# INFS_SP5_2023
# Predictive Analytics
# Assignment 1

Enna Kris

**Huining Huang huahy057@mymail.unisa.edu.au**

**Lingjun Ji Jiyly006@mymail.unisa.edu.au**

## Part 1: Introduction

Healthcare insurance is crucial for citizens' access to quality healthcare and national well-being. But it's threatened by healthcare insurance fraud, which involves deceptive activities among medical providers, patients, and insurance companies, posing a significant challenge in the healthcare sector. Insurance companies are frequently on the receiving end of these bad practices, which in turn has caused them to hike the prices of their insurance premiums, making healthcare costs surge periodically.[1] It is pretty evident that patients and their healthcare information can easily be exploited which later can hamper the overall cost.[1] Healthcare fraud has far-reaching consequences. It jeopardizes the healthcare system's stability, erodes patient trust in insurance, and can lead to unnecessary expenses and health risks from unnecessary treatments. It also poses financial risks to insurance companies and damages the reputation of all healthcare providers.

In order to detect and avoid the fraud, data mining techniques are applied.[2] Data mining aids in uncovering fraud patterns and unusual behavior in extensive medical data, helping the healthcare system spot unnecessary procedures and ensure appropriate care for patients. Predictive models using historical data can also identify potential fraudulent providers, enabling early detection and prevention of medical insurance fraud. This leads to cost reduction for insurance companies and, in turn, lower insurance premiums.

In this paper, we will examine academic papers on healthcare technologies and algorithms to improve our approach and feature extraction strategies. We will also use data mining to analyze a detailed dataset for detecting medical fraud. This dataset consists of medical insurance claims, containing provider details, billing information, patient records, procedure data, and fraud indicators. Our goal is to extract meaningful features from this dataset and build a predictive model to identify potential medical fraud providers.

## Part 2: Related Work

This section will review three academic papers that analyze and address healthcare fraud prediction from different perspectives, providing valuable insights for the current task.

First, the paper titled "Healthcare Provider Summary Data for Fraud Classification"[3] by J. M. Johnson emphasizes the critical role of feature engineering and dataset construction in healthcare fraud detection. The author leverages the latest publicly available data from CMS and introduces two new labeled Medicare

Part B datasets for supervised learning. Their research demonstrates that, through careful selection and construction of the SbP feature set, significant improvements can be achieved in the performance of practical healthcare fraud detection models. This is of paramount importance to our project as we need to carefully consider how to design and select the most informative features to enhance the accuracy of our model.

Furthermore, the paper titled "A Comparative Analysis of Fraud Detection in Healthcare using Data Balancing & Machine Learning Technique"[4] emphasizes the use of data balancing and machine learning techniques to enhance healthcare fraud detection. The experimental results from authors Nikita Agrawal et al. indicate that machine learning models oversampling the imbalanced dataset using two data balancing techniques, namely Class Weighing Scheme (CWS) and Adaptive Synthetic Oversampling (ADASYN), outperform the imbalanced dataset. This finding prompts us to consider adopting data balancing techniques in healthcare fraud detection to improve model performance metrics.

Lastly, in the paper titled "Predicting health insurance claim frauds using supervised machine learning technique"[5] authors Veena K et al. propose a method for healthcare fraud detection using the decision tree classifier algorithm. They compare the accuracy of four algorithms—logistic regression, random forest, decision tree classifier, and naive Bayes—in fraud detection. Their experimental results show that the decision tree classifier exhibits outstanding accuracy, reaching 97.03% in fraud detection. This finding provides a strong starting point for us to consider whether we should incorporate the decision tree classifier into our model to enhance the accuracy of our predictive model.

Through in-depth examination of these papers, we have gained insights into the pivotal roles played by feature engineering, data balancing and decision tree classifiers techniques in healthcare fraud prediction. These insights will guide us in formulating sound data preprocessing, feature extraction plans and methodological strategies for our project, ensuring that we leverage the experiences and successes of these previous works to address the challenges of healthcare fraud prediction more effectively.

---

# Part 3: Data Exploration and Feature Engineering

## 1. Data Source and Description

Unfortunately, the Kaggle page does not provide metadata for the dataset, necessitating external research to grasp the variables and the data thoroughly. Noteworthy insights into the dataset can be found in a blog post authored by Pulkit Ratna Ganjeer, available at the following link: Pulkit Ratna Ganjeer's Blog Post.

The table below are significant takeaways from the blog that were not highlighted on the Kaggle page:

**Table 1: Data Source and Description**

## Mapping Data Details

| Sections | Details |
|---|---|
| Mapping Data | Contains the mapping of Provider's unique id and the class label signifying whether the Provider is fraud or not. For the Test data, only the Provider's unique id and the class label is given. |

## Beneficiary Data Details

| Sections | Details |
|---|---|
| IP Reimbursement | Annual amount reimbursed for the treatment of the beneficiary when admitted to the hospital. |
| IP Deductible | Annual premium amount paid to the Insurance Agency towards the treatment of the beneficiary when admitted to the hospital. |
| OP Reimbursement | Annual amount reimbursed for the treatment of the beneficiary when visited the hospital but not admitted. |
| OP Deductible | Annual premium amount paid to the Insurance Agency towards the treatment of the beneficiary when he visited the hospital but was not admitted. |

## Inpatient and Outpatient Data Details

| Sections | Details |
|---|---|
| InscReimbursed | Amount reimbursed by the Payer (Insurance Agency) for the healthcare services provided to the beneficiary. |
| Physicians-related columns | Columns showing the physicians who attended the beneficiary/patient, operated the patient, and any other physicians if any. |
| Admission and Discharge Date (Inpatient Data) | Columns showing the dates on which the beneficiary was admitted to the hospital and when he was discharged. |
| Claim Diagnosis Codes 1-10 | ClmDiagnosisCode_1: Diagnosis code identifying the beneficiary's principal diagnosis. ClmDiagnosisCode_2-10: Diagnosis code in the 2nd, 3rd, and so on, till the 10th position identifying the condition(s) for which the beneficiary is receiving care. |
| Claim Procedure Codes 1-6 | Codes that indicate the principal or other procedures performed during the period covered by the institutional claim. |
| DiagnosisGroupCode (Inpatient Data) | Code to classify hospital cases according to certain groups, also referred to as DRGs, which are expected to have similar hospital resource use (cost). |
| DeductibleAmtPaid | Amount the beneficiary has to pay as part of the claim, and the rest of the amount is paid by the insurance company. It is equal to the total claim amount minus the reimbursed amount. |

# 2. Exploratory Data Analysis (EDA) and Data Preprocessing

## 2.1 Overview of Train and Test Datasets and Data Cleaning

In the initial analysis, we examined the training dataset, which has 5410 entries, primarily focusing on two key variables: “Provider” and “PotentialFraud.” The “Provider” is a unique identifier for healthcare providers, and “PotentialFraud” indicates possible fraudulent activities, marked as “Yes” or “No.” Notably, there’s an imbalance in the “PotentialFraud” distribution, with a higher proportion of non-fraudulent cases at 90.64% compared to 9.36% fraudulent ones. Importantly, the training and test datasets are distinct, with no common providers, enabling accurate model evaluation on unseen data.

