability with a lot of hyperparameters can change it to optimize performance.

By using grid search it shows that the best parameters for xgboost are colsample_bytree: 1.0, gamma: 0.5, learning_rate: 0.1, max_depth: 5, min_child_weight: 1, subsample: 0.8.

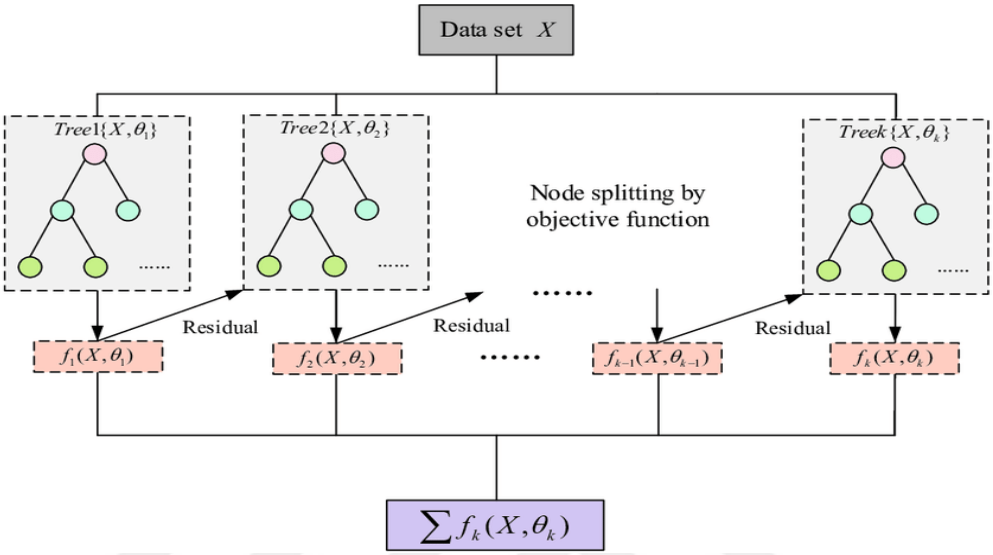

Figure 12 XGBoost Model

Figure 12 (researchgate.net, 2020) shows how XGBoost works.

7. Feature Selections

A crucial stage in the machine learning process is feature selection, this involves picking a subset of relevant features from the initial collection of features. This procedure is necessary to decrease overfitting, increase interpretability, and improve model performance. There are several feature selection methods, and the one chosen will rely on several variables, including the dataset, the specific issue at hand, and the intended model output.

Feature importance approach

Feature importance refers to methods that assign a number to each input feature in a model. The score assigned to each feature indicates its "importance". A higher score suggests that the characteristic will influence the model used to forecast a specific variable more. Feature importance scores can be calculated for regression and classification problems.

The scores are helpful and applicable to several scenarios in a predictive

modeling challenge, including:

- Get better knowledge of the data.
- improved comprehension of a model.
- cutting down on the quantity of input features.

After applying the Feature importance approach in all 5 machine learning models the figures below show the important features for each model. And use these features to train the models again to see the effect of the train just the important features instated of all the dataset.

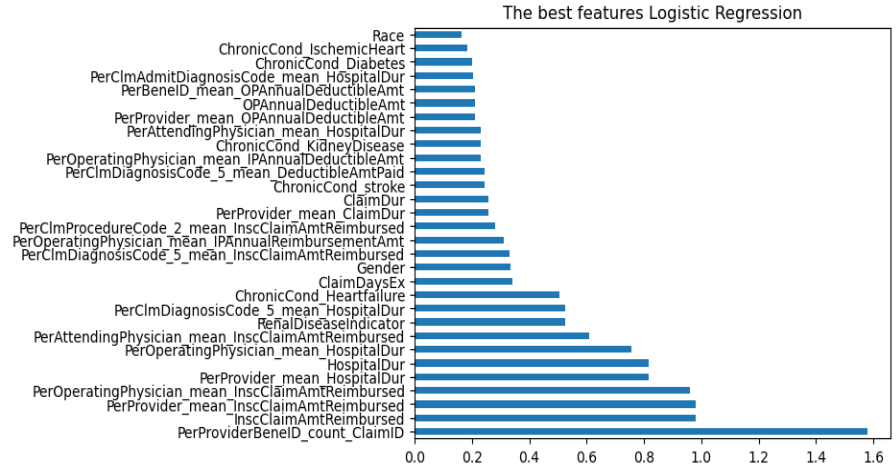

Figure 13 Most Important Features of Logistic Regression

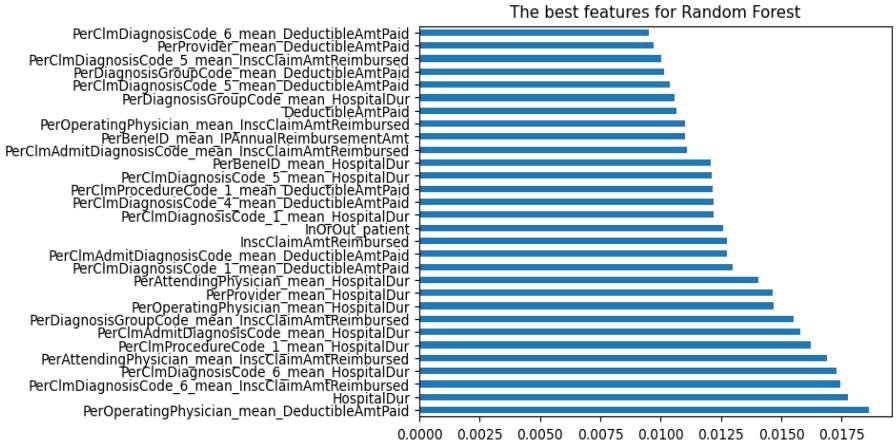

Figure 14 Most Important Features of Random Forest

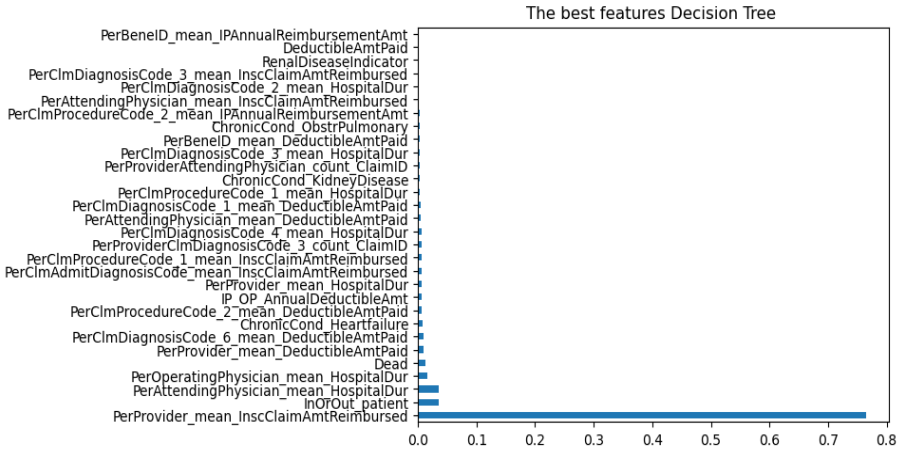

Figure 15 Most Important Features of Decision Tree

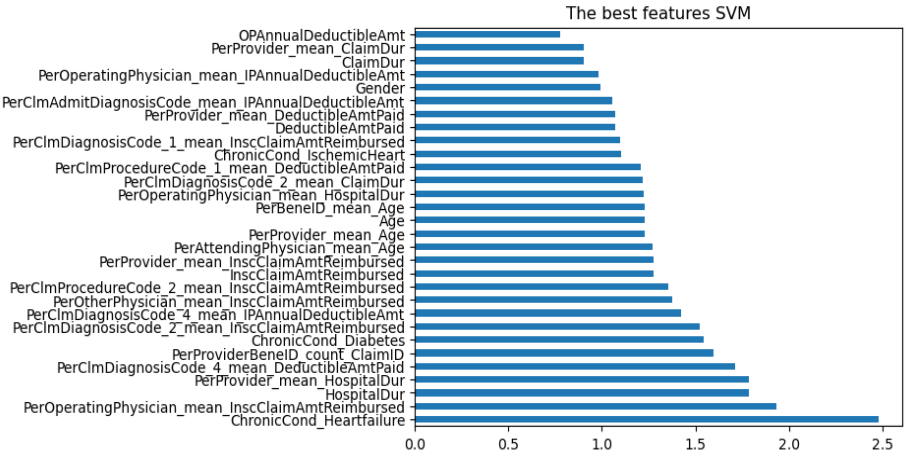

Figure 16 Most Important Features of SVM

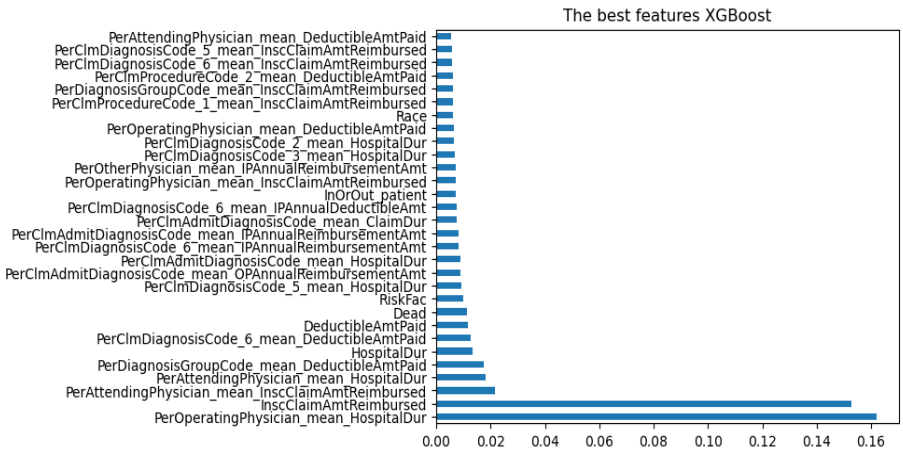

Figure 17 Most Important Features of XGBoost

8. Ensemble Learning

In this research 5 ML Models are used that's why by using Ensemble learning which is a machine learning technique that combines multiple individual models to create a more powerful model that can make better predictions.

