# Fraudulent Healthcare Providers detection using Machine Learning Algorithms

MSc Research Project
Data Analytics

Aniket Jambukar
Student ID: x20185014

School of Computing
National College of Ireland

Supervisor: Noel Cosgrave

**National College of Ireland**
**Project Submission Sheet**
**School of Computing**

| | |
|---|---|
| **Student Name:** | Aniket Jambukar |
| **Student ID:** | x20185014 |
| **Programme:** | Data Analytics |
| **Year:** | 2021 |
| **Module:** | MSc Research Project |
| **Supervisor:** | Noel Cosgrave |
| **Submission Due Date:** | 16/12/2021 |
| **Project Title:** | Fraudulent Healthcare Providers detection using Machine Learning Algorithms |
| **Word Count:** | 6714 |
| **Page Count:** | 20 |

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

**ALL** internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

| | |
|---|---|
| **Signature:** | |
| **Date:** | 16th December 2021 |

**PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:**

| | |
|---|---|
| Attach a completed copy of this sheet to each project (including multiple copies). | □ |
| **Attach a Moodle submission receipt of the online project submission**, to each project (including multiple copies). | □ |
| **You must ensure that you retain a HARD COPY of the project**, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer. | □ |

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

| **Office Use Only** | |
|---|---|
| Signature: | |
| Date: | |
| Penalty Applied (if applicable): | |

# Fraudulent Healthcare Providers detection using Machine Learning Algorithms

Aniket Jambukar
x20185014

## Abstract

Healthcare services are one of the basic needs of everyone in society. Frauds in healthcare not only lose the integrity of the services but also impacts everyone financially as there is a rise in insurance premiums and healthcare expenditures. NHCAA states in 2018 itself in the USA up to 10% of total healthcare expenditure was reported as fraud and the loss was estimated to be $300 billion. Upcoding in procedures and providing unnecessary services also impacts the medical history of individuals in fraud committed by healthcare providers. This research project aims to detect fraudulent healthcare providers using machine learning algorithms. A statistical approach of MANOVA and ANOVA f-test were carried to select the best features for model building. To handle highly imbalanced healthcare datasets hybrid sampling SMOTETomek technique was used. Random forest, SVM, XGBoost, and LogisticGAM supervised machine learning models were implemented, as evaluation matrix accuracy and class 1 recall were considered. From the comparison and evaluation of four models built, LogisticGAM had the highest fraud class 1 recall of 88%.

# 1 Introduction

Healthcare is one of the basic needs that every human seeks for living a healthy life. Healthcare providers have gained a lot of respect from society for all medication facilities they provide to keep treating the sick by dealing with illness, diseases, challenges in the facilities. With advancements in technology and researches, healthcare providers have been able to provide with best treatments to cure diseases. The Healthcare industry is one of the highest revenue-generating industries as well as also receives a maximum part of the government's budgets. The government's focus on the healthcare sector as it provides one of the basic needs for all, hefty researches are funded by the governments benefiting everyone to a healthy lifestyle. While healthcare providers' mission is to provide the best medical treatments to everyone, few dishonest providers tend to commit fraud and make profits affecting everyone in many ways and also damaging the integrity of healthcare systems (NHCAA; 2021). Billions of dollars are lost every financial year due to healthcare frauds losing their integrity and impacting the society and economy. The expenditure made on healthcare was $3.6 trillion in the USA alone in 2018 where billions of claims were made for health insurance. The loss was placed up to 10% by enforcement agencies and governments which could be $300 billion and more than that (NHCAA; 2021) (Copeland et al.; 2012).

This insurance fraudulent claims in healthcare are a chunk from billions of claims but they come with heavy price impacting financially and values of healthcare systems. Healthcare fraud is an organized crime and can be charged up to 10 years of imprisonment in the USA while if there is any form of injury to patients that could be doubled to imprisonment of 20 years and if it results in death could be sentenced to imprisonment for life (NHCAA; 2021). Common practices in healthcare provider fraud are billing of unnecessary expensive services, providing more services not required in standard treatments, conducting unnecessary medical tests, changes in medical statements to cover more medical services (Thornton et al.; 2015). These unnecessary medical tests and services could also result in serious health issues as well as impact the medical history of an individual making complications in future medications and treatments. Insurance premiums are increasing every year as insurance companies face huge losses due to fraudulent claims rising expenses of healthcare services (Wilson; 2009).

## 1.1 Background and Motivation

A large amount of data is generated by the healthcare sector from various sources like medical history, treatments records, insurance claims, financial details of healthcare providers, etc. The use of this data from various sources results in creating big data and data warehouses that are widely used for statistical methods to analyze data and build efficient models for fraud detections (Copeland et al.; 2012). Auditing strategy is a time-consuming process as claim reviewing experts review claims for every case from the samples of cases from different types of claims. This strategy is not efficient to identify fraud claims from millions of cases from different sources (Seo and Mendelevitch; 2017).

Correlation and regression analytical methods are used in statistical strategies for fraud identification make it very efficient than the auditing strategy of reviewing individual claims. In supervised methods, experts label the claims data in fraud and non-fraud that is used to process through algorithms in data mining. Imbalance of legit and fraud claims in any claims datasets causes the supervised algorithms to misclassify the fraud claims making it more challenging to build efficient supervised algorithms. Even with data challenges, supervised algorithms like Bayesian networks, Neural networks, Fuzzy logic, and Decision trees are widely used in fraud detection (Copeland et al.; 2012). Supervised algorithms with less computational price provide efficient fraud detection for large-sized datasets (Dua and Bais; 2014).

## 1.2 Specification of Research Project

With the urge of big data in healthcare and to identify fraudulent healthcare providers that claimed fraud insurance, this research project aims to implement and evaluate different Supervised machine learning algorithms for fraud detection of healthcare providers. Data from different sources need to be gathered and clubbed to the healthcare provider's data to build a dataset for learning the provider fraud detection. Healthcare datasets for fraud detection are highly imbalanced datasets where non-fraudulent providers weigh more than fraudulent providers, this challenges the supervised machine learning algorithms on their learning part of fraud detection. To handle imbalances datasets hybrid sampling techniques used can balance the majority and minority classes for balancing datasets to train machine learning models.

### 1.2.1 Research Question

*"How precisely Supervised machine learning algorithms Random Forest, SVM, XGBoost, and LogisticGAM can identify fraudulent healthcare providers from insurance claims data?"*

## 1.3 Objectives of Research Project

The objective of the research project is to build and evaluate supervised machine learning models to identify fraudulent healthcare providers. The datasets for this research project were referred from Kaggle [1] that are distributed in four datasets. Every dataset to be pre-processed and new features created would add more value to the datasets to build precise machine learning models to answer the research question. Healthcare datasets used to identify fraud are highly biased the important objective is also to handle the bias data using sampling techniques. To identify fraudulent healthcare providers four machine learning models were built and evaluated.