The test dataset, with 1353 entries, has a similar structure, using the “Provider” column to identify providers. We found that the “Provider” column in the training dataset is unique across entries, indicating no duplicates, and both datasets have no duplicate rows or missing values. This suggests well-organized data that doesn’t require initial cleaning or deduplication.

Moving forward, we’ll consider strategies like transforming the “PotentialFraud” variable for classification tasks. Since the test dataset lacks this column, it’s suitable for validation based on insights from the training dataset. Strategies like cross-validation and addressing data imbalance will be essential to build a robust machine learning model while avoiding overfitting.

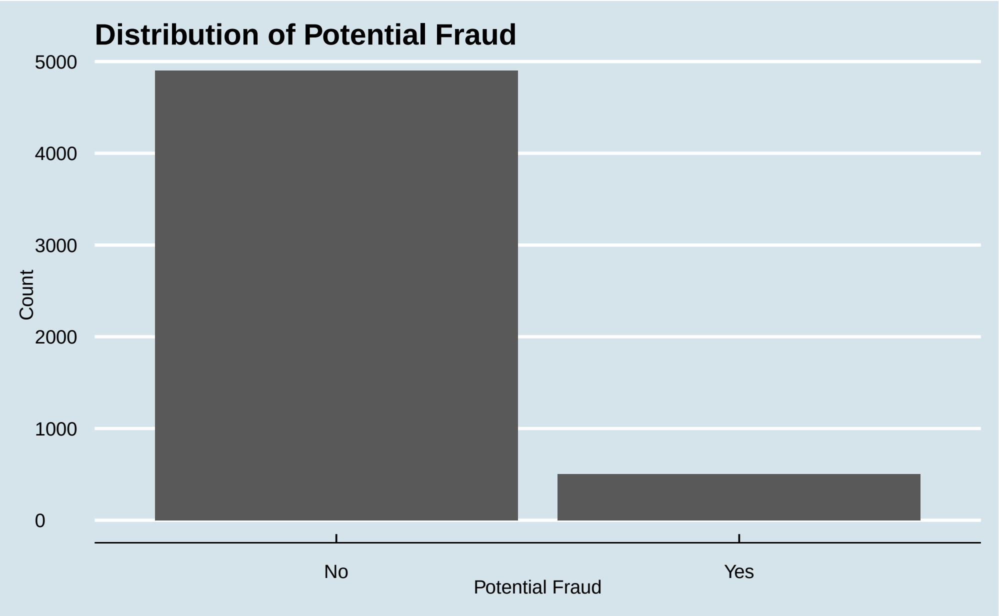


**Figure 1:** Distribution of Potential Fraud in the Train dataset.

---

## 2.2 Beneficiary Data Overview and Data Pre-processing

### 2.2.1 Demographic Details

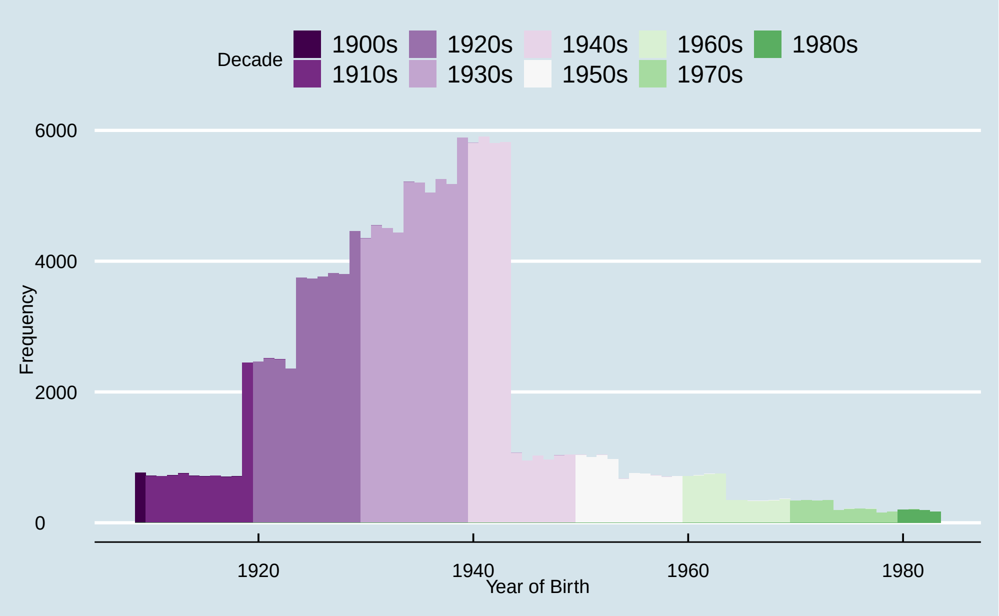


**Figure 2:** Distribution of Year of Birth in the Train_Beneficiary dataset.

The distribution of birth years in the demographic details of beneficiaries exhibits a significant skew, primarily concentrated in the decades of the 1920s, 1930s, and 1940s. This allows us to identify the predominant age groups in the dataset, serving as important predictive variables for analyzing health conditions and financial implications.

### 2.2.2 Gender and Race

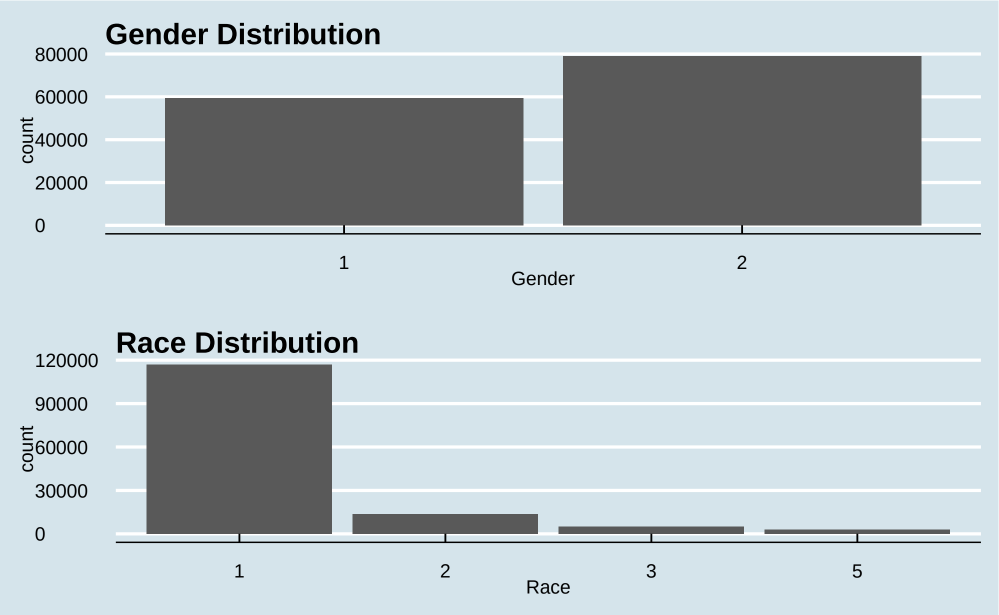


**Figure 3.** The distribution of Gender and Race in the Train_Beneficiary dataset.

Analyzing the gender and race variables reveals a limited yet distinct categorization. The gender variable bifurcates into two categories, while the race variable encompasses three unique levels. This categorical data is poised to be a critical asset in predictive modelling, potentially unveiling patterns or trends that exhibit correlation with gender or race factors.