Hard voting and soft voting are two methods for combining predictions from multiple machine learning models in ensemble learning. Both methods are used in classification tasks where each model predicts the class label for a given instance.

a. Hard Voting

In machine learning, hard voting is an ensemble learning strategy that is especially useful when dealing with categorization difficulties. In hard voting, a majority vote determines the final prediction after several models (classifiers) are trained separately on the same dataset (medium.com, 2023). And hard voting is works like that:

- Training Multiple Models: use the same training set of data to train several different classifiers. These classifiers might be variants of the same algorithm with altered hyperparameters or alternative algorithms altogether.
- Prediction: every trained model predicts the class label on its own when applied to fresh, untainted data.
- Voting: a majority vote determines the final prediction. The ensemble predicts a class based on which classifiers earn the greatest number of votes.

b. Soft Voting

One kind of ensemble learning technique used in machine learning, especially for classification issues, is soft voting. It belongs to the larger class of ensemble methods, which combine several different models to get a forecast. A final forecast is made by summing the class probabilities that the models anticipate for each instance in the soft voting scenario (medium.com, 2023). And this is the process of soft voting:

- Individual Model Predictions: For a given input instance, each model in the ensemble generates predictions on its own.
- Probabilities by Class: Each model generates a probability distribution over all possible classes for the input instance, as opposed to just one class prediction. The probabilities depict the level of confidence the model has for every class.
- Averaging Probabilities: For every class, the average of the class probabilities from every individual model is calculated. Taking the mean or weighted

mean of the probabilities is a common method for doing this.

- Final Prediction: Next, the class with the highest average probability is chosen to be the ensemble's final forecast.

E. Deep Learning

Deep learning is a branch of machine learning that specializes in deep neural networks, which are multi-layered neural networks. These networks automatically extract features at various levels of abstraction, allowing them to learn intricate hierarchical representations of data.

Artificial Neural Network

Artificial Neural Networks contain Units, or neurons. when the units are arranged in a series of layers they form Artificial Neural Network of a system. The layer can contain a few or many units, depending on the number of sophisticated neural networks required to find the dataset's hidden patterns. Typically, artificial neural networks have three layers: hidden, output, and input. The input layer is where external data enters the neural network for analysis or instruction. After that, the data passes via one or more hidden layers, which convert the input into usable data for the output layer. Lastly, the Artificial Neural Networks' reaction to the supplied input data is presented as an output by the output layer (geeksforgeeks.org, 2023).

Units are connected from one layer to another in most neural networks. The weights assigned to each of these relationships indicate how much effect one unit has upon the others. The neural network gains more and more knowledge about the data as it moves from one unit to the next, ultimately producing an output from the output layer (geeksforgeeks.org, 2023).

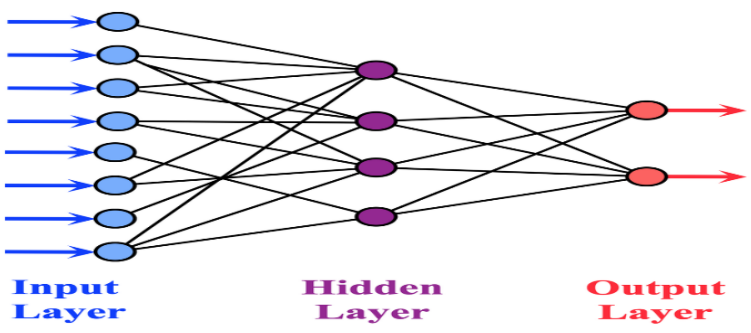

Figure 18 Artificial Neural Network Architecture

The model that used in this research was built like this first in input layer by using dense layer with 256 units and relu activation function, then hidden layers with 128 units and relu activation function, finally the output layer with 1 units and sigmoid activation function I fit the model with 100 epoch and got accuracy 96.7 and 96.27 validation accuracy, and for the training loss is 10.0, and validation loss is 12.0.

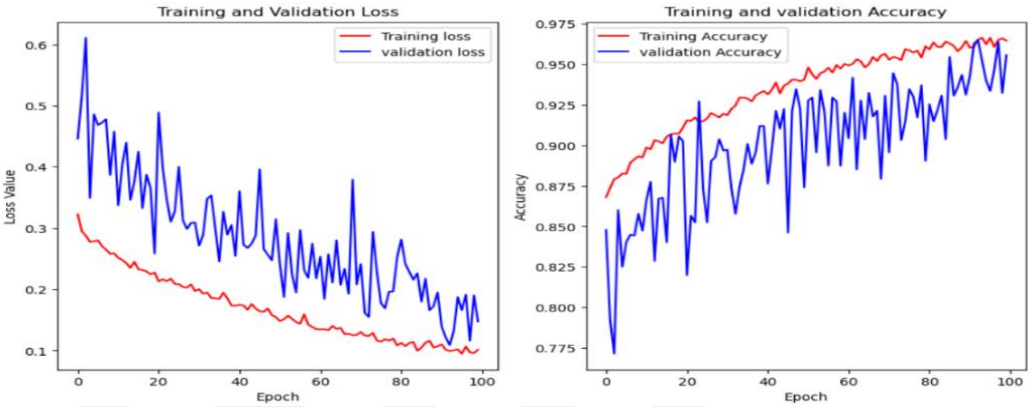

Figure 19 Training and Validation Accuracies and Losses

Figure 19 shows the accuracy and loss value of the training and validation results of the proposed model.

F. Principal Component Analysis (PCA)

Principal component analysis, or PCA, is a dimensionality reduction approach that is commonly used to reduce the dimensionality of large data sets by converting a large group of variables into a smaller one that retains the majority of the information in the larger set.

Accuracy naturally degrades when a data collection contains fewer variables; nonetheless, the key to dimensionality reduction is to sacrifice accuracy for simplicity. Smaller data sets make exploration and visualization easier, and since they have fewer processing variables, machine learning algorithms can evaluate data points faster and more easily.

And this is how PCA works:

- Standardization: standardize the continuous beginning variable range such that each one makes an equal contribution to the analysis. so that they have a mean of 0 and a standard deviation of 1. More particular, because PCA is

highly sensitive to the variances of the initial variables, normalization must be completed before PCA. In other words, the variables with larger ranges will prevail over those with narrower ranges if the initial variables' ranges differ noticeably, showing in biased findings. Therefore, this issue can be averted by translating the data to comparable scales.

- Compute the Covariance Matrix: This step's objective is to ascertain if a relationship exists among each variable in the input dataset and the ways in which they deviate from the mean relative to each other. Because variables might occasionally include redundant information due to strong correlations. Therefore, to determine these relationships, we compute the covariance matrix. The covariance matrix has all possible initial variable pairs as entries and is a symmetric matrix.
- Compute the eigenvectors and eigenvalues: to ascertain the principal components of the data, eigenvalues and eigenvectors from the covariance matrix must be calculated. and order them.
- Feature Vector: is a matrix containing the components' eigenvectors that we decide to keep as columns.
- Recast the data along the principal component’s axes: The objective is to use the feature vector produced by employing the eigenvectors of the covariance matrix to redirect the data from the original axis to the ones suggested by the principal components. To accomplish this, multiply the feature vector's transpose by the original data set's transpose (builtin.com, 2023).

By using PCA and setting n_components (dimensions) to 2, indicating that PCA should reduce the dimensionality of the data to two principal components where each data point is represented by a pair of values corresponding to these two principal components. The PCA score was 71.042.

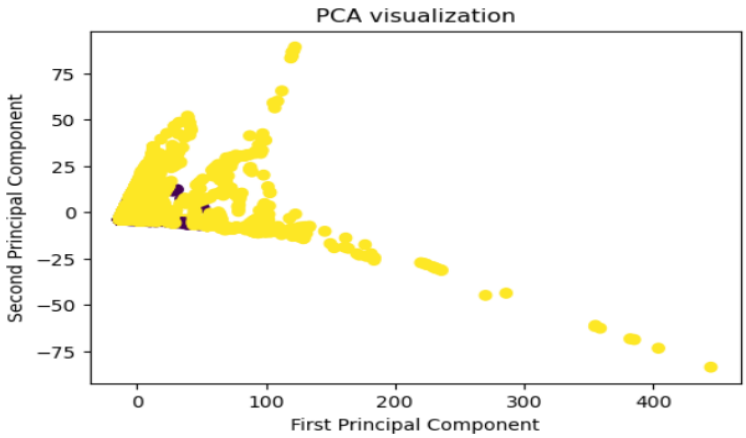

Figure 20 PCA Scatter Plot Visualization

In the scatter plot visualization resulting from PCA with 2 components, each data point corresponds to an observation in the original dataset. The position of each point in the plot is determined by the values of its two principal components. Points that are close together in the plot share similar characteristics or patterns in the original data space, while points that are far apart represent dissimilar observations.

IV. RESULT AND DISCUSSION

A. Confusion Matrix

A table utilized in classification to evaluate a machine learning model's execution is called a confusion matrix. It's a helpful tool for evaluating a classifier's performance and determining how well it performs across various classification which displays the number of true positive, false positive, true negative, and false negative predictions made by a model on a classification task (blogspot.com, 2019).

Some key terms in a confusion matrix:

- True Positives (TP): The quantity of cases that were accurately identified as the target class (i.e., correctly anticipated as positive).
- True Negative (TN): The quantity of cases that were accurately classified as negative (i.e., as a class other than the target class) and correctly predicted as negative.
- False positive (FP): The number of erroneous predictions that an example is positive that is, a negative class that is mistakenly recognized as positive.
- True Negative (TN): The number of accurate predictions that a given example is negative, or the number of negative classes that are correctly classified as negative (blogspot.com, 2019).