## 1.4 Contributions of Research Project

For highly biased healthcare datasets identifying fraud is challenging and from various sampling techniques, this research project used the hybrid SMOTETomek sampling technique. A major contribution from this research to identify fraudulent healthcare providers was a novel generalized additive model (GAM) for the classification, their functions of flexible predictors that discover the underlying patterns that are hidden in datasets, and functions of predictors regularization that avoids models from overfitting.

Further, the report is divided into 6 sections as follows: section 2 is the related work of fraud detection in the insurance domain, section 3 is about scientific methodology followed for the project, section 4 is about the design specifications of the project, section 5 is about the implementation details of the project, section 6 is about the evaluation of models built, section 7 is the conclusion and future work of research project.

# 2 Related Work

## 2.1 Introduction

Fraud by healthcare providers has impacted the integrity of healthcare services as well the expenditure and cost of insurances as well as overall expenses in the healthcare sector (NHCAA; 2021). In addition to financial losses, medical facilities are not received by the needy once where required while unnecessary treatments and services are provided to others to claim more insurance for the greed of profits. These types of frauds in healthcare are committed where many other bodies like physicians are involved apart from hospitals which makes it a more difficult problem to identify fraud (Rashidian et al.; 2012). The data that is generated by the insurance companies is large and where manual auditing procedures won't be an efficient method and an appropriate approach to detect healthcare provider's fraud. Machine learning and big-data techniques provide a more robust and efficient approach for the analysis of big data to identify patterns of frauds in large medical insurance datasets.

[1] https://www.kaggle.com/rohitrox/healthcare-provider-fraud-detection-analysis

Highly imbalanced datasets with minority class of fraud observations make it a challenging part to build efficient models to predict potential fraud healthcare providers (Dua and Bais; 2014). Various data mining techniques have been implemented to detect fraud in different financial sectors like banks, insurance, Securities and commodities, and other fields related to finance. Different approaches like classification, clustering, outlier detection, predictions, regression, and visualizations have been used in applications to identify as well as predict frauds (Ngai et al.; 2011). Techniques used to handle imbalanced datasets are described in the below subsection 2.2 and supervised machine learning models used to detect fraud are described in subsection 2.3.

## 2.2 Sampling Techniques for Imbalanced Healthcare Datasets

Healthcare data for fraud prediction comes with a high-class imbalance where fraud labels are very few compared to non-fraud. Machine learning models trained on these high imbalanced datasets and due to fewer observations of fraud their predictions are biased towards non-fraud and tend to mislead. The learning rate of fraud classes is affected due to an imbalance in classes. To balance these classes several sampling techniques are used where majority and minority classes are balanced to improve the performance of models. The Random undersampling technique randomly reduces the majority class discarding them till the balance with the minority class is reached. This reduces the overall required computational resources that make analysis on the large datasets easy and manageable. Random undersampling discarding the majority class randomly may lose important information from the dataset that can also affect the learning rate of models (Bauder et al.; 2018) (Mishra; 2017) (More; 2016).

The Random Oversampling technique increases minority classes by randomly adding multiple instances from available data. This increases the minority class size but by increasing the duplicate samples from available instances. With the increase in data of minority classes the required computational resources would also increase. An increase in the size of the dataset can also be an issue for the learner with the size of the dataset increased. As data is duplicated by randomly adding more instances from the training dataset, it can affect more to the highly skewed dataset. In Random Oversampling when instances are duplicated to increase the size of minority class and get a balanced availability of classes, it can tend models to overfit with the minority classes (Bauder et al.; 2018) (Bauder and Khoshgoftaar; 2018a).

Synthetic Minority Oversampling Technique (SMOTE) with k-nearest neighbors and replacement artificially generates instances to increase the minority class size. The difference is calculated between the considered sample and its k-nearest neighbors and then multiplied with a number that is randomly selected between 0 and 1 and finally, this generated vector is added to the instance considered. Instead of adding duplicated instances like random oversampling, SMOTE artificially creates instances from observations of present instances. As this technique creates new instances artificially and does not add duplicates compared to random oversampling, it is less prone to overfitting. Borderline-SMOTE is a modified approach where it performs SMOTE on instances of the minority class that are in the border of the decision region of the minority class (Bauder et al.; 2018) (Mishra; 2017)(Bauder and Khoshgoftaar; 2018a). SMOTETomek is a hybrid sampling technique that increases instances in the minority class using SMOTE as well as reduces the number of instances in the majority class. To decrease the number of instances in the majority class it uses Tomek undersampling method(Alam et al.; 2020).

Tomek identifies the samples with the nearest pair but from categories that are different. Samples that are nearest neighbors but from different categories, either one of them is noise or possibilities of both samples are on the boundary of the different classes. After removing the Tomek link that identified samples with the nearest neighbor from different classes, the samples with the nearest neighbor can now belong to the same class to be classified better (Zheng et al.; 2021).

## 2.3 Supervised Machine Learning Models in Fraud Detection

### 2.3.1 Random Forest Classifier

Random Forest has good abilities for classification and has been superior to other classifiers on balanced as well as imbalanced datasets of various fraud identification problems. It is an ensemble type of method where classification is made from multiple numbers of unpruned decision trees built and finally combining the results of the trees (Xuan et al.; 2018). Random forest algorithm creates several datasets randomly by using sampling for decision trees to be trained. Entropy is a measure of features that are uncertain while information gain finds the more informative features. By use of entropy as well as information gain at every node in the tree Random forest selects the best features from the classes. With feature selections, the Random forest emphasizes more on information gain and less entropy (Bauder and Khoshgoftaar; 2017) (Bauder and Khoshgoftaar; 2018b).

Random forest is resistant to overfit and also found to give a better estimation for the generalization error, as every tree is independently trained, for huge datasets with more number of features Random forest still training is fast (Campus; 2018). With its improved variant from mechanisms of the Decision tree, the advantages of accuracy and higher dimensionality make Random forest a better algorithm for classification even on highly imbalanced datasets to identify fraud and anomalies. The Random forest generally is a better learner when it comes to fraudulent detection in medicare datasets (Bauder and Khoshgoftaar; 2017). Credit card fraud detection dataset with high imbalance where fraud labels are only 0.173%, on this imbalanced datasets performance of models is highly affected by sampling techniques, features selection methods and the algorithms used to detect fraud. With an imbalanced dataset, Random forest performed better compared to other models and could achieve an accuracy of 98.6% (Campus; 2018).

### 2.3.2 Support Vector Machines

SVM is one of the robust algorithms for classification where it constructs a hyperplane that separates the data by maximizing the datasets margin. The training set is classified which is done by grouping the training sets into categories. While SVM functions linear classification it can also function non-linear classification by using kernel trick. Every support vector creates subsets of the data provided for training the classification model and the data points that are beyond this are eliminated for classification (Bauder and Khoshgoftaar; 2016). To detect fraud in health insurance claims a hybrid approach combining Evolving Clustering Method and SVM was proposed as SVM has eased for training and quality for generalization (Rawte and Anuradha; 2015). For the class imbalanced problem in automobile insurance claims fraud, ADASYN was employed and SVM obtained the highest detection of fraud rate on both balanced and unbalanced datasets. On a balanced dataset, the sensitivity achieved for SVM was 94.52% (Subudhi and Panigrahi; 2018).