### 2.2.3 Geographical Disparity in State and County

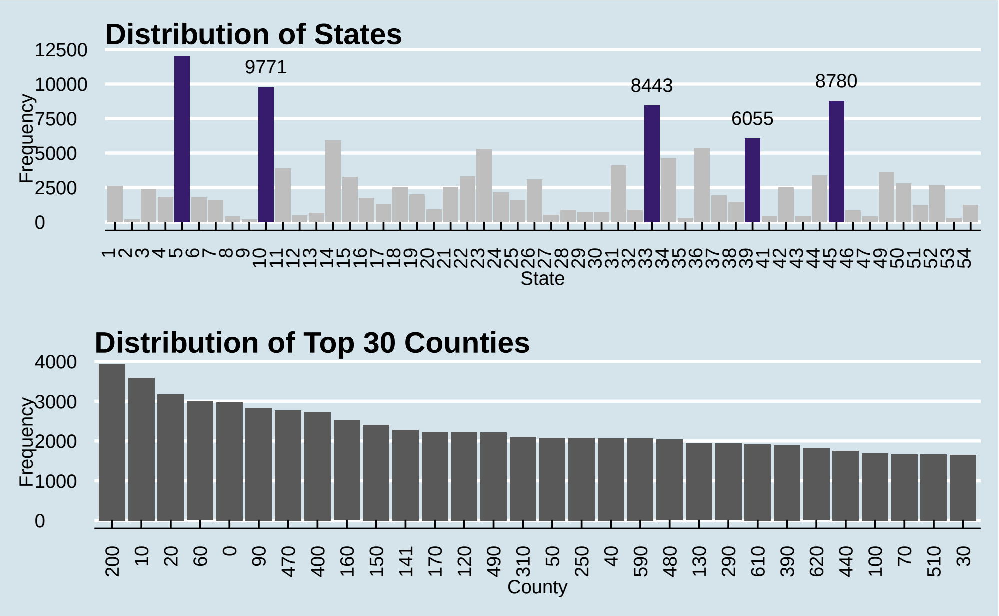
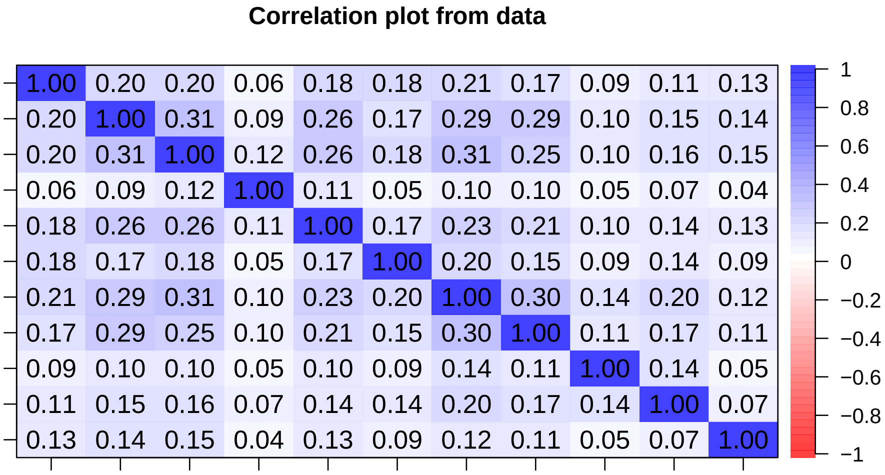


**Figure 4.** The distribution of States and Counties in the Train_Beneficiary dataset.

The dataset unveils considerable geographical disparities, representing 54 unique states and 314 distinct counties. The state data is predominantly clustered around labels 5, 10, 45, 33, and 39, and the prominent counties are designated as 200, 10, 20, 60, and 0. This geographical data can be a linchpin in assessing healthcare accessibility and uncovering potential fraud patterns localized within specific regions.

### 2.2.4 Health Conditions



| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 1.00 | 0.20 | 0.20 | 0.06 | 0.18 | 0.18 | 0.21 | 0.17 | 0.09 | 0.11 | 0.13 |
| 0.20 | 1.00 | 0.31 | 0.09 | 0.26 | 0.17 | 0.29 | 0.29 | 0.10 | 0.15 | 0.14 |
| 0.20 | 0.31 | 1.00 | 0.12 | 0.26 | 0.18 | 0.31 | 0.25 | 0.10 | 0.16 | 0.15 |
| 0.06 | 0.09 | 0.12 | 1.00 | 0.11 | 0.05 | 0.10 | 0.10 | 0.05 | 0.07 | 0.04 |
| 0.18 | 0.26 | 0.26 | 0.11 | 1.00 | 0.17 | 0.23 | 0.21 | 0.10 | 0.14 | 0.13 |
| 0.18 | 0.17 | 0.18 | 0.05 | 0.17 | 1.00 | 0.20 | 0.15 | 0.09 | 0.14 | 0.09 |
| 0.21 | 0.29 | 0.31 | 0.10 | 0.23 | 0.20 | 1.00 | 0.30 | 0.14 | 0.20 | 0.12 |
| 0.17 | 0.29 | 0.25 | 0.10 | 0.21 | 0.15 | 0.30 | 1.00 | 0.11 | 0.17 | 0.11 |
| 0.09 | 0.10 | 0.10 | 0.05 | 0.10 | 0.09 | 0.14 | 0.11 | 1.00 | 0.14 | 0.05 |
| 0.11 | 0.15 | 0.16 | 0.07 | 0.14 | 0.14 | 0.20 | 0.17 | 0.14 | 1.00 | 0.07 |
| 0.13 | 0.14 | 0.15 | 0.04 | 0.13 | 0.09 | 0.12 | 0.11 | 0.05 | 0.07 | 1.00 |