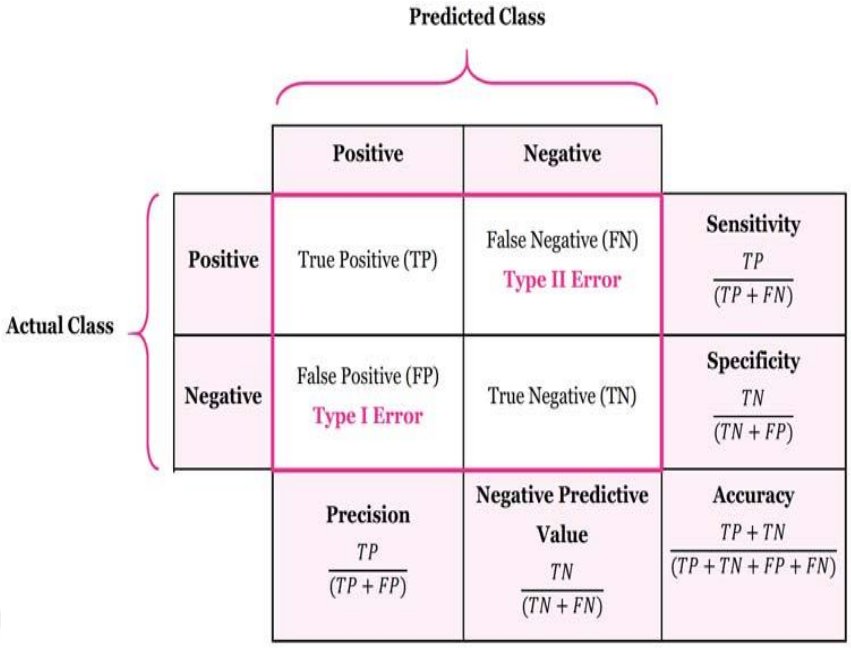

Figure 21 Confusion Matrix Explanation

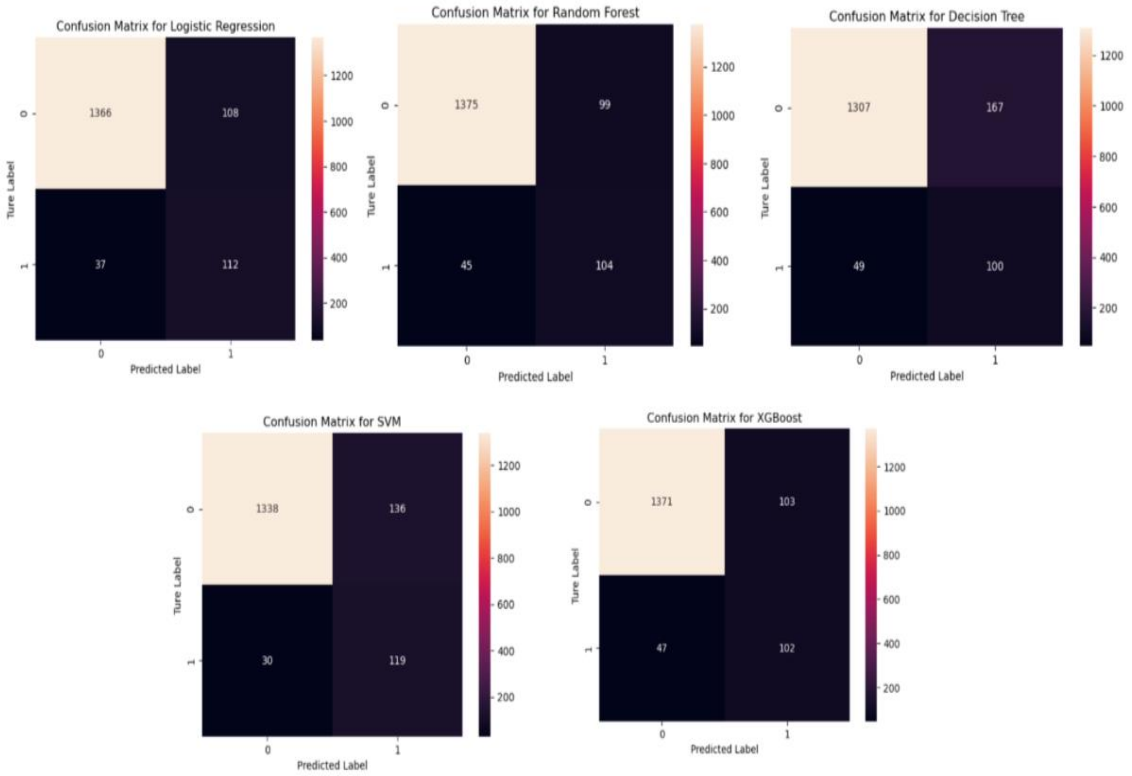

Figure 22 Confusion Matrices of All Models

Figure 22 shows the confusion matrix for all machine learning models.

B. Evaluation Matrecies

Accuracy

Calculates the percentage of correctly classified examples (both positive and negative) out of all occurrences, reflecting the overall correctness of the model.

And this is the formula of accuracy:

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

Precision

The ratio of all accurately classified positive examples to all projected positive examples is known as precision. It demonstrates correctness attained in positive prediction.

And this is the formula of precision:

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

Recall

Evaluates how well the model can represent every positive example. It is the percentage of actual positive instances out of all true positive predictions. A low percentage of false negatives is indicated by high recall.

And this is the formula of recall:

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

F1-Score

The F1 Score represents the harmonic means of recall and precision. It gives a balanced measure that takes into account FP and FN. It's especially helpful if there is an unequal class distribution.

And this is the formula of F1-score:

$$\text{F1Score} = \frac{2 \text{ x precision x recall}}{\text{precision} + \text{recall}}$$

Roc Curve and AUC Score

Using the AUC-ROC curve, the classification issues' performance is examined at a variety of threshold values. In contrast to ROC, which is a probability curve, AUC is a statistic that measures separability. It demonstrates how well the model can discriminate between classes. The higher the AUC, the better the model predicts 0 and 1 classes as zero and one. Likewise, the capacity of a model to distinguish between patients who have the condition and those who do not is indicated by a higher AUC (towardsdatascience.com, 2018).

TPR is placed on the y-axis and FPR is plotted on the x-axis to create the ROC curve.

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}$$

$$\mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{TN} + \mathrm{FP}}$$

Table 1 Evaluation Metrics of ML Models

ML Model	Accuracy	Precision	Recall	F1 – score	AUC score
LR	91.2	50.9	75.2	60.7	92.4
RF	91.1	51.2	69.8	59.1	92.1
DT	86.7	42.5	67.1	50.1	87.1
SVM	89.8	47.7	79.9	59.9	91.2
XGBoost	90.8	49.8	68.5	57.6	92.4

In Table 1 it shows the evaluation metrics for all machine learning models According to the results logistic regression has the best results in accuracy, f1-score and auc score among all machine learning models. whereas random forest has the best result in precision and svm has the best result in recall. And we can see that the decision tree shows the lowest performance according to its results.

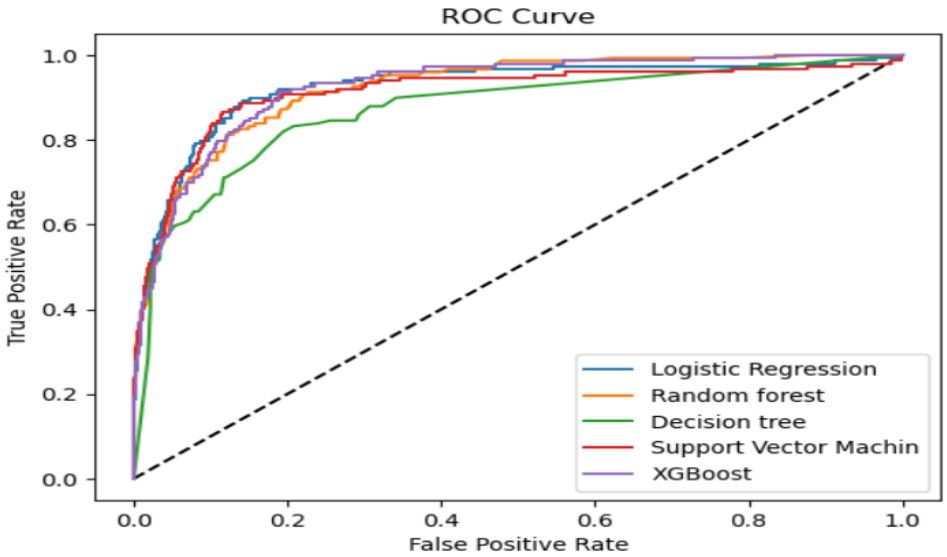

Figure 23 ROC Curves of All Models

In Figure 23 a model is well-classified if it can accurately distinguish between positive and negative cases, as shown by an AUC value of 1. Conversely, an AUC value of 0.5 points that the performance of the model is equivalent to that of random guesswork, where the false positive rate and true positive rate are identical. AUC values between 0.5 to 1 are typical, with larger values denoting superior performance. A number nearer 1 denotes a higher degree of classification capacity, whereas a score of 0.5 indicates that the model has no discriminatory power at all.

Table 2 Hard and Soft Voting Results.

ML Model	Accuracy	Precision	Recall	F1 – score	Auc score
Hard Vote	90.8	50.0	72.4	59.2	82.6
Soft Vote	90.9	52.2	72.4	59.3	82.6

In Table2. hard vote and soft vote which are methods to combine all ML models achieve high accuracy rates of around 90.8% to 90.9%. However, Soft Vote tends to have better precision and F1-score compared to Hard Vote, while maintaining similar recall and AUC scores.