### 2.3.3 XGBoost

XGBoost is an extreme gradient boosting algorithm where every tree that grows from the previous tree learns from the errors of the previous tree making it more robust in classifying more precisely. The combination of predictions from several multiple learners increases its learning power giving more accurate classification. It is also well known for its high performance where its decreased execution time compared to other machine learning algorithms (Akbar et al.; 2020). XGBoost when applied on a highly imbalanced dataset of auto insurance claims to detect fraudulent claims achieved an accuracy of 99.25% with a recall core of 0.992 where the Decision tree model could achieve an accuracy of 92.99% and recall score of 0.929 (Dhieb et al.; 2020).

As insurance datasets are large and require high computational speeds to detect fraudulent claims, a comparative study of performance between XGBoost and CatBoost was carried on a large medicare insurance dataset (Hancock and Khoshgoftaar; 2020). On this highly imbalanced dataset, the performance of both methods was measured on basis of AUC and running time. When data or Medicare Part B dataset is aggregated there is no significant difference statistically in the AUC score of both the GBDT's, but when no data is aggregated Catboost's mean AUC score is 0.9080 while XGBoost the mean AUC score is 0.8616. For the Medicare Part B dataset for fraud detection, 250 trees are used in their ensembles which was part of the research and this contribution was important as the number of trees tends to resource consumption in the GBDT (Hancock and Khoshgoftaar; 2020). While handling an imbalance dataset to detect fraudulent healthcare providers this paper implies the use of undersampling technique and study experimented Random forest and XGBoost. The accuracy of both models was the same of 85% but the class 1 recall of XGBoost was 4% more significantly of 85% (Akbar et al.; 2020). This research project refers to the same dataset of Healthcare Provider fraud detection that was used in the research published by (Akbar et al.; 2020).

### 2.3.4 Generalized Additive Models

Günther et al. (2014) experimented with customer churn prediction on a dataset of an insurance company using the GAM approach. GAM approach can be used to get the realistic description for independent and dependent variables relation. Chang et al. (2021) researched on simulated datasets as well as real datasets were used to investigate quantitatively and qualitatively different GAM algorithms. The findings were that when very few features are used to make predictions GAM can be unfair to the minority subpopulations as it can miss all the patterns from the data. The results from experiments suggest that tree-based GAMs have the best balance of accuracy, sparsity, as well as fidelity, and hence they are the most trustworthy GAM models.

## 2.4 Conclusion

From sampling techniques, SMOTETomek is a hybrid sampling technique that increases instances of minority class using SMOTE and reduces instances of majority classes using Tomek undersampling which gives more balanced datasets that makes it more suitable for highly imbalanced datasets of healthcare fraud. To identify fraudulent healthcare providers LogisticGAM for classification would be a more suitable approach as they give more underlying patterns to identify relations for independent and dependent variables that can get better predictions in fraud class.

# 3 Methodology

## 3.1 Fraudulent Healthcare Providers detection Methodology

To detect fraudulent healthcare providers using supervised machine learning algorithms this research project follows the KDD-based methodology to discover machine learning models that can precisely detect fraud healthcare providers. The methodology followed in the fraudulent healthcare providers detection research project is shown in below figure 1 and was refereed from (Luo; 2008). This scientific methodology has five important parts of collecting raw data from sources to facilitate analysis, data pre-processing and transformation for analyzing raw data, data mining in which applying machine learning models on data to detect frauds, and evaluating the results of models to derive knowledge from predictions of models. Details of each block in methodology are described in detail in the below sub-sections.

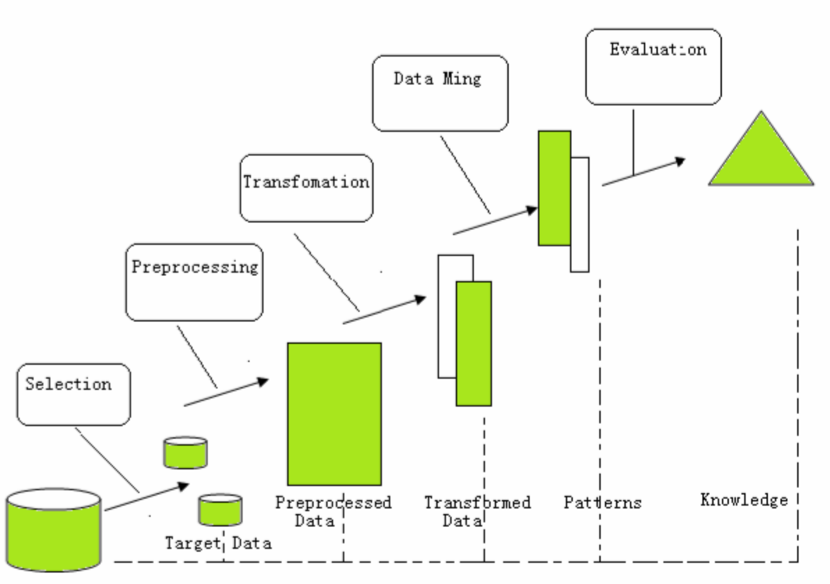


Figure 1: Knowledge Discovery in Database (KDD) Methodology

## 3.2 Data Selection

Insurance data are widely distributed as per the applications that make data collection more important to collect relevant data for analysis, predictions, and decision making. Data selection is the first step of the project where the datasets were referred from the Kaggle portal consisting of four csv training files. The first file is the training file that contains the provider ID and potential fraud label stating if the provider is fraud or not. The second file is the beneficiary file in which details of all beneficiaries like medical and personal details are available. The third file is the inpatient file in which details of claims as well as treatments given to patients admitted are available. The fourth file is the outpatient file in which details of claims as well as treatments given to patients visited are available. These four files need to be collected and further processed and merged to create a master dataset for research of detecting fraudulent providers.

## 3.3 Data Pre-processing and Transformation

Data pre-processing is the most important process for this research question, as data is distributed into different datasets. Merging datasets with correct attributes to achieve quality data that can be used for analysis is challenging. Detailed data needs to be aggregated and formed into a valuable dataset. This stage of data pre-processing is to clean, merge and structure the four raw datasets into a master dataset which can be used to analyze and build machine learning models. Raw datasets also contain null values that need to be handled and new features to be created to bring knowledge and values to the dataset. Once data is cleaned and new features are created four datasets have common attributes like beneficiary ID, claim ID, and provider ID that can be used to merge all datasets to provider ID transforming to a master dataset for analysis. Further EDA and statistical analysis for feature selection are carried to build precise machine learning models to detect fraud. A statistical approach for feature selection would lead to selecting the right features for model building and predictions. Imbalanced datasets in healthcare applications make the research more challenging. Selecting the right sampling technique to handle imbalanced datasets can avoid overfitting and better predictions.