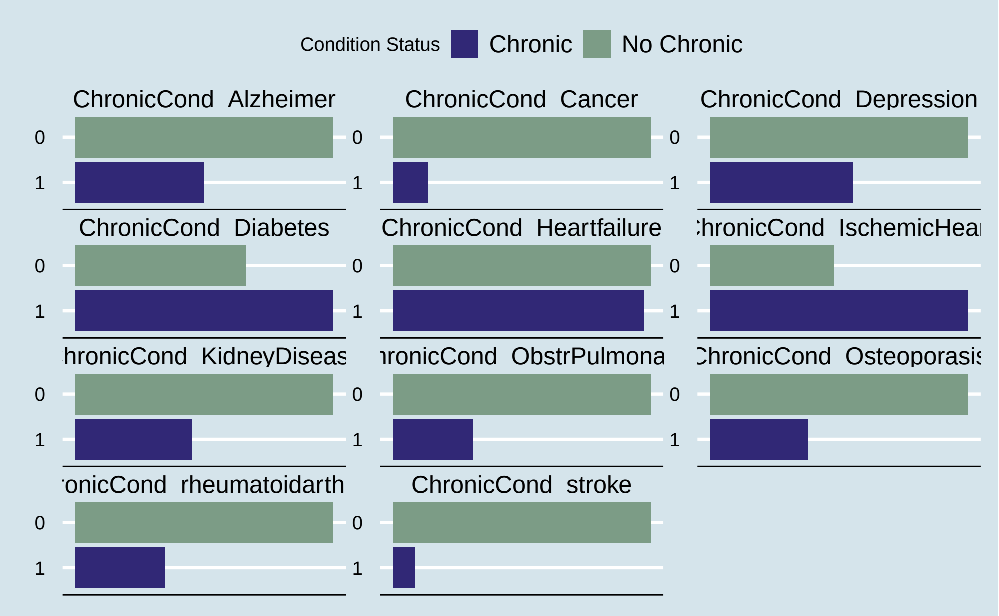
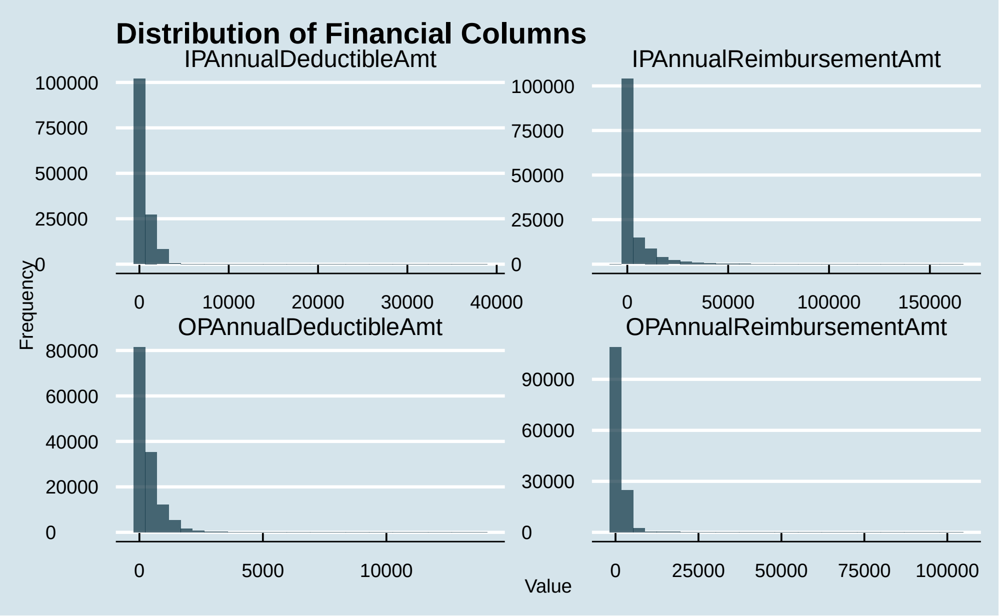


**Figure 5.** The correlation distribution of Health Conditions in the Train_Beneficiary dataset.

The analysis of chronic health conditions in the dataset did not reveal significant correlations between different diseases, indicating their independent occurrence. This independence is beneficial in predictive modeling as it avoids issues of multicollinearity and enhances model stability. Furthermore, a more in-depth analysis showed a higher prevalence of certain chronic conditions such as diabetes, heart failure, and ischemic heart disease. This finding helps us understand the common health issues among the beneficiary population.

### 2.2.5 Financial Details



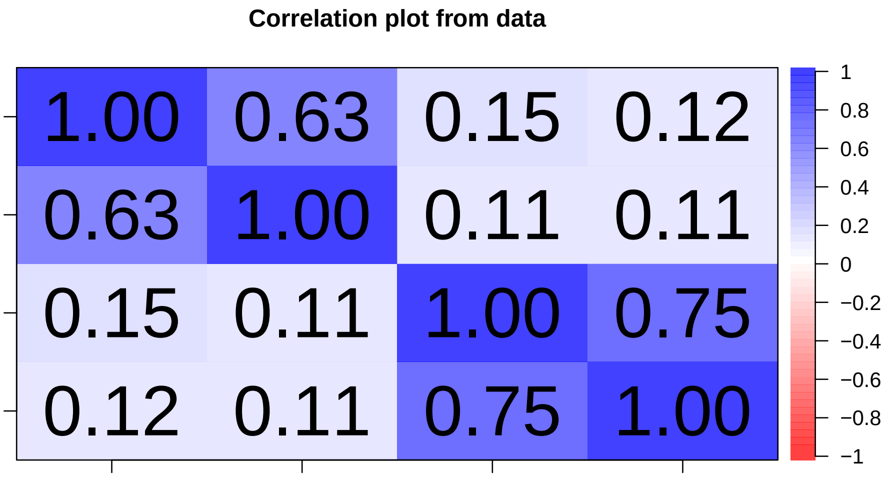


**Figure 6.** The Distribution and correlation of Financial Columns in the Train_Beneficiary dataset.

The financial attributes section demonstrates a right-skewed distribution, with most beneficiaries having lower annual reimbursements and deductibles, while only a very small minority has higher values. This reflects the general health condition of the beneficiaries. However, the high-value outliers require further scrutiny to uncover potential fraudulent activities. Furthermore, the significant correlations among various financial columns highlight the crucial role of annual reimbursements and deductibles in fraud detection.

### 2.2.6 Critical Event Information-Date of death

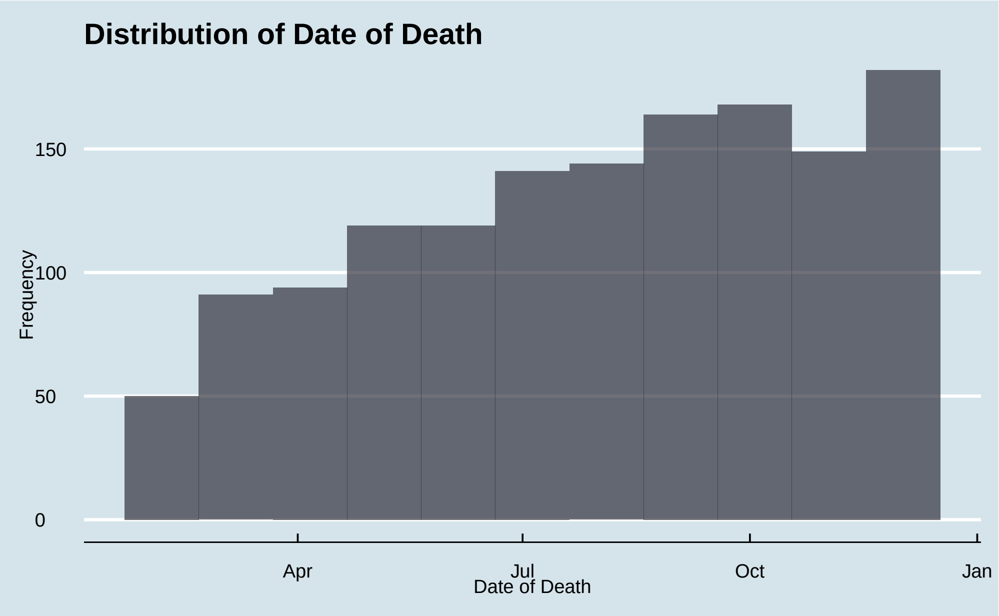


**Figure 7.** The Distribution of Date of Death in the Train_Beneficiary dataset.

The dataset signifies a uniformity in the recorded year of death, pinpointed to the year 2009. This trend perhaps is a consequence of the dataset's restriction to the data documented in the year 2009, with a noteworthy concentration of death occurrences in December and the latter half of the year.