Table 3 Evaluation Metrics of ML Models with Feature Selection

ML Model	Accuracy	Precision	Recall	F1 – score	Auc score
LR	87.2	41.4	72.5	51.9	90.8
RF	90.9	50.2	69.1	58.2	92.5
DT	89.8	46.4	68.5	53.3	88.8
SVM	91.2	46.8	77.1	67.2	92.3
XGBoost	90.3	48.0	72.5	57.8	92.4

Table3. shows all the models, especially the decision tree model, perform

noticeably better when feature selections are used. However, when it comes to feature selection, random forest and logistic regression do not yield better results.

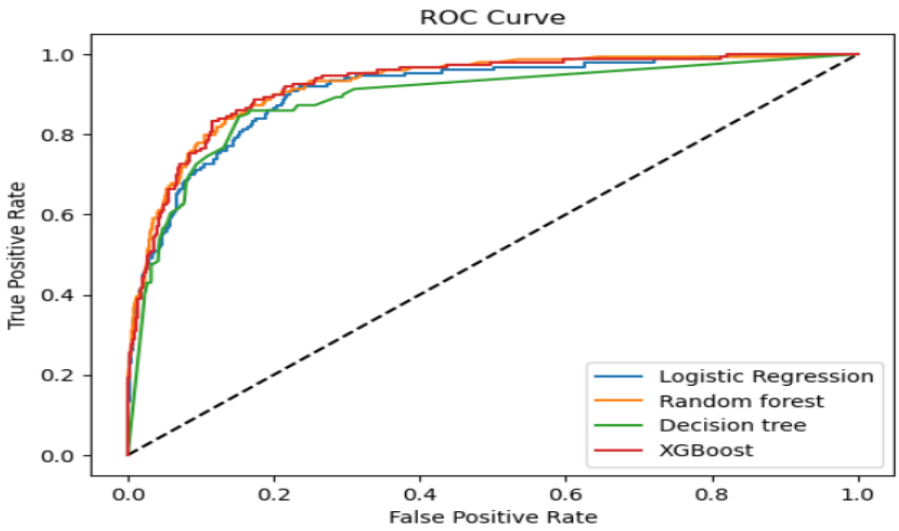

Figure 24 ROC Curves of All Models with Feature Selection

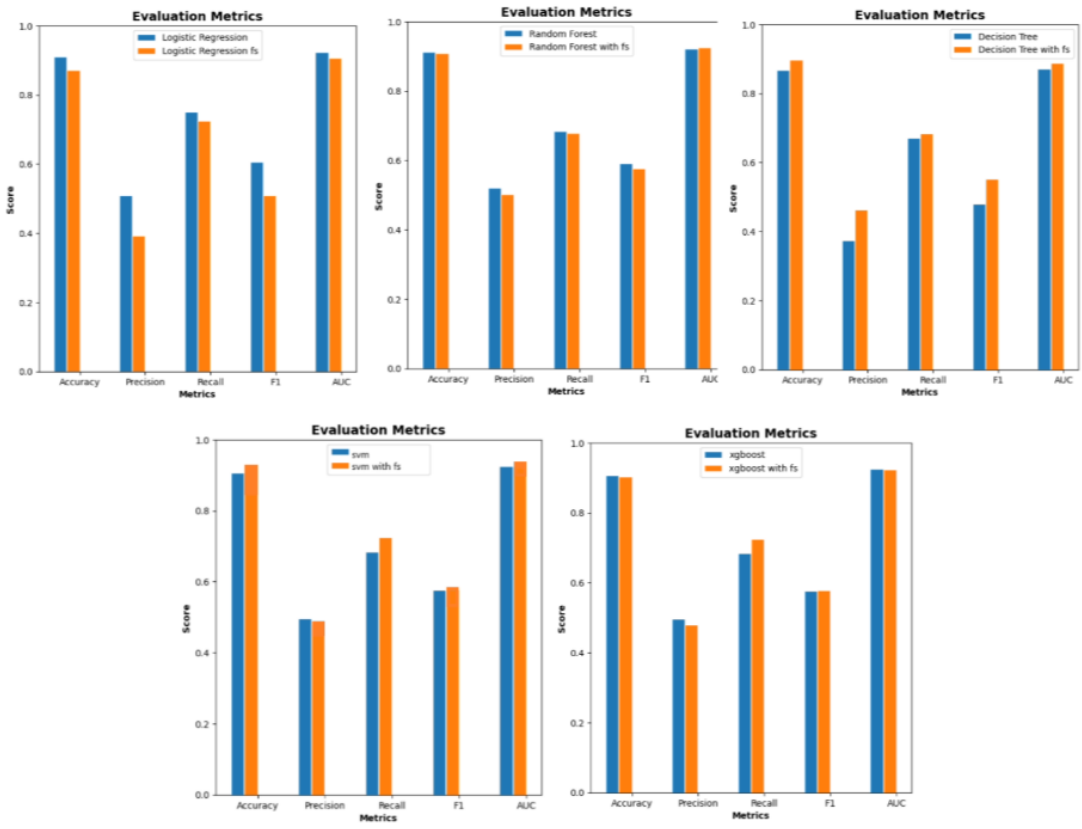

Figure 25 Comparisons

Figure 25 shows the graphs of all evaluation Metrics which compare each model with the model with feature selections the blue line represents the models, and the orange line represents the models with feature selections we can see that all the

models work much better with feature selections specially decision tree. On the other hand, logistic regression and random forest does not show a better result in feature selections.

Table 4 Runing Time

ML Model	Runing Time	Runing Time With Feature Selection
LR	0.069 sec	0.046 sec
RF	5.469 sec	5.096 sec
DT	7.290 sec	0.709 sec
SVM	7.756 sec	1.322 sec
XGBoost	2.044 sec	0.115 sec

In Table 4 the running time of the models with feature selection is faster than the models that train the hole data and it's clear that logistic regression is the fastest one for both cases because the combination of simplicity, efficient optimization algorithms, linear complexity, and potential for parallelization makes logistic regression a fast and scalable choice for many classification tasks.

ANN Result

	precision	recall	f1-score	support
0	0.96	0.94	0.95	1474
1	0.49	0.62	0.55	149
accuracy			0.91	1623
macro avg	0.73	0.78	0.75	1623
weighted avg	0.92	0.91	0.91	1623

Figure 26 ANN Classification Report

The classification report in figure 26 shows the performance of the ANN that made it in this research we can see that the average accuracy is 91% its very close to other classifiers that used in the research.

C. Discussion

choose logistic regression as the primary fraud detection model for health insurance claims due to its simplicity, interpretability, and consistent performance across research criteria. Logistic regression models provide a simple explanation of the predictions, making it easier to explain the factors that contribute to fraudulent claims. Furthermore, through techniques such as SHAP (SHapley Additive

exPlanations), logistic regression models can provide insight into the importance of different factors to predict fraud, increase transparency, understanding and even, performance consistently outperform other models in various metrics such as accuracy, f1-score and auc score. which showed its effectiveness in accurately detecting fraudulent activities. This combination of definition and efficiency makes logistic regression the best choice for fraud detection in health insurance claims.

D. Testing

	Provider	PotentialFraud	predicted_label
0	PRV51001	0	0
1	PRV51003	1	1
2	PRV51004	0	0
3	PRV51005	1	1
4	PRV51007	0	0
5	PRV51008	0	0
6	PRV51011	0	0
7	PRV51012	0	0
8	PRV51013	0	0
9	PRV51014	0	0
10	PRV51015	0	0
11	PRV51016	0	0
12	PRV51017	0	0
13	PRV51021	1	1
14	PRV51023	0	0
15	PRV51024	0	0
16	PRV51025	0	0
17	PRV51026	0	0
18	PRV51027	0	0
19	PRV51029	0	0
20	PRV51030	0	1

Provider	predicted_label
PRV51002	0
PRV51006	0
PRV51009	0
PRV51010	0
PRV51018	0
PRV51019	0
PRV51020	0
PRV51022	0
PRV51028	0
PRV51033	0
PRV51034	0
PRV51039	0
PRV51050	0
PRV51051	0
PRV51069	0
PRV51073	1
PRV51079	1
PRV51080	0
PRV51085	0
PRV51087	0
PRV51088	0

Figure 27 The Result of Testing the Model

Figure 27 shows when after applying the logistic regression model on train and test datasets the table on the left side is the training dataset, we can see the columns potential fraud which already exists on the dataset and predicted_label that it is created after applying the model on it and almost show the same results. On the right side is the test dataset and after applying the model it shows the predicted_lable of these providers which 0 means non fraud and 1 is fraud.

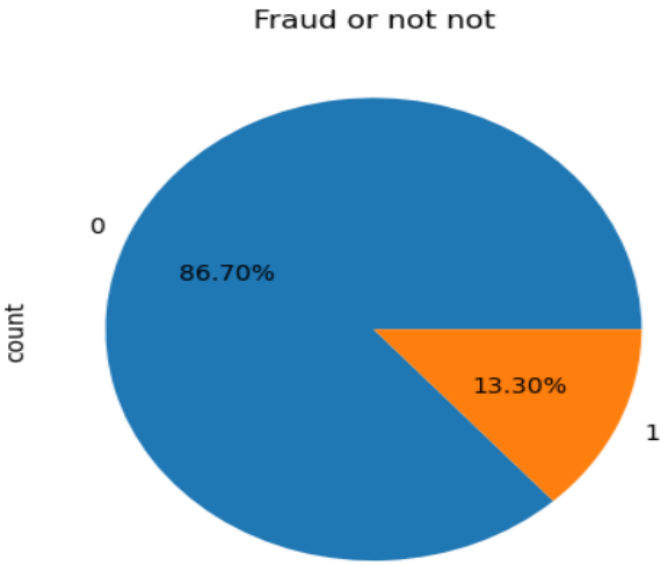

Figure 28 Pie Chart Test Result

In Figure 28 shows the pie chart of test dataset of the providers whether they are fraud or not and we can see 86.7% of them are not fraud and 13.3% are fraud.