## 3.4 Data Mining

Once the data is transformed as per the requirement next step comes data mining in which model building is a key component in the whole process. As the data is transformed can be used to seek the outlines of benefits. This phase varies based on the targeted output, in this research project of healthcare provider fraud detection requires the grouping of data and machine learning models building that results in fraud detection. Random forest classifies based on results combined of all trees from multiple decisions tree built with enhanced mechanisms of Decision trees makes it a good learner and a suitable algorithm for fraud classification. SVM for its generalized is suitable for fraud classifications, XGBoost learns from multiple learners and is more accurate in classifications for fraud applications. GAM in fraud classification problems is more suitable because they can learn of non-linear relationships between the dependent and independent variables to better predict.

## 3.5 Evaluation

Different evaluation parameters are used to evaluate the models used for predictions to define which model would bring more value to the solution of the problems. Efficient and precise models would bring more value as well as robust the process of predictions and solutions to the problems. Parameters accuracy and class 1 recall are used to evaluate the performance of the models in this research project.

## 3.6 Knowledge

From four supervised machine learning models built the evaluation using accuracy and class 1 recall would result in which model would precisely identify detect fraudulent healthcare providers. A novel approach of the SMOTETomek sampling technique for imbalanced datasets, and evaluations of the LogisticGAM, and three other machine learning models would contribute to the research question.

# 4 Design Specifications

The dataset used in this research project consists of four datasets that need to be cleaned, merged, and structured to facilitate analysis and model building. The architectural design for this research project is as below figure 2.

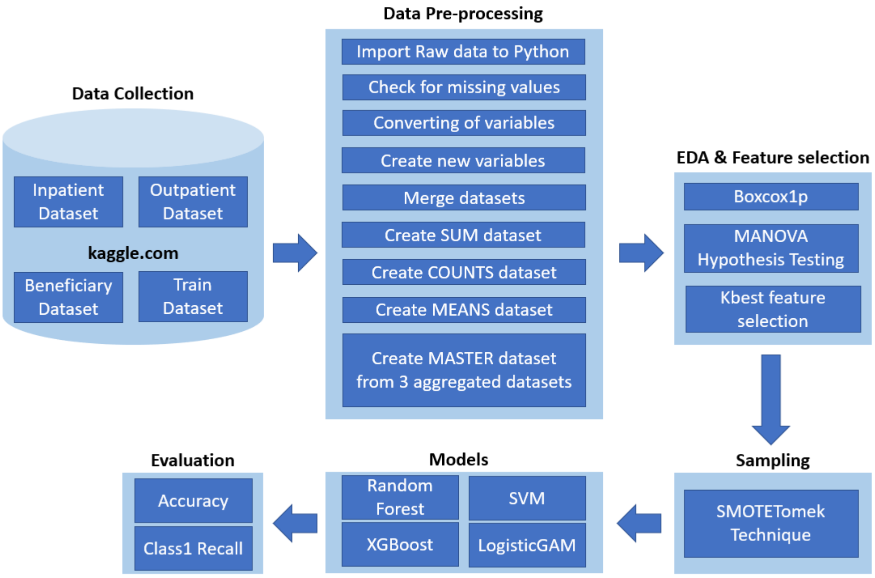


Figure 2: Architectural Design

## 4.1 Data collection and Pre-processing

The data collected for detecting fraudulent healthcare providers from Kaggle consists of four datasets. Each dataset needs to be checked for null values and needs to be cleaned to get a more valuable dataset. The potential fraud label for providers in the training file has to be merged with all details from the other three datasets to create a single master dataset. The first file with beneficiary details contains details of all patients about the disease indicators as well as personal details of the date of birth, date of death, gender, race, and more which has 138,556 beneficiaries and 25 features. If a patient was admitted for medical treatments there claims data are present in the inpatient dataset that has 40474 claims with 30 features. Patients who only visited for treatments and were not admitted are present in the outpatient dataset that has 517737 claims with 27 features. Beneficiary details need to be merged into every detail of claims in the inpatient and outpatient dataset. This dataset needs to be aggregated and merged to provider labels where 5410 distinct providers are available.

The architecture of all four datasets is as below figure 3. The first stage of design would be to merge the inpatient and outpatient dataset into a single dataset that will contain all the insurance claims of all visits to the healthcare providers. Once the dataset of claims is formed details of beneficiaries would be merged to them that will contain details of all beneficiaries and the claims made. This dataset will have multiple records of claims that were made by the provider as the number of beneficiaries have visited the provider. The training data has a distinct provider ID and label of potential fraud i.e only one record for every provider and if the provider is potential fraud.

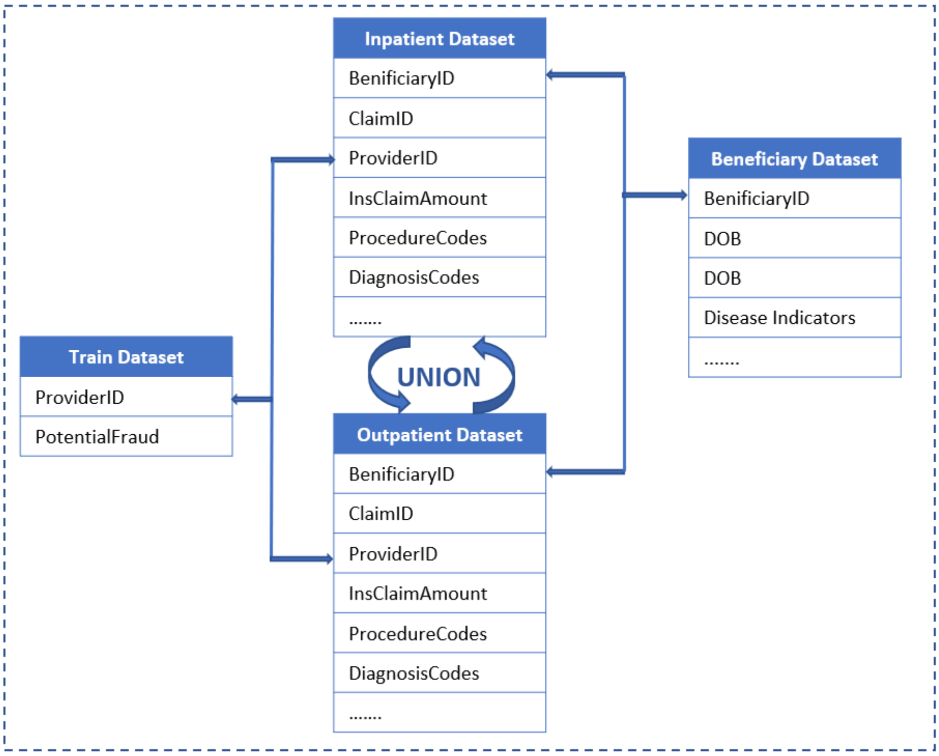


Figure 3: Data architecture

## 4.2 Feature engineering, EDA, and Feature selection

As four datasets are merged into single dataset new features are to be created that can add more value to the dataset. Features would be created a few while data cleaning and pre-processing while few after merging into a single dataset. EDA to check all details in the dataset that will help in selecting the best features. To conduct hypothesis testing MANOVA was carried out as the dataset has thirty continuous variables and one target variable with two classes of potential fraud and not-fraud. A correlation test was carried to check if features are correlated with other features. To select the best features Kbest features selection method was used.