### 2.3 Inpatient and Outpatient Data Overview and Data Pre-processing

In the endeavor to unravel nuanced insights into healthcare fraud detection, the analysis extends to a comprehensive review of inpatient and outpatient data, facilitating a meticulous approach to identifying unusual patterns that may indicate fraudulent activities. This section explicates the methodology and observations gleaned from data pre-processing and analysis.

#### 2.3.1 Merging Inpatient and Outpatient Datasets

The merging of Train_Inpatient and Train_Outpatient data is intended to present the medical journeys of patients. This merger includes 558,211 rows and 30 columns, allowing for a detailed examination of each beneficiary's medical encounters. It reveals patterns and correlations that may remain hidden when analyzed separately. This helps identify unusual claims patterns and marks cases for further investigation, thereby enhancing the integrity of the healthcare sector.

#### 2.3.2 BeneID, ClaimID, Provider

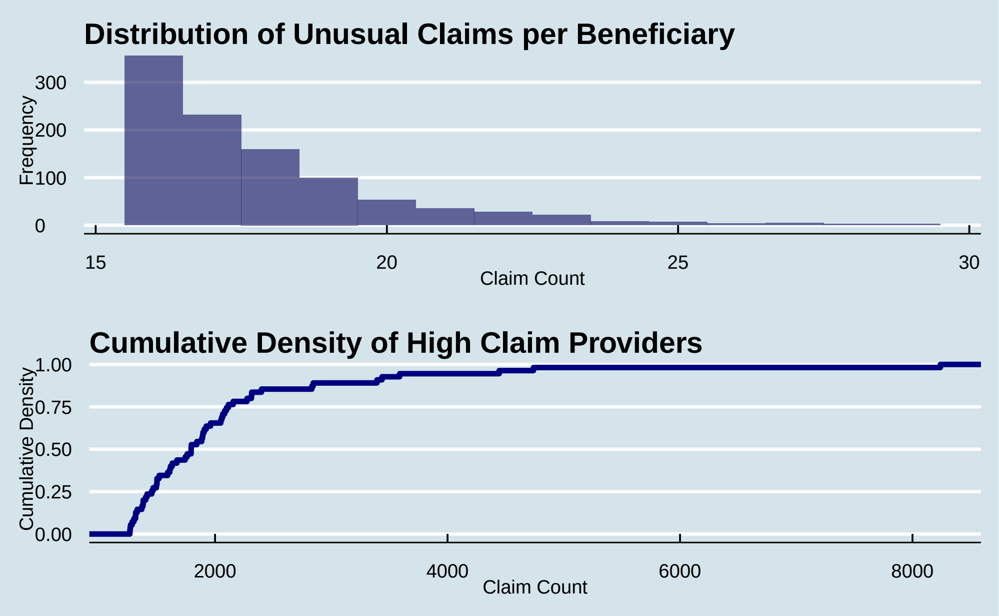


**Figure 8.** The Distribution of Unusual Claims per Beneficiary and Cumulative Density of High Claim Providers in the Train_Inpatient and Train_Outpatient dataset.

Examining the dataset, we find 138,556 unique beneficiaries, 558,211 unique claims, and 5,410 different providers. Notably, the average claim value per beneficiary is around 4.029, which helps identify unusual claim patterns and improves fraud detection precision. Additionally, filtering out data entries above the 99th percentile narrows down beneficiaries and providers with high claim counts for closer inspection.

The Cumulative Density Plot (CDP) shows a transition zone, indicating a shift from common to higher claim counts, possibly suggesting a gradual increase in unusual claims. This area, especially when claim counts exceed 4,000, is crucial for further investigation, highlighting providers who significantly deviate from the norm.

#### 2.3.3 Reimbursed and Deductible Amounts

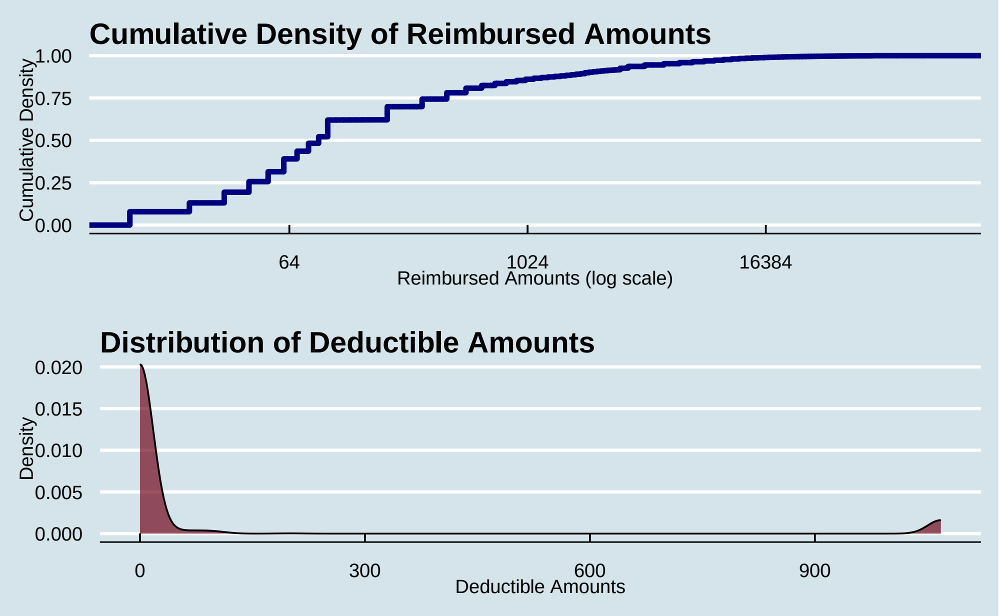


**Figure 9.** The Cumulative Density of Reimbursed Amounts and Distribution of Deductible Amounts in the Train_Inpatient and Train_Outpatient dataset.

The data indicates that most claims have low reimbursement amounts, but sometimes there are very high amounts, which might be due to fraud or errors. In medical provider fraud detection, these irregularities could be red flags and need closer scrutiny to prevent fraud or billing mistakes. Also, many people have zero deductible amounts, meaning they don't have to pay deductibles. However, some individuals have unusually high deductibles, which are quite different from the norm. This emphasizes the need for strong methods to spot these unusual cases.

## 2.4. Merge Dataset and Data Cleaning

This dataset has 558,211 records and 54 features. 'ClaimID' identifies each claim, and 'BeneID' distinguishes each beneficiary for monitoring their medical situations and claims.

To prepare the data for analysis, we made some changes. We converted binary values (0 and 1) in the 'Gender' column into numbers. In the 'RenalDiseaseIndicator' column, 'Y' became '1' to show renal disease, while others stayed the same.

We filled empty values in the 'DeductibleAmtPaid' column with '0'. Lastly, in the 'PotentialFraud' column, 'Yes' turned into '1' for potential fraud, and everything else became '0'. These steps made the dataset consistent and ready for analysis.