V. EXPLAINABLE AI

The goal of explainable AI is to convert complicated "black box" AI models and system outputs into intelligible representations, or simpler or naturally explainable models (Nicodeme, 2020: 20-23). Explainable AI refers to an AI model's ability to provide human-comprehensible explanations for its decisions or forecasts.

The capacity to comprehend and analyze the predictions produced by machine learning models is known as "model explainability." It is especially crucial for intricate models where the decision-making process might not be immediately obvious, like ensemble models or deep neural networks. Model explanations can help build confidence, make debugging easier, and shed light on how certain aspects affect predictions.

The definitions of interpretability, explainability, and transparency terms. are illustrate below:

- Interpretability is the ability to convey meaning to a human being in a way that they can comprehend.
- Explainability is related to the idea of explanation as a human-machine interface that is both a human-understandable representation of the decision maker and an accurate proxy for the decision maker.
- Transparency: If a model can be understood on its own, it is said to be transparent. Transparent models are categorized into three groups based on their varying degrees of understandability: simulatable models, decomposable models, and algorithmically transparent models (Arrieta, et al., 2020: 82-115).

One of XAI libraries is SHAP which provides a particular approach within the realm of XAI.

A. SHAP

Shapley Additive Explanations (SHAP) is a technique for creating artificial intelligence that is used with machine learning models to interpret their predictions. It can be applied to elucidate any model's predictions. Because of this, SHAP is a very well-liked and helpful XAI library for analyzing your machine learning models.

SHAP is a game theoretic method for interpreting any machine learning model's output. It makes use of the traditional Shapley values from game theory and their extensions to establish a connection between local explanations and optimal credit distribution. Shapley values offer a mechanism to divide a value among contributors in an equitable manner. SHAP values are employed in the machine learning environment to ascribe each feature's contribution to the prediction for a specific instance (medium.com, 2023).

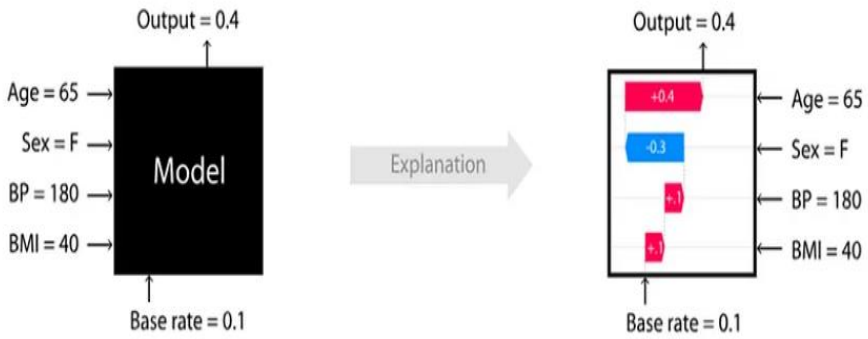

Figure 29 SHAP

Figure 29 (medium.com, 2023) It provides a way to understand the importance of features in generating predictions from black-box models, meaning models where the internal workings are not easily interpretable.

Model debugging using SHAP values.

Finding and fixing problems that arise during the training and evaluation stages of machine learning models is a crucial activity known as "model debugging." This is the area in which SHAP values become quite helpful. They support us in the following ways (neptune.ai, 2023):

- Testing the robustness of the model
- Identifying model bias

- Investigating model behavior
- Determining characteristics that influence prediction.

The advantages of using shap in insurance claims fraud detection:

- Enhances Interpretability: SHAP provides insights into factors affecting individual predictions in insurance claims fraud detection.
- Identifies Feature Importance: SHAP identifies high-impact factors contributing to prediction, making fraud detection systems more efficient.
- Justifies Decisions: Quantifying the impact of each aspect on prediction results helps justify decisions made by the fraud detection model.
- Facilitates Early Detection of Anomalies: SHAP aids in rapid anomaly detection by highlighting unusual patterns or discrepancies.
- Improves Model Improvements: SHAP insights can refine existing fraud detection models, improving accuracy and robustness.

B. SHAP Graphs

The SHAP library (graph) is a powerful tool for interpreting machine learning models. It provides insights into the contributions of individual features to model predictions.

1. Beeswarm Plot

One kind of scatter plot that shows individual data points along one axis without overlapping is called a beeswarm plot. The distribution of SHAP values for each feature across a dataset may be shown in the context of SHAP using a beeswarm Plot, which makes it possible to comprehend how different features affect the predictions made by the model.

The purpose of the beeswarm plot is to present a concise overview of the relationship between the most important features in a dataset and the model's output. A single dot on each feature represents each occurrence of the explanation provided. Dots "pile up" along each feature row to indicate density; the x position of the dot is dictated by the feature's SHAP value, the values on the y-axis are organized by features. The features in the above plot are ordered by mean SHAP. The feature

value determines the color of the dots in each group, the greater feature values are redder. In x-axis the right-hand side means class 1 which is more fraudulent, and the left-hand side means class 0. Red color means high feature values means high numeric values; blue color means low feature values means low numeric values (medium.com, 2023).

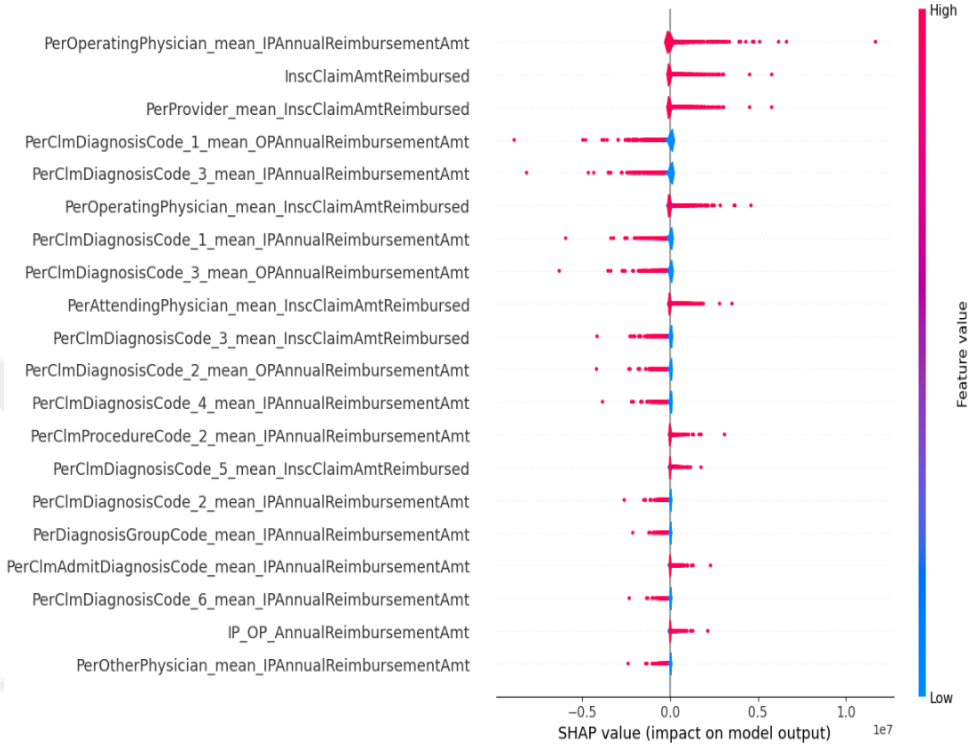

Figure 30 Beeswarm Plot

In Figure 30 explain the features and how they effect fraudulent providers for example preoperatingphyscian_mean_ipannualreimbursementamt when this feature is increased the probability to be more fraud provider. On the other hand, for perclmdiagnosiscode_1_mean_Opannualreimbursementamt when this feature is increased the probability to be more non fraud provider and so on.

2. Waterfall Plot

A waterfall plot shows how components are added one after the other in a graphical manner. A waterfall plot, when used in conjunction with SHAP, shows how each feature contributes to the variation between the model's output and a reference (often the model's baseline output or mean forecast). It facilitates seeing the relative contributions of each feature to the final prediction. With positive contributions at the top and negative contributions at the bottom, waterfall plots

display the features in a top-down manner. The expected value of the model output appears at the bottom of a waterfall plot, and each row illustrates how each feature's positive (red), or negative (blue) contribution shifts the value based on the anticipated model result throughout the dataset that serves as the model's background for this prediction (medium.com, 2023).

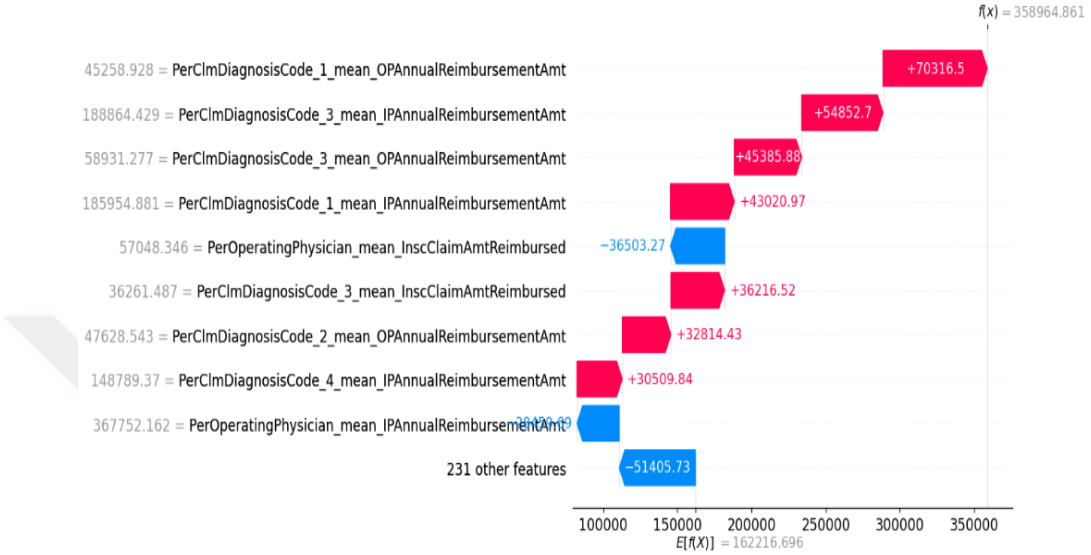

Figure 31 Waterfall Plot

In Figure 31 E[f(x)] =1622163696 gives the average predicted number of potential frauds across all 5410 records. f(x)=358964.861 is the predicted number of potential frauds for this particular record. The SHAP values are all the values in between.