## 4.3 Sampling Techniques

Referred dataset for this research project is highly imbalanced where only 9.35% of potential fraud are available in the dataset. Sampling techniques are used either to increase instances in minority class or reduce instances from majority class to balance the dataset. This balancing of classes using sampling techniques would reduce the biases of the models and also avoid models from overfitting. The undersampling technique would only reduce the instances from the majority class that can also delete important observations reducing the quality of data while oversampling could add noise to data. In research of (Akbar et al.; 2020) implemented undersampling on this dataset but this research proposes the use of SMOTETomek a hybrid sampling technique that will increase the instances in the minority class using SMOTE and reduce the instances in the majority class using Tomek undersampling that will balance the dataset more efficiently.

### 4.4 Model Building and Evaluation

The Random forest for this research would be a better learner for the imbalanced class 1 fraud healthcare providers. As Random forest, it is an ensemble method that demonstrates to make a classification based on the combination of multiple tree results that makes it a better classifier and results in better predictions on balanced and imbalanced datasets(Bauder and Khoshgoftaar; 2017). Support vector machines and XGBoost have proved to have high accuracy and sensitivity with a better ability to classify fraud detection problems and were selected for fraud classification in the research project(Subudhi and Panigrahi; 2018)(Akbar et al.; 2020). As GAM learns from linear as well as non-linear relationships, LogisticGAM would make better predictionsGünther et al. (2014). This research focuses more on identifying fraudulent healthcare providers, to evaluate the four machine learning algorithms metrics used are accuracy and class 1 recall that is the recall of fraud class healthcare providers.

## 5 Implementation

### 5.1 Data Pre-processing and Feature Engineering

Beneficiary dataset with all details of beneficiaries of disease and personal details were processed to create new features. With features date of birth and date of death age of beneficiaries were calculated, for beneficiaries that date of death is not available maximum date of death from the dataset was considered. All features with labels were converted with values 1 and 0 for yes and no respectively. The inpatient dataset has details of all patients that were admitted for medication. Null values in all physicians columns were replaced with characters missing and the total number of physicians attended for every claim was calculated. Deductible amount with null values was replaced with 0 as other records have only one amount of 1068. The total number of days the patient was admitted was calculated. The outpatient dataset with null values of the physician was replaced with characters missing and the total number of physicians attended for every claim was calculated. Inpatient and outpatients datasets were merged into a single dataset with all common columns and other columns to form a union of two datasets using the outer function, this dataset has 558211 records with 32 features. The beneficiary dataset was merged with Beneficiary ID creating a new dataset with 558211 records and 58 features.

From the above dataset merged with inpatient, outpatient, and beneficiary datasets Claim procedure code 6 was dropped as the entire column is null. The total number of unique diagnosis codes and procedures carried out and claimed in the claims were calculated. Features for attending physicians, operating physicians, other physicians, and diagnosis group code present were created. New features diagnosis group code was created and data merged all together in a dataset consisting of 558211 records and 64 features. The dataset created formed by merging three detailed datasets needs to be aggregated to providers ID so that it can be merged to every provider with the label of potential fraud.

The total count of beneficiaries that visited providers and the total number of claims made by the provider were calculated for all providers. The number of months in parts A and B coverage and age of patients was aggregated to every provider by taking the mean of these features. The total sum for every provider's amounts of insurance claimed was summed to providers ID that gives details of the amounts that were reimbursed to every provider. The total number of physicians from the entire claims data for every provider was calculated by summing the count of physicians. From claims data, the total number of diseases of all patients treated, diagnoses carried and procedures followed were summed up to every provider. Three aggregated datasets one with counts of beneficiary and claims counts, second with the sum of all amounts, physician and disease indicators, and third with mean of part months coverage and age were merged to the provider's ID. Three datasets that are merged to provider labels datasets using inner function results to a final dataset with 5410 records with 34 features with details of every provider of amounts claimed, beneficiary and claims count, the number of procedures and diagnosis carried, physicians, and a label of the provider being potential fraud. Feature with providers ID that has unique IDs of all providers was dropped. This aggregated final dataset is further used for EDA and feature selection.

## 5.2 EDA and Feature selection

Aggregated and cleaned dataset formed after data pre-processing was analyzed to get the importance of features in the dataset from EDA. All 32 features in the dataset are continuous variables with only one only target variable of potential fraud is categorical. Below figure 4 shows the individual barplot of potential fraud vs Inpatient and Outpatient Annual Deductible and Reimbursement Amounts which shows that there is a Potential Fraud when the amounts are more as there mean is more. Same plots were also plotted for Beneficiary ID count, Claim ID counts, Total physicians count, and Insurance Claim amount reimbursed where it showed that there there is a Potential Fraud when the amounts and count are more as there mean is more. We can conclude that there is potential fraud when there are multiple beneficiaries with more claims made and even reimbursement amounts are more.

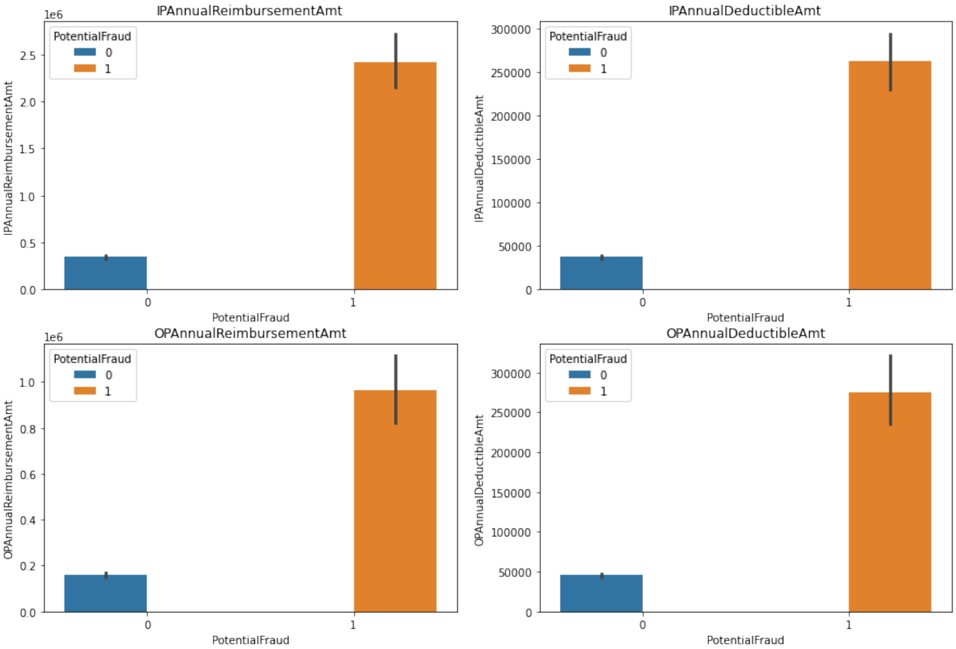


Figure 4: Inpatient and Outpatient Annual Deductible and Reimbursement Amounts vs Potential Fraud Analysis

The below figure 5 shows the count plot for part A and B insurance coverage in months where maximum counts of coverage are for 12 months. As maximum claims made by providers have 12 months coverage and these are maximumly biased from all values in the feature they cannot be used to train models.