Table 4: Summary Statistics for Variables in Train Dataset

| skim_type | skim_variable | n_missing | numeric.mean | numeric.sd | numeric.p0 | numeric.p50 | numeric.p100 |
|---|---|---|---|---|---|---|---|
| Date | ClaimStartDt | 0 | NA | NA | NA | NA | NA |
| Date | ClaimEndDt | 0 | NA | NA | NA | NA | NA |
| Date | DOB | 0 | NA | NA | NA | NA | NA |
| Date | DOD | 554080 | NA | NA | NA | NA | NA |
| character | BeneID | 0 | NA | NA | NA | NA | NA |

| skim_type | skim_variable | n_missing | numeric.mean | numeric.sd | numeric.p0 | numeric.p50 | numeric.p100 |
|---|---|---|---|---|---|---|---|
| character | Provider | 0 | NA | NA | NA | NA | NA |
| character | ClaimID | 0 | NA | NA | NA | NA | NA |
| character | AttendingPhysician | 1508 | NA | NA | NA | NA | NA |
| character | OperatingPhysician | 443764 | NA | NA | NA | NA | NA |
| character | OtherPhysician | 358475 | NA | NA | NA | NA | NA |
| character | AdmissionDt | 517737 | NA | NA | NA | NA | NA |
| character | ClmAdmitDiagnosisCode | 0 | NA | NA | NA | NA | NA |
| character | DischargeDt | 517737 | NA | NA | NA | NA | NA |
| character | DiagnosisGroupCode | 517737 | NA | NA | NA | NA | NA |
| character | ClmDiagnosisCode_1 | 10453 | NA | NA | NA | NA | NA |
| character | ClmDiagnosisCode_2 | 195606 | NA | NA | NA | NA | NA |
| character | ClmDiagnosisCode_3 | 315156 | NA | NA | NA | NA | NA |
| character | ClmDiagnosisCode_4 | 393675 | NA | NA | NA | NA | NA |
| character | ClmDiagnosisCode_5 | 446287 | NA | NA | NA | NA | NA |
| character | ClmDiagnosisCode_6 | 473819 | NA | NA | NA | NA | NA |
| character | ClmDiagnosisCode_7 | 492034 | NA | NA | NA | NA | NA |
| character | ClmDiagnosisCode_8 | 504767 | NA | NA | NA | NA | NA |
| character | ClmDiagnosisCode_9 | 516396 | NA | NA | NA | NA | NA |
| character | ClmDiagnosisCode_10 | 553201 | NA | NA | NA | NA | NA |
| character | RenalDiseaseIndicator | 0 | NA | NA | NA | NA | NA |
| factor | Race | 0 | NA | NA | NA | NA | NA |
| factor | State | 0 | NA | NA | NA | NA | NA |
| factor | County | 0 | NA | NA | NA | NA | NA |
| logical | ClmProcedureCode_6 | 558211 | NA | NA | NA | NA | NA |
| numeric | InscClaimAmtReimbursed | 0 | 997.01 | 3821.53 | 0.00 | 80.00 | 125000.00 |
| numeric | DeductibleAmtPaid | 0 | 78.29 | 273.81 | 0.00 | 0.00 | 1068.00 |
| numeric | ClmProcedureCode_1 | 534901 | 5896.15 | 3050.49 | 11.00 | 5363.00 | 9999.00 |
| numeric | ClmProcedureCode_2 | 552721 | 4106.36 | 2031.64 | 42.00 | 4019.00 | 9999.00 |
| numeric | ClmProcedureCode_3 | 557242 | 4221.12 | 2281.85 | 42.00 | 4019.00 | 9999.00 |
| numeric | ClmProcedureCode_4 | 558093 | 4070.26 | 2037.63 | 42.00 | 4019.00 | 9986.00 |
| numeric | ClmProcedureCode_5 | 558202 | 5269.44 | 2780.07 | 2724.00 | 4139.00 | 9982.00 |
| numeric | PotentialFraud | 0 | 0.38 | 0.49 | 0.00 | 0.00 | 1.00 |
| numeric | ClaimDuration | 0 | 1.73 | 4.90 | 0.00 | 0.00 | 36.00 |
| numeric | Gender | 0 | 0.58 | 0.49 | 0.00 | 1.00 | 1.00 |
| numeric | ChronicCond_Alzheimer | 0 | 1.60 | 0.49 | 1.00 | 2.00 | 2.00 |
| numeric | ChronicCond_Heartfailure | 0 | 1.41 | 0.49 | 1.00 | 1.00 | 2.00 |
| numeric | ChronicCond_KidneyDisease | 0 | 1.59 | 0.49 | 1.00 | 2.00 | 2.00 |
| numeric | ChronicCond_Cancer | 0 | 1.85 | 0.36 | 1.00 | 2.00 | 2.00 |
| numeric | ChronicCond_ObstrPulmonary | 0 | 1.69 | 0.46 | 1.00 | 2.00 | 2.00 |
| numeric | ChronicCond_Depression | 0 | 1.57 | 0.50 | 1.00 | 2.00 | 2.00 |
| numeric | ChronicCond_Diabetes | 0 | 1.29 | 0.46 | 1.00 | 1.00 | 2.00 |
| numeric | ChronicCond_IschemicHeart | 0 | 1.24 | 0.43 | 1.00 | 1.00 | 2.00 |
| numeric | ChronicCond_Osteoporasis | 0 | 1.68 | 0.47 | 1.00 | 2.00 | 2.00 |
| numeric | ChronicCond_rheumatoidarthritis | 0 | 1.69 | 0.46 | 1.00 | 2.00 | 2.00 |
| numeric | ChronicCond_stroke | 0 | 1.90 | 0.30 | 1.00 | 2.00 | 2.00 |
| numeric | IPAnnualReimbursementAmt | 0 | 5227.97 | 11786.27 | -8000.00 | 0.00 | 161470.00 |
| numeric | IPAnnualDeductibleAmt | 0 | 568.76 | 1179.17 | 0.00 | 0.00 | 38272.00 |
| numeric | OPAnnualReimbursementAmt | 0 | 2278.23 | 3881.85 | -70.00 | 1170.00 | 102960.00 |
| numeric | OPAnnualDeductibleAmt | 0 | 649.70 | 1002.02 | 0.00 | 340.00 | 13840.00 |