E.g. the preclmdiagnosiscode_3_mena_ipannualreimbursementAmt has increased the predicted number of potential frauds by 70316.5.

3. Force Plot

A comprehensive visual depiction of the SHAP values for every feature for a particular instance is offered by a Force Plot in SHAP. It illustrates how each feature affects the variation between the expected model output (mean prediction or baseline) and the model's output for a particular instance. Knowing the direction and strength of each feature's influence on a given forecast is helpful. The positive contributions are on the left and the negative contributions are on the right in SHAP force plots, which display the features from left to right (medium.com, 2023).

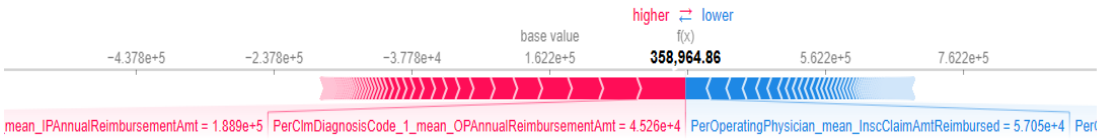

Figure 32 Force Plot

Figure 32 can consider force plot to be a simplified waterfall plot. We begin with the same base value of 162,200 and examine how each attribute contributes to the final forecast of 358964.86. so, the features in red increase the probability of fraud provider and the features in blue decrease it.

Stacked Force Plot

Stacked Force plot is the vertical perspective of force plot when analyzing misclassified instances and learning more about the causes behind those misclassifications, the stacked force plot is quite helpful. This makes it possible to comprehend the model's decision-making process better and helps identify areas that need more research or development.

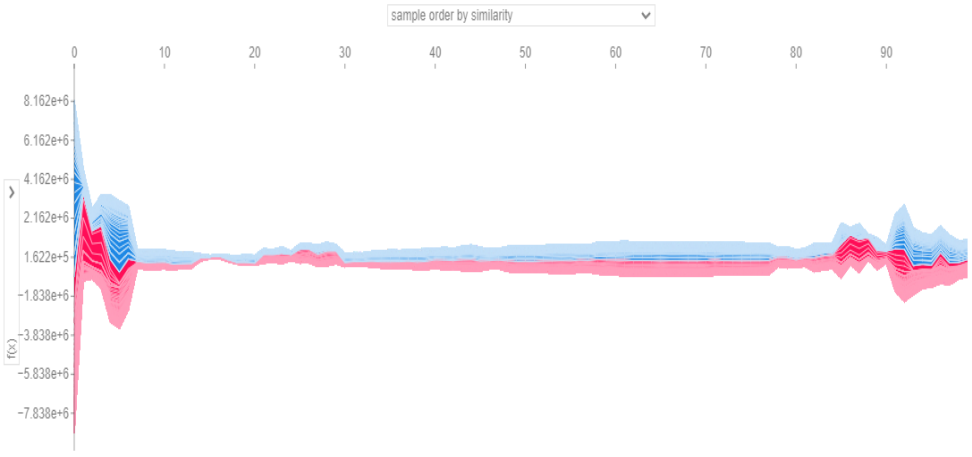

Figure 33 Stacked Force Plot

4. Bar Plot

To see the average absolute SHAP values for every feature over the course of the dataset, a bar plot is frequently utilized. It aids in determining which features are most important by summarizing each feature's overall influence on model predictions (medium.com, 2023).

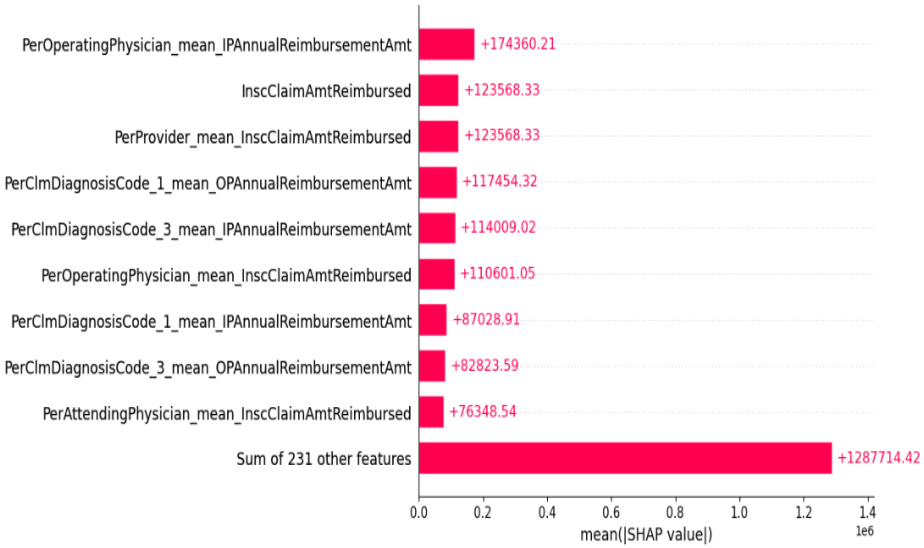

Figure 34 Bar Plot

In Figure 34 preopertingphysican_mean_ipannualremibursementamt had the highest mean SHAP value.

5. Heatmap Plot

A heatmap is plot in SHAP shows the SHAP values for several features over a group of instances, illuminating the various ways in which features affect predictions made by a dataset. Each cell's color intensity reveals the degree and direction of the feature's influence on the prediction for that occurrence. A display with the examples on the x-axis, the inputs on the y-axis, and the SHAP values encoded on a color scale is produced by passing a matrix of SHAP values to the heatmap plot function. The samples are arranged by default in a hierarchical clustering order determined by the similarity of their explanations (medium.com, 2023).

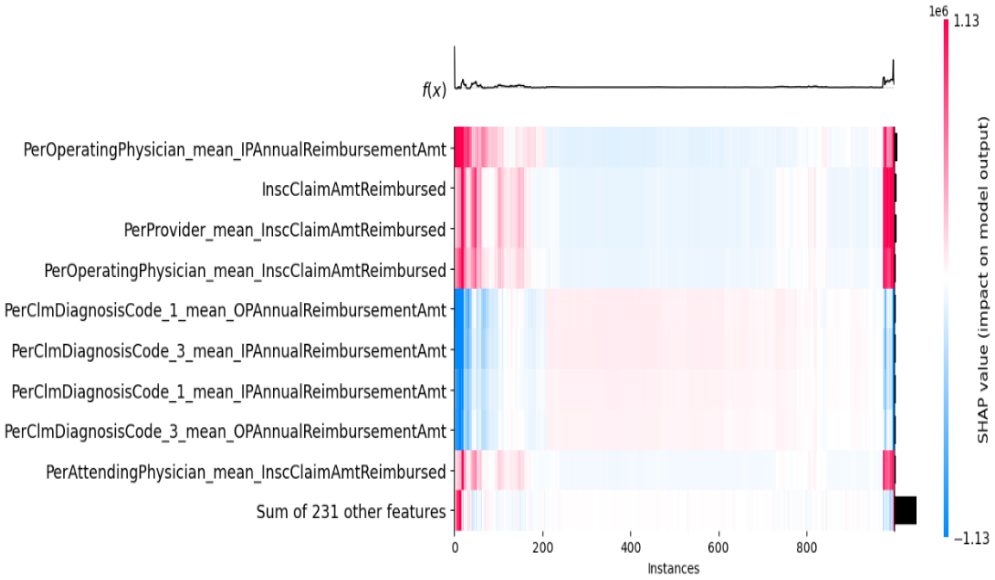

Figure 35 Heatmap Plot

C. SHAP Results

Because SHAP provides information on how each feature affects a model's prediction, it is essential for fraud detection. By identifying the primary characteristics impacting questionable claims with SHAP, insurers can improve the interpretability of their fraud detection algorithms. The results of SHAP for each graph are written under the graphs.

VI. DEPLOYMENT

A. Flask

A simple and adaptable web framework for Python web application development is called Flask. It adheres to the WSGI (Web Server Gateway Interface) standard and is made to be straightforward and simple to use. It is categorized as a microframework since it doesn't need any specific libraries or tools. Flask gives developers the freedom to select extra libraries and components according to their needs while still offering the fundamental tools required to create a web application. It lacks any form validation, database abstraction layer, or other features for which existing third-party libraries provide common functionality.