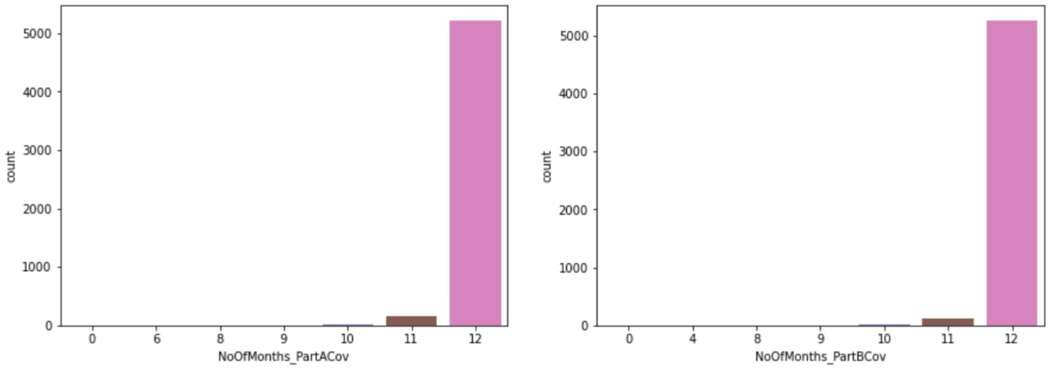


Figure 5: Count Plot of Part A and B Coverage

All features in the dataset are highly left-skewed and need to be normally distributed. To distribute all features normally boxcox1p transformation technique was used. Many features in the dataset have 0 values because of which they cannot be transformed using the log transformation method. Boxcox1p transformation adds 1 constant to all values and then transforms the data. Yeo-Johnson transformer technique was used as it distributes the data symmetrically. All features were transformed and normally distributed, the below figure 6 shows an example of the distribution of feature Insurance claim amount reimbursed that is left-skewed and its normal distribution after boxcox1p transformation.

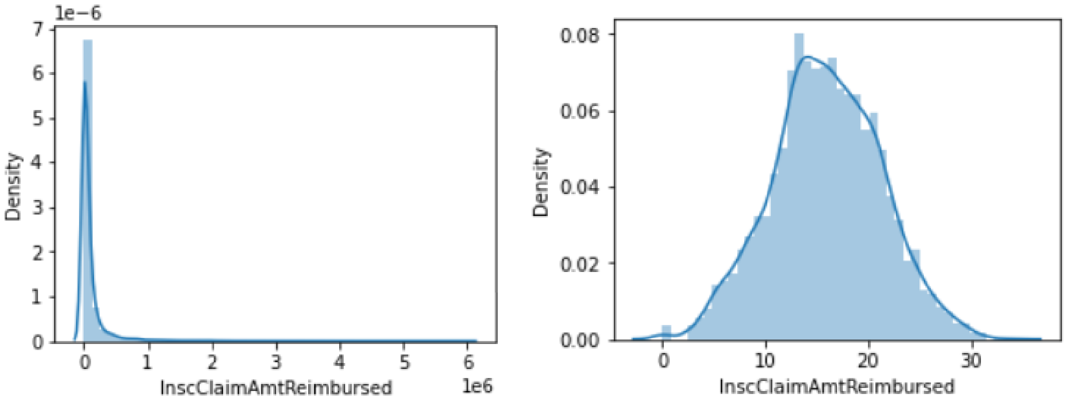
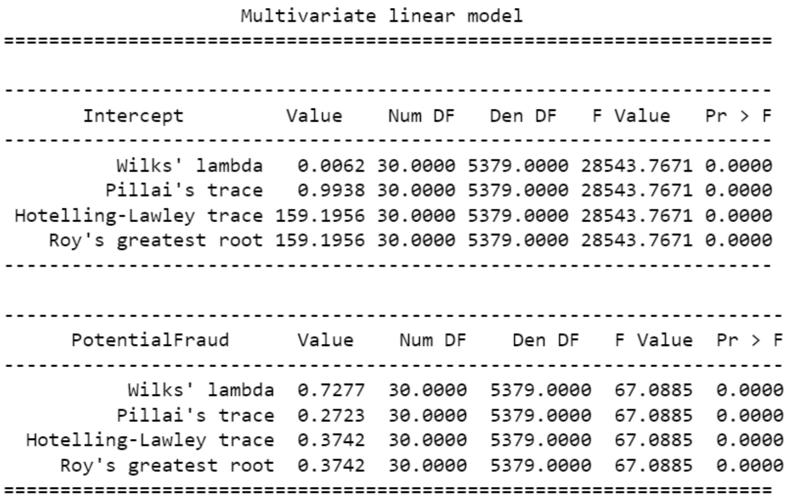


Figure 6: Left skewed and bocox1p transformed normal distribution of Insurance claim amount reimbursed

To check significance in the group differences for all variables in the dataset MANOVA hypothesis testing was carried out. The null hypothesis says that there is no significant difference between combined variables with potential fraud, and the alternate hypothesis says there is a significant difference between combined variables with potential fraud. From below figure 7 of MANOVA test Wilks'lambda we see that the p-value is less than 0.05 so we reject the null hypothesis.



```
                    Multivariate linear model
==================================================================
                                                                  
------------------------------------------------------------------
       Intercept         Value    Num DF    Den DF    F Value   Pr > F
------------------------------------------------------------------
          Wilks' lambda   0.0062 30.0000 5379.0000 28543.7671 0.0000
         Pillai's trace   0.9938 30.0000 5379.0000 28543.7671 0.0000
 Hotelling-Lawley trace 159.1956 30.0000 5379.0000 28543.7671 0.0000
    Roy's greatest root 159.1956 30.0000 5379.0000 28543.7671 0.0000
------------------------------------------------------------------
                                                                  
-------------------------------------------------------------------
      PotentialFraud      Value     Num DF     Den DF    F Value  Pr > F
-------------------------------------------------------------------
           Wilks' lambda  0.7277  30.0000  5379.0000  67.0885  0.0000
          Pillai's trace  0.2723  30.0000  5379.0000  67.0885  0.0000
  Hotelling-Lawley trace  0.3742  30.0000  5379.0000  67.0885  0.0000
     Roy's greatest root  0.3742  30.0000  5379.0000  67.0885  0.0000
==================================================================
```

Figure 7: MANOVA Test

There is a significant difference between the group of all variables with potential fraud, now to check if there is a significant difference of every variable with potential fraud univariate analysis was carried out using ANOVA. Post hoc analysis was carried out for every variable, from all variables age variable p-value was less than alpha and it tells that there is no significant difference of it with the potential fraud class. The age variable was dropped as it does not have a significant difference and would not contribute to the model.