# 3. Missing Data Analysis and Intervention

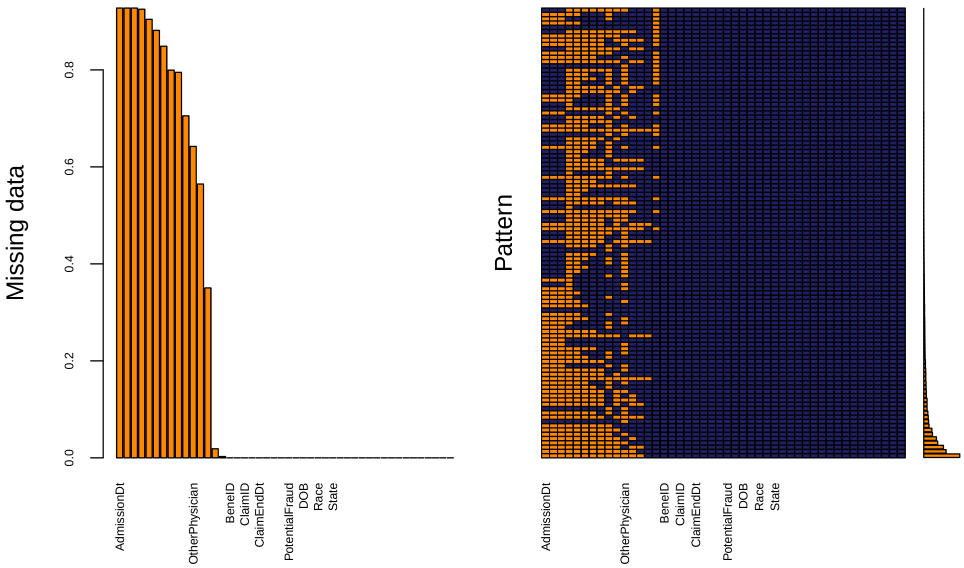
### 3.1.1. Management of Missing Data

In proficiently navigating missing data within the healthcare fraud detection paradigm, a structured methodology was employed. A significant threshold of 95% was instituted for potential column elimination, ensuring data reliability whilst diminishing irrelevant fluctuations.

Noteworthy percentages of missing data in the 'AdmissionDt' and 'DischargeDt' columns (approximately 92.75%) are logically accounted for, as they encapsulate substantial events from 2009. These columns are essential in formulating the 'length of stay' feature, a crucial aspect in fraud detection examinations.

Furthermore, the 'ClmDiagnosisCode_2-10' columns encapsulate a range of medical conditions, warranting the varying missing values. Here, applying domain knowledge prudently is vital in determining their retention, consequently augmenting the fraud detection model's accuracy and efficiency.

Columns possessing over 30% missing values are provisionally segregated from the original dataset, while columns with less than 30% missing values are retained for imputation.



```
##
##  Variables sorted by number of missings:
##                Variable       Count
##             AdmissionDt 0.927493367
##             DischargeDt 0.927493367
##      DiagnosisGroupCode 0.927493367
##      ClmDiagnosisCode_9 0.925091050
##      ClmDiagnosisCode_8 0.904258426
##      ClmDiagnosisCode_7 0.881448055
##      ClmDiagnosisCode_6 0.848817024
##      ClmDiagnosisCode_5 0.799495173
##     OperatingPhysician 0.794975377
##      ClmDiagnosisCode_4 0.705244074
##          OtherPhysician 0.642185482
##      ClmDiagnosisCode_3 0.564582210
##      ClmDiagnosisCode_2 0.350415882
##               ClmDiagnosisCode_1 0.018725894
##               AttendingPhysician 0.002701487
##                           BeneID 0.000000000
##                         Provider 0.000000000
##                          ClaimID 0.000000000
##                     ClaimStartDt 0.000000000
##                       ClaimEndDt 0.000000000
##           InscClaimAmtReimbursed 0.000000000
##            ClmAdmitDiagnosisCode 0.000000000
##                DeductibleAmtPaid 0.000000000
##                   PotentialFraud 0.000000000
##                    ClaimDuration 0.000000000
##                              DOB 0.000000000
##                           Gender 0.000000000
##                             Race 0.000000000
##            RenalDiseaseIndicator 0.000000000
##                            State 0.000000000
##                           County 0.000000000
##            ChronicCond_Alzheimer 0.000000000
##         ChronicCond_Heartfailure 0.000000000
##        ChronicCond_KidneyDisease 0.000000000
##               ChronicCond_Cancer 0.000000000
##       ChronicCond_ObstrPulmonary 0.000000000
##           ChronicCond_Depression 0.000000000
##             ChronicCond_Diabetes 0.000000000
##        ChronicCond_IschemicHeart 0.000000000
##         ChronicCond_Osteoporasis 0.000000000
##  ChronicCond_rheumatoidarthritis 0.000000000
##               ChronicCond_stroke 0.000000000
##         IPAnnualReimbursementAmt 0.000000000
##            IPAnnualDeductibleAmt 0.000000000
##         OPAnnualReimbursementAmt 0.000000000
##            OPAnnualDeductibleAmt 0.000000000
```

**Figure 10.** The Distribution of Missing Data in the Train_claim dataset.

### 3.1.2. Procedure for Missing Data Imputation

From the observed pattern of missing data, it is evident that the missing values are not absent completely at random (MCAR). These missing values bear relations to other variables, albeit not with the missing data variable. Hence, the Multiple Imputation by Chained Equations (MICE) method will be deployed to address the missing values less than 30%. This approach leverages other variables within the dataset to derive missing values, employing the mice package in R for implementation.

- m = 5, denotes the generation of five imputed datasets, providing a spectrum of values instead of a singular prediction.
- maxit = 50, indicates the iterative rounds undertaken for missing values imputation.
- The mean matching method is adopted for imputation.

---

## 4. Feature Selection

In this section, filter methods are utilized for selecting pertinent features, leveraging the mlr3 package to facilitate the implementation of this approach. Through a rigorous analysis, it becomes discernible which features serve as substantial predictors in the model, while also identifying those which offer limited contribution to the prediction of the target variable, based on the criteria of information gain.

Upon examining the scores, it is evident that certain features manifest high scores, implying their potential as significant predictors in the model. Specifically, features such as ‘ClaimID’, ‘Provider’, and ‘AttendingPhysician’ have surfaced as potential powerhouse predictors, holding substantial influence in determining the outcome of the target variable.

In contrast, features with scores nearing or equating to zero demonstrate negligible impact on the target variable. It has been observed that features including ‘ChronicCond_Cancer’, ‘ChronicCond_Depression’, and ‘Gender’, exert little to no influence on the predictive capability of the model, adhering to the parameters set by the information gain criterion.

Moving forward, a strategic approach will be adopted where features manifesting high scores will be given priority during the model development phase. This strategy is grounded on data-driven insights and aims to enhance the model’s predictive accuracy by focusing on significant predictors, thereby avoiding potential noise generated by irrelevant features.