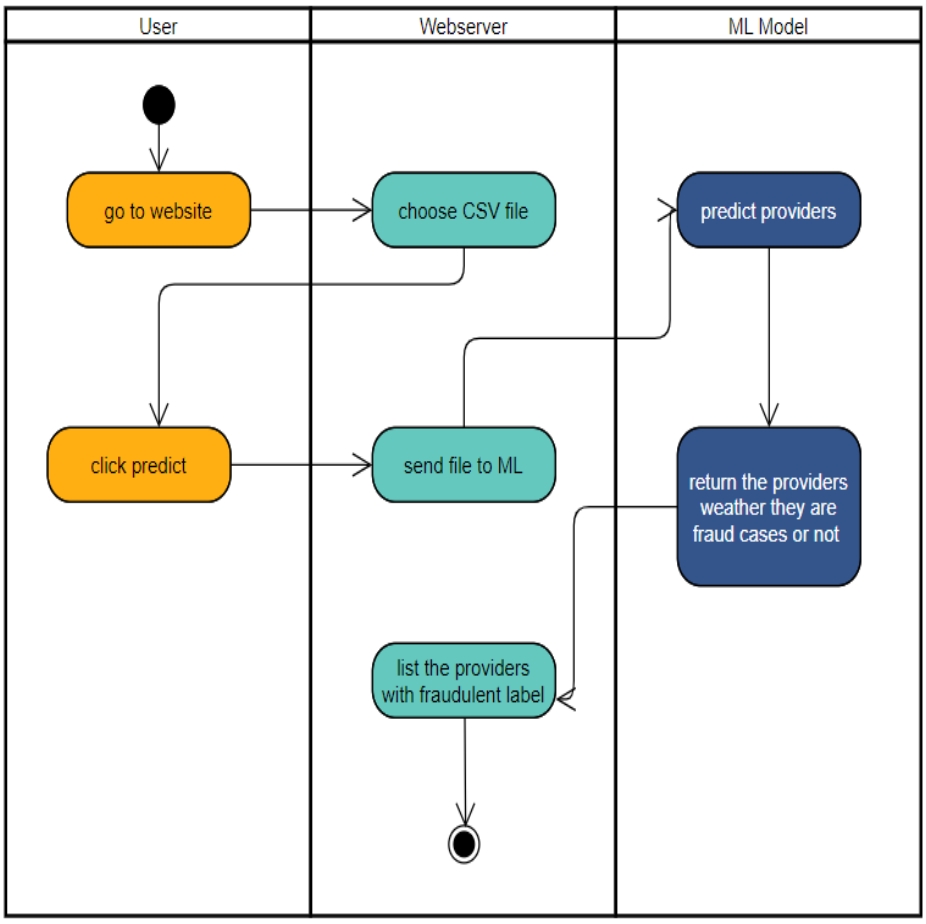

Figure 36 Activity Diagram

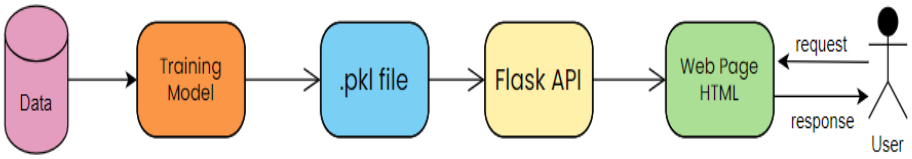

Figure 37 Flask Flowchart

Figure 36 shows how Flask can be used to deploy a machine learning model and present its results through an HTML template by steps:

- Machine Learning Model: first to begin with, by trained machine learning model.
- Pickled Model File (PKL): Once there is a trained model, it should typically save it to a file for later use. In Python, one common way to save a model is by using the pickle module to serialize it and save it to a file with pkl extension.
- Flask Application: create a Flask web application. Flask is a micro web framework for Python, which helps to build web applications. by writing a Python script that defines a Flask application.
- Python File: In this Python file, load the pickled model into memory. This involves importing the necessary libraries, loading the pickled model file, and defining routes for the Flask application.
- HTML Template: Next, create an HTML template that defines the structure and layout of the web page where you want to display your machine learning results. This HTML template can include placeholders where you'll dynamically insert the results computed by your machine learning model.
- Integration: In Flask application, specify routes that handle HTTP requests from clients (web browsers). When a client accesses a specific URL Flask will call the corresponding Python function (view function) that you've defined.
- Request Handling: Within these Python functions, you'll handle the incoming requests. For example, if a client submits a form with image data, you'll

extract that data, preprocess it if necessary, and pass it to your machine learning model for prediction.

- Prediction: the machine learning model will make predictions based on the input data it receives. Once the prediction is made, the results will format appropriately.
- Rendering HTML: Finally, render the HTML template, passing the results of the prediction as variables. Flask will merge these variables with the HTML template, generating a complete HTML page, which is then sent back to the client's browser for display.
- Client Interaction: The client's browser will receive the HTML response from the Flask application and render it for the user to see. They can interact with the web page, submit new requests, and receive updated predictions as per the model's responses.

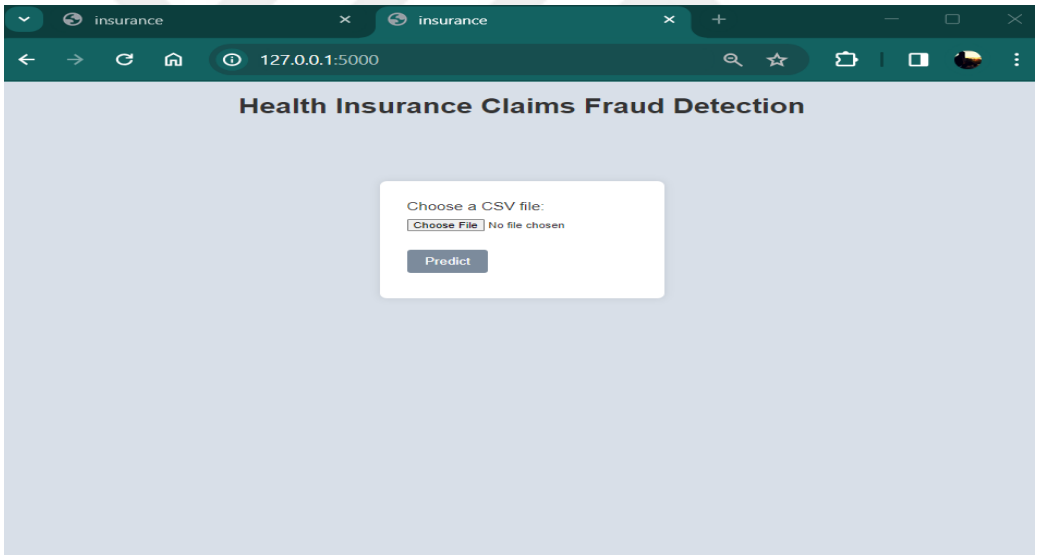

Figure 38 Web Interface

Figure 38 shows the HTML template of the web application which when we choose a file of providers and click predict it will show each provider and the predication whether fraud is or not.

Figure 39 Results

VII. CONCLUSION AND FUTURE WORK

In conclusion, the research addressed the crucial issue of insurance claim fraud detection, aiming to enhance the efficiency and accuracy of identifying fraudulent activities within insurance claims. Despite encountering challenges such as class imbalance within the Medicare dataset from Kaggle consisting of four files These files consider the beneficiary information, outpatient, and inpatient claims, and by using EDA to visualize the data to get more understanding about this dataset. Also, by applying preprocessing techniques to get ready to work with the dataset. And, by creating a dataset that is a combination of these datasets. The dataset is used for supervised classification problem which classifies the Healthcare Provider whether fraud or not. successfully applied the SMOTE Technique to mitigate this imbalance and ensure robust model performance. Leveraging various machine learning (ML) algorithms including Logistic Regression (LR), Random Forest (RF), Decision Trees (DT), Support Vector Machines (SVM), and XGBoost, alongside ensemble learning techniques such as hard and soft voting, conducted a comprehensive evaluation of fraud detection performance. Additionally, apply a feature selection approach to compare the outcomes and enhance the model's performance. Furthermore, investigates the relationship between deep learning and fraud detection using artificial neural networks. Also, applying the PCA technique to reduce the dimensionality of the data to two principal components. The results show that Logistic Regression (LR) performs best among the models considered with an accuracy of 91.2%, f1-score 60.7%, and auc score of 92.4%. To enhance understanding of the decision-making process in the LR model, by using Explainable AI (XAI) methodologies like SHAP (SHapley Additive exPlanations) method. The SHAP study provided valuable insight into the importance of various factors in determining the likelihood of insurance claim fraud. And by using SHAP plots like beeswarm, waterfall, force, bar, and heatmap plots These insights not only enhance interpretation but also lay the foundation for model reconstruction and synthesis.

For future work, the aim is to work with more Explainable Artificial

Intelligence (XAI) techniques to significantly improve performance and further our understanding of model decisions. More specifically, want to use techniques like G-REX and Local Interpretable Model-agnostic Explanations (LIME) to explore the models' internal mechanisms in greater detail. With LIME, interpretability can be improved by providing insights into specific predictions through the approximation of their decision bounds. Furthermore, G-REX will help comprehend the models' overall behavior, which will improve feature selection and model refining. By incorporating these XAI methods, hoping to provide model outputs that are more transparent and dependable, opening the door to improved performance and interpretability in upcoming projects.

VIII. REFERENCES

ARTICLES

ARRIETA, A. B., DÍAZ-RODRÍGUEZ, N., DEL SER, J., BENNETOT, A., TABIK, S., BARBADO, A., & ... & HERRERA, F. (2020). "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI". **Information fusion**, 58, 82-115.

BAUDER, R. A., & KHOSHGOFTAAR, T. M. (2017, December). "Medicare fraud detection using machine learning methods". **international conference on machine learning and applications (ICMLA)**, 16th, 858-865.

BAUDER, R. A., KHOSHGOFTAAR, T. M., RICHTER, A., & HERLAND, M. (2016, November). "Predicting medical provider specialties to detect anomalous insurance claims". (T. M. KHOSHGOFTAAR, Ed.) **international conference on tools with artificial intelligence (ICTAI)**, 28th, 784-790.

DEBENER, J., HEINKE, V., & KRIEBEL, J. (2023). "Detecting insurance fraud using supervised and unsupervised machine learning". **Journal of Risk and Insurance**, 90(3), 743-768.

DHIEB, N., GHAZZAI, H. B., & MASSOUD, Y. (2020). "A secure ai-driven architecture for automated insurance systems: Fraud detection and risk measurement". 8, 58546-58558.