As all the dependent features are numerical and the independent target variable is categorical to select the best features ANOVA f-test is used where features that are independent of the categorical target variable will be removed. Figure 8 shows the plot of feature scores from the ANOVA f-test where eight features with the highest f-scores were selected to build models. The final dataset contained 5410 records with 9 features.

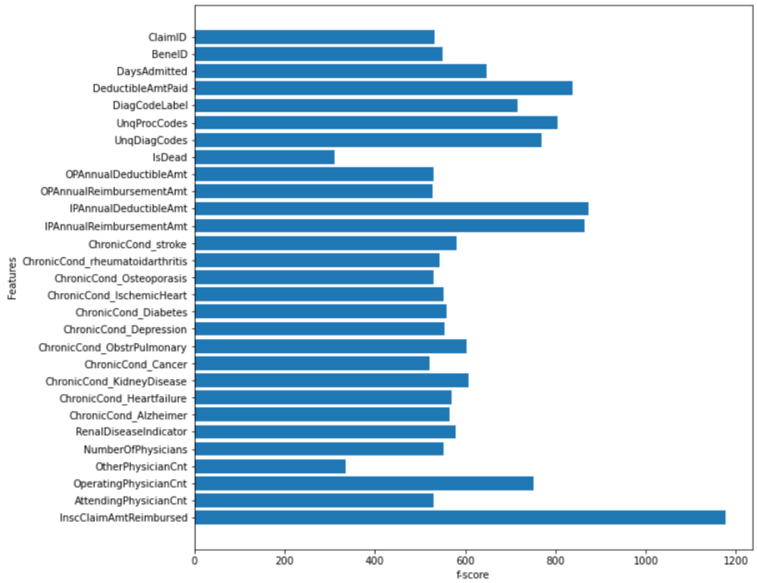


Figure 8: ANOVA f-test scores plot of features

## 5.3 SMOTETomek Sampling Technique

The dataset was split into 70% train and 30% test in which amongst 3487 providers non-fraud class has 3426 data while fraud class has only 361 data in the training dataset. This class imbalance will be prone to bias the model predictions to non-fraud. To handle this imbalance SMOTETomek hybrid technique was used in which the very few instances from the majority class of non-fraud were removed and new instances were added to minority class fraud. After applying the sampling technique class non-fraud has 3378 instances and the same number of instances in the fraud class. This balanced dataset improved the accuracy of the models, as well as the learning rate, was better than the imbalanced dataset where more recall score was achieved.

## 5.4 Models Implementation

Four models Random Forest, Support Vector Machines (SVM), XGBoost, and LogisticGAM were implemented in this research project. The data was split into the ratio of 70:30. In highly imbalanced datasets for this research question achieving more recall scores to precisely detect the fraud and non-fraud providers and compared to previous researches to achieve more accuracy and recall score has experimented. GAM algorithms predicting functions can be regularized that avoids overfitting while predicting functions can also underlying patterns hidden in data while they are also interpretable. These three features in GAM make it more robust for binary classification. For evaluation, the matrix selected is accuracy and class 1 recall.

# 6 Evaluation

The focus of this research project is to identify how precisely can supervised machine learning algorithms can detect potential fraudulent healthcare providers so with accuracy, class 1 recall is the important evaluation matrix to evaluate the implemented machine learning models.

## 6.1 Random Forest Model Evaluation

The evaluations of the Random forest model from the classification report of the model are shown in table 1. The Random forest model had an accuracy of 91% and class 1 recall of 72%.

Table 1: Random forest model Classification report

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.97 | 0.92 | 0.95 | 1478 |
| 1 | 0.48 | **0.72** | 0.58 | 145 |
| accuracy | | | **0.91** | 1623 |
| macro avg | 0.73 | 0.82 | 0.76 | 1623 |
| weighted avg | 0.93 | 0.91 | 0.91 | 1623 |

## 6.2 Support Vector Machines Model Evaluation

The evaluations of the SVM model from the classification report of model are shown in table 2. The SVM model had an accuracy of 87% and class 1 recall of 77%.

Table 2: SVM model Classification report

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.98 | 0.88 | 0.92 | 1478 |
| 1 | 0.39 | **0.77** | 0.51 | 145 |
| accuracy | | | **0.87** | 1623 |
| macro avg | 0.68 | 0.83 | 0.72 | 1623 |
| weighted avg | 0.92 | 0.87 | 0.89 | 1623 |

## 6.3 XGBoost Model Evaluation

The evaluations of the XGBoost model from the classification report of model are shown in table 3. The XGBoost model achieved an accuracy of 87% and class 1 recall of 86%. The class 1 recall of the XGBoost model was better than the SVM model, it could identify fraudulent providers more precisely.

Table 3: XGBoost model Classification report

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.98 | 0.88 | 0.93 | 1478 |
| 1 | 0.40 | **0.86** | 0.55 | 145 |
| accuracy | | | **0.87** | 1623 |
| macro avg | 0.69 | 0.87 | 0.74 | 1623 |
| weighted avg | 0.93 | 0.87 | 0.89 | 1623 |

## 6.4 LogisticGAM Model Evaluation

The evaluations of the LogisticGAM model from the classification report of model are shown in table 4. The LogisticGAM model had an accuracy of 87% and class 1 recall of 88%. The class 1 recall of the LogisticGAM model was more than above three models.

Table 4: LogisticGAM model Classification report

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.99 | 0.87 | 0.92 | 1478 |
| 1 | 0.40 | **0.88** | 0.55 | 145 |
| accuracy | | | **0.87** | 1623 |
| macro avg | 0.69 | 0.87 | 0.73 | 1623 |
| weighted avg | 0.93 | 0.87 | 0.89 | 1623 |

## 6.5 Comparative Evaluation of Models

The below table 5 shows the comparative evaluation of machine learning models built with their respective accuracy and class 1 recall. Random forest had high accuracy but the class 1 recall was very less compared to other models. SVM and XGBoost had the same accuracy but the class 1 recall score of XGBoost was better than SVM. LogisticGAM had same accuracy as SVM and XGBoost but the class 1 recall was highest of 88%.

Table 5: Evaluations of Models

| **Model** | **Accuracy** | **Class 1 Recall** |
|---|---|---|
| Random Forest | 0.91 | 0.72 |
| SVM | 0.87 | 0.77 |
| XGBoost | 0.87 | 0.86 |
| LogisticGAM | 0.87 | **0.88** |

## 6.6 Discussion

Data pre-processing is important for healthcare datasets that are distributed and highly biased which makes it challenging to build clean and structured datasets. Aggregating and feature creation add more values to the datasets to build better-performing models. Biased datasets need appropriate sampling technique approaches as undersampling may lose important data and oversampling may add noise to the datasets. Four models were built Random Forest, SVM, XGBoost, and LogisticGAM to identify fraudulent healthcare providers. From these four models built the accuracy of three models SVM, XGBoost, and LogiticGAM was the same but their different class 1 recall scores would also vary in efficiently identifying the fraudulent providers. LogisticGAM had the highest class 1 recall score of 88% that from the models built and was an important contribution to the research question. Various approaches of aggregation of data, sampling techniques would vary the performance of models in identifying the fraudulent healthcare providers.