GUPTA, R. Y., MUDIGONDA, S. S., & BARUAH, P. K. (2021). "A comparative study of using various machine learning and deep learning-based fraud detection models for universal health coverage schemes". **International Journal of Engineering Trends and Technology**, 69(3), 96-102.

HANCOCK, J. T., & KHOSHGOFTAAR, T. M. (2021). "Gradient boosted decision tree algorithms for medicare fraud detection". **SN Computer Science,** 2(4), 268.

HANCOCK, J. T., BAUDER, R. A., Wang, H., & KHOSHGOFTAAR, T. M. (2023). “Explainable machine learning models for Medicare fraud detection”. **Journal of Big Data**, 10(1), 154.

HAQUE, M. E., & TOZAL, M. E. (2021). “Identifying health insurance claim frauds using mixture of clinical concepts”. **IEEE Transactions on Services Computing**, 15(4), 2356-2367.

HERLAND, M., KHOSHGOFTAAR, T. M., & BAUDER, R. A. (2018). “Big data fraud detection using multiple medicare data sources”. **Journal of Big Data**, 5(1), 1-21.

JOHNSON, J. M., & KHOSHGOFTAAR, T. M. (2019). “Medicare fraud detection using neural networks”. **Journal of Big Data**, 6(1), 1-35.

JOHNSON, J. M., & KHOSHGOFTAAR, T. M. (2022, August). “Healthcare Provider Summary Data for Fraud Classification”. IEEE 23rd **International Conference on Information Reuse and Integration for Data Science (IRI)**, 236-242.

JOHNSON, M. E., & NAGARUR, N. (2016). “Multi-stage methodology to detect health insurance claim fraud”. **Health care management science**, 19, 249-260.

LU, F., & BORITZ, J. E. (2005, October). “Detecting fraud in health insurance data: Learning to model incomplete Benford’s law distributions”. **In European Conference on Machine Learning**, 633-640.

LUNDBERG, S. M., & LEE, S. I. (2017). "A unified approach to interpreting model predictions". **Advances in neural information processing systems**, 30.

NABRAWI, E., & ALANAZI, A. (2023). “Fraud Detection in Healthcare Insurance Claims Using Machine Learning”. **Risks**, 11(9), 160.

NICODEME, C. (2020, June). “Build confidence and acceptance of AI-based decision support systems-Explainable and liable AI” . **13th international conference on human system interaction (HSI)**, 20-23.

PENG, Y., KOU, G., SABATKA, A., CHEN, Z., KHAZANCHI, D., & SHI, Y. (2006, October). “Application of clustering methods to health insurance fraud detection”. **International Conference on Service Systems and**

Service Management, 1, 116-120.

RAGHAVAN, P., & El GAYAR, N. (2019, December). “Fraud detection using machine learning and deep learning". **international conference on computational intelligence and knowledge economy (ICCIKE)**, 334-339.

RAWTE, V., & ANURADHA, G. (2015, January). “Fraud detection in health insurance using data mining techniques”. **International Conference on Communication, Information & Computing Technology (ICCICT)**, 1-5.

RUKHSAR, L., BANGYAL, W. H., NISAR, K., & NISAR, S. (2022). “Prediction of insurance fraud detection using machine learning algorithms”. **Mehran University Research Journal of Engineering & Technology**, 41(1), 33-40.

SALDAMLI, G., REDDY, V., BOJJA, K. S., GURURAJA, M. K., DODDAVEERAPPA, Y., & TAWALBEH, L. (2020, April). “Health care insurance fraud detection using blockchain”. **seventh international conference on software defined systems (SDS)**, 145-152.

SETTIPALLI, L., & GANGADHARAN, G. R. (2023). “WMTDBC: An unsupervised multivariate analysis model for fraud detection in health insurance claims”. **Expert Systems with Applications**, 215.

SMITA, K., PRANATHI, D., PRAVALIKA, D., SUPRAJA, E., & HARIKA, G. . (2023). “Detection of Fraudulent Medicare Providers using Decision Tree and Logistic Regression”. **Journal of Cardiovascular Disease Reasearch**, 14(7).

SUN, C., L. Q., LI, H., SHI, Y., ZHANG, S., & GUO, W. (2018). “Patient cluster divergence based healthcare insurance fraudster detection”. 7, 14162-14170.

THORNTON, D., VAN CAPELLEVEEN, G., POEL, M., VAN HILLEGERSBERG, J., & MUELLER, R. M. (2014, April). “Outlier-based Health Insurance Fraud Detection for US Medicaid Data”. **ICEIS**(2), 684-694.

VERMA, A., TANEJA, A., & ARORA, A. (2017, August). “Fraud detection and frequent pattern matching in insurance claims using data mining techniques”. **tenth international conference on contemporary computing (IC3)**, 1-7.

DISSERTATIONS

DANGERS, L. . (2022). “Fraud: and anomaly detection in healthcare: an unsupervised machine learning approach” . Doctoral dissertation.

JI, Y. (2021). “Explainable AI methods for credit card fraud detection: Evaluation of LIME and SHAP through a User Study”.

ELECTRONIC SOURCES

URL-1 “What is exploratory data analysis (EDA)?”. (2024). Retrieved from IBM: https://www.ibm.com/topics/exploratory-data-analysis

URL-2 “Pandas dataframe.groupby() Method”. (2023, September 4). Retrieved from geeksforgeeks: https://www.geeksforgeeks.org/python-pandas-dataframe-groupby/

URL-3 dholakiya, p. (2023, April 26). SMOTE (Synthetic Minority Over-sampling Technique)”. Retrieved from medium: medium, https://medium.com/@parthdholakiya180/smote-synthetic-minority-over-sampling-technique-4d5a5d69d720,

URL-4 “What is Hyperparameter Tuning?”. (2024). Retrieved from AWS: https://aws.amazon.com/what-is/hyperparameter-tuning/

URL-5 “Logistic Regression in Machine Learning”. (2024, April 11). Retrieved from geeksforgeeks: https://www.geeksforgeeks.org/understanding-logistic-regression/

URL-6 Kanade, V. (2022, April 18). “What Is Logistic Regression? Equation, Assumptions, Types, and Best Practices”. Retrieved from spiceworks: https://www.spiceworks.com/tech/artificial-intelligence/articles/what-is-logistic-regression/

URL-7 “Random Forest based prediction process”. (2023). Retrieved from ReasearchGate: https://www.researchgate.net/figure/Random-forest-

based-prediction-process_fig3_371178775

URL-8 “Decision Tree”. (2023, August 20). Retrieved from geeksforgeeks: https://www.geeksforgeeks.org/decision-tree/

URL-9 “Decision Tree Algorithm Implementation”. (2021). Retrieved from ResearchGate: https://www.researchgate.net/figure/Decision-Tree-Algorithm-Implementation_fig4_351897936

URL-10 “Support Vector Machine (SVM) Algorithm”. (2023, June 10). Retrieved from geeksforgeeks: https://www.geeksforgeeks.org/support-vector-machine-algorithm/

URL-11 Alam, B. (2021, December 29). “Supervised Learning Algorithms Explained [Beginners Guide]”. Retrieved from golinuxcloud: golinuxcloud, https://www.golinuxcloud.com/supervised-learning-algorithms/

URL-12 “xgboost”. (2023, February 6). Retrieved from geeksforgeeks: https://www.geeksforgeeks.org/xgboost/

URL-13 “XGBoost". (2020, September 1). Retrieved from researchgate: https://www.researchgate.net/figure/Flow-chart-of-XGBoost_fig3_345327934,

URL-14 Ahmed, I. (2023, May 31). “What is Hard and Soft Voting in Machine Learning”. Retrieved from medium: https://ilyasbinsalih.medium.com/what-is-hard-and-soft-voting-in-machine-learning-2652676b6a32

URL-15 Jaadi, Z. (2024, February 23). “A Step-by-Step Explanation of Principal Component Analysis (PCA)”, builtin". Retrieved from builtin: https://builtin.com/data-science/step-step-explanation-principal-component-analysis

URL-16 “artificial neural networks and its applications”. (2023, June 2). Retrieved from geeksforgeeks: https://www.geeksforgeeks.org/artificial-neural-networks-and-its-applications/

URL-17 “CONFUSION MATRIX”. (2019, April 29). Retrieved from blogspot: https://manisha-sirsat.blogspot.com/2019/04/confusion-matrix.html,

URL-18 Narkhede, S. (2018, June 26). “Understanding AUC - ROC Curve". Retrieved from towardsdatascience : https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5\

URL-19 Choudhary, I. (2023, August 30). “All You Need to Know About SHAP for Explainable AI?”. Retrieved from medium: https://medium.com/@shahooda637/all-you-need-to-know-about-shap-for-explainable-ai-8ad35a05e6ec

URL-20 “How to Use SHAP Values to Optimize and Debug ML Models”. (2023, August 28). Retrieved from Neptune: https://neptune.ai/blog/shap-values

RESUME

Name Surname **:** Ghina Özdemir

EDUCATION **:**

- **Bachelor** **:** 2022, Istanbul Kultur University, Engineering, Computer Engineering
- **M.A** **:** 2024, Istanbul Aydin University, Engineering, Computer Engineering

PUBLICATIONS FROM DISSERTATION, PRESENTATIONS AND PATENTS:

- Özdemir, G. (2024). EXPLAINABLE AI IN HEART DISEASE PREDICTION. IRJMETS, 6(4). doi:https://www.doi.org/10.56726/IRJMETS54626
- Wasim Raed, M., Huseyinov, I., Ozdemir, G., Kotenko, I., & Fedorchenko, E. (2023, October). An IoT-Based Smart Home for Elderly Suffering from Dementia. In The Proceedings of the International Conference on Smart City Applications (pp. 362-371). Cham: Springer Nature Switzerland.
- Alrefaai, S., Özdemir, G., & Mohamed, A. (2022, June). Detecting Phishing Websites Using Machine Learning. In 2022 International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA) (pp. 1-6). IEEE.