# 7 Conclusion and Future Work

The summary, the aim of the research project was achieved by answering the research question. To answer the research question four machine learning models were built and evaluated, from evaluations it can be stated that LogisticGAM which had the highest class 1 recall score can precisely identify fraudulent healthcare providers compared to other models built. The evaluations gained for each model depict that fraud predictions are effective with classification methods.Finance transactions details of healthcare providers if added to data could add more value and contribute more in solutions to the fraud problems. Fraud healthcare providers will commit fraud in new ways, identifying new patterns from insurance claims and updating the learning of models is important to identify fraud. Changes in insurance and claims policies would make concept and data drifts which need to be addressed to avoid degradation of performance of models.

# Acknowledgement

I would specially thank my supervisor Noel Cosgrave who supported and guided me through my research project. I would also like to thank my wife who supported me during my academics.

# References

Akbar, N. A., Sunyoto, A., Rudyanto Arief, M. and Caesarendra, W. (2020). Improvement of decision tree classifier accuracy for healthcare insurance fraud prediction by using extreme gradient boosting algorithm, *2020 International Conference on Informatics, Multimedia, Cyber and Information System (ICIMCIS)*, pp. 110–114.

Alam, T. M., Shaukat, K., Hameed, I. A., Luo, S., Sarwar, M. U., Shabbir, S., Li, J. and Khushi, M. (2020). An investigation of credit card default prediction in the imbalanced datasets, *IEEE Access* **8**: 201173–201198.

Bauder, R. A. and Khoshgoftaar, T. M. (2016). A novel method for fraudulent medicare claims detection from expected payment deviations (application paper), *2016 IEEE 17th International Conference on Information Reuse and Integration (IRI)*, pp. 11–19.

Bauder, R. A. and Khoshgoftaar, T. M. (2017). Medicare fraud detection using machine learning methods, *2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pp. 858–865.

Bauder, R. A. and Khoshgoftaar, T. M. (2018a). The effects of varying class distribution on learner behavior for medicare fraud detection with imbalanced big data, *Health information science and systems* **6**(1): 1–14.

Bauder, R. A., Khoshgoftaar, T. M. and Hasanin, T. (2018). Data sampling approaches with severely imbalanced big data for medicare fraud detection, *2018 IEEE 30th International Conference on Tools with Artificial Intelligence (ICTAI)*, pp. 137–142.

Bauder, R. and Khoshgoftaar, T. (2018b). Medicare fraud detection using random forest with class imbalanced big data, *2018 IEEE International Conference on Information Reuse and Integration (IRI)*, pp. 80–87.

Campus, K. (2018). Credit card fraud detection using machine learning models and collating machine learning models, *International Journal of Pure and Applied Mathematics* **118**(20): 825–838.

Chang, C.-H., Tan, S., Lengerich, B., Goldenberg, A. and Caruana, R. (2021). How interpretable and trustworthy are gams?, *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery amp; Data Mining*, KDD '21, Association for Computing Machinery, New York, NY, USA, p. 95–105.
**URL:** *https://doi.org/10.1145/3447548.3467453*

Copeland, L., Edberg, D., Panorska, A. K. and Wendel, J. (2012). Applying business intelligence concepts to medicaid claim fraud detection, *Journal of Information Systems Applied Research* **5**(1): 51.

Dhieb, N., Ghazzai, H., Besbes, H. and Massoud, Y. (2020). A secure ai-driven architecture for automated insurance systems: Fraud detection and risk measurement, *IEEE Access* **8**: 58546–58558.

Dua, P. and Bais, S. (2014). Supervised learning methods for fraud detection in healthcare insurance, *Machine learning in healthcare informatics*, Springer, pp. 261–285.

Günther, C.-C., Tvete, I. F., Aas, K., Sandnes, G. I. and Borgan, Ø. (2014). Modelling and predicting customer churn from an insurance company, *Scandinavian Actuarial Journal* **2014**(1): 58–71.

Hancock, J. and Khoshgoftaar, T. M. (2020). Performance of catboost and xgboost in medicare fraud detection, *2020 19th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pp. 572–579.

Luo, Q. (2008). Advancing knowledge discovery and data mining, *First International Workshop on Knowledge Discovery and Data Mining (WKDD 2008)*, pp. 3–5.

Mishra, S. (2017). Handling imbalanced data: Smote vs. random undersampling, *International Research Journal of Engineering and Technology (IRJET)* **4**(8).

More, A. (2016). Survey of resampling techniques for improving classification performance in unbalanced datasets, *arXiv preprint arXiv:1608.06048* .

Ngai, E., Hu, Y., Wong, Y., Chen, Y. and Sun, X. (2011). The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature, *Decision Support Systems* **50**(3): 559–569. On quantitative methods for detection of financial fraud.
**URL:** *https://www.sciencedirect.com/science/article/pii/S0167923610001302*

NHCAA (2021). The challenge of health care fraud. [Accessed on 04 Dec 2021].
**URL:** *https://www.nhcaa.org/tools-insights/about-health-care-fraud/the-challenge-of-health-care-fraud/*

Rashidian, A., Joudaki, H. and Vian, T. (2012). No evidence of the effect of the interventions to combat health care fraud and abuse: a systematic review of literature.

Rawte, V. and Anuradha, G. (2015). Fraud detection in health insurance using data mining techniques, *2015 International Conference on Communication, Information Computing Technology (ICCICT)*, pp. 1–5.

Seo, J. and Mendelevitch, O. (2017). Identifying frauds and anomalies in medicare-b dataset, *2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 3664–3667.

Subudhi, S. and Panigrahi, S. (2018). Effect of class imbalanceness in detecting automobile insurance fraud, *2018 2nd International Conference on Data Science and Business Analytics (ICDSBA)*, pp. 528–531.

Thornton, D., Brinkhuis, M., Amrit, C. and Aly, R. (2015). Categorizing and describing the types of fraud in healthcare, *Procedia Computer Science* **64**: 713–720.

Wilson, J. H. (2009). An analytical approach to detecting insurance fraud using logistic regression, *Journal of Finance and accountancy* **1**: 1.

Xuan, S., Liu, G., Li, Z., Zheng, L., Wang, S. and Jiang, C. (2018). Random forest for credit card fraud detection, *2018 IEEE 15th International Conference on Networking, Sensing and Control (ICNSC)*, pp. 1–6.

Zheng, H., Sherazi, S. W. A. and Lee, J. Y. (2021). A stacking ensemble prediction model for the occurrences of major adverse cardiovascular events in patients with acute coronary syndrome on imbalanced data, *IEEE Access* **9**: 113692–113704.