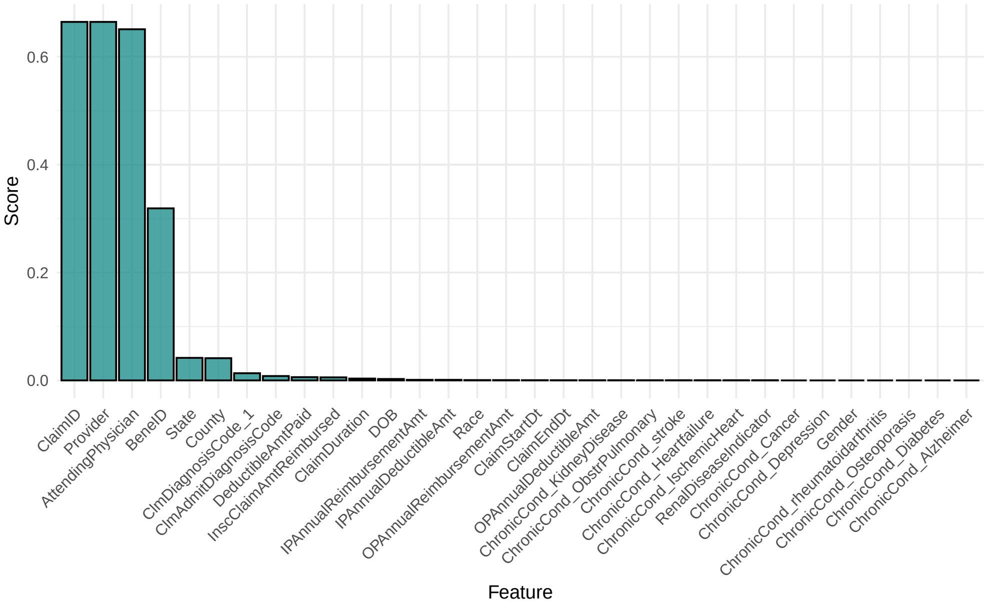


**Figure 11.** The Feature Importance Plot in the Train_claim dataset.

**Table 3.** Selected Features from the Feature Selection Process.

# Selected Features

| Sections | Details |
|---|---|
| BeneID | Unique identifiers for beneficiaries. Categorical variable. |
| Provider | Unique identifiers for healthcare providers. Categorical variable. |
| ClaimID | Unique identifiers for claims. Categorical variable. |
| ClaimStartDt and ClaimEndDt | Start and end dates of claims, ranging from November 27, 2008, to December 31, 2009. Datetime variables. |
| InscClaimAmtReimbursed | Amount reimbursed for insurance claims. Continuous variable with a range from 0 to 125,000. |

| Sections | Details |
|---|---|
| AttendingPhysician | Names or identifiers of attending physicians. Categorical variable. |
| ClmAdmitDiagnosisCode | Diagnosis codes related to admission. Categorical variable. |
| DeductibleAmtPaid | Deductible amount paid, ranging from 0 to 1,068. |
| ClmDiagnosisCode_1 | Additional diagnosis codes with a substantial number of missing values (10,453). Categorical variable. |
| PotentialFraud | Binary variable indicating potential fraud (1) or no fraud (0). Not missing. |
| ClaimDuration | Duration of claims, ranging from 0 to 36. |
| DOB | Date of birth of beneficiaries, ranging from January 1, 1909, to December 1, 1983. Datetime variable. |
| Gender | Binary variable representing the gender of beneficiaries (0 or 1). |
| Race | Categorical variable with values 1, 2, 3, or 5. |
| RenalDiseaseIndicator | Binary variable indicating the presence (1) or absence (0) of renal disease. |
| State | Categorical variable with multiple levels. |
| County | Categorical variable with multiple levels. |
| Chronic Conditions | Binary variables indicating the presence or absence of various chronic conditions. |
| IPAnnualReimbursementAmt and IPAnnualDeductibleAmt | Likely annual reimbursement amount and deductible amount for inpatient services. |
| OPAnnualReimbursementAmt and OPAnnualDeductibleAmt | Likely annual reimbursement amount and deductible amount for outpatient services. |

## 5. Correlation analysis

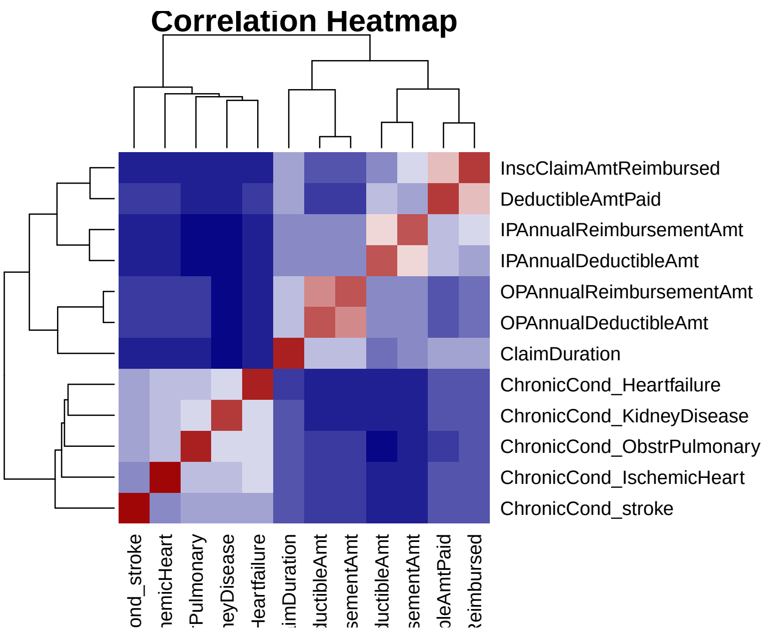


**Figure 12.** The Correlation Heatmap for the selected features.

The analysis of different variables in the dataset shows different relationships. We found a strong positive relationship between insurance claim amounts and deductible amounts (correlation coefficient approximately

0.654). There's a moderate relationship (ranging from 0.226 to 0.384) among annual reimbursements, deductibles for inpatient services, and reimbursed amounts. Also, there's a weak relationship (less than 0.2) between annual reimbursements and deductibles for outpatient services.

The data suggests that longer claim durations might have a small impact on reimbursed amounts and deductible payments, with ranging from 0.217 to 0.257. These different levels of relationships give us insights into how different factors are connected and can help us make better predictions.

## 6. Feature engineering

### 6.1. Temporal Features, Medical Condition and Other Features

The narrative strategically develops crucial temporal features fundamental to building a robust predictive model. Initially, it delineates the formation of the 'Age at the Time of Claim' and 'Claim Processing Time' variables through arithmetic manipulations on existing columns in the 'Train_claim_cleaned' data frame, illustrating this using R programming snippets.

Subsequently, the focus shifts to synthesizing a composite feature that encapsulates the severity of a patient's chronic conditions by aggregating data points from relevant columns, thus offering a detailed view of patients' health trajectories over time.

Lastly, the discussion expands to the generation of varied features, crafted through advanced statistical and mathematical approaches. Here, transformative data science techniques take center stage, facilitating the creation of new variables through interaction features, polynomial features, and categorical binning, among others, thereby enhancing the data's dimensionality and paving the way for an in-depth analytical expedition.

### 6.2. Features Transformation

The transformation of features address the statistical nuances of skewness and kurtosis which play a pivotal role in moulding the efficacy of machine learning algorithms. The analytical spotlight here is on the marked skewness and kurtosis evident in the data, advocating for transformations, predominantly log transformations, to normalize the data distribution. This stride not only mitigates skewness but serves as a bulwark against the adverse effects of outliers, steering the data towards a more standardized landscape.

**Table 4.** Skewness and Kurtosis Analysis for Features Transformation.

# Skewness Data Details

| Variable | Skewness Value |
|---|---|
| InscClaimAmtReimbursed | 9.49 |
| DeductibleAmtPaid | 3.33 |
| ClaimDuration | 3.16 |
| IPAnnualReimbursementAmt | 3.99 |
| IPAnnualDeductibleAmt | 8.16 |
| OPAnnualReimbursementAmt | 5.52 |
| OPAnnualDeductibleAmt | 4.36 |
| MissingDiagnosisCode | 7.10 |

## Kurtosis Data Details

| Variable | Kurtosis Value |
|---|---|
| InscClaimAmtReimbursed | 136.58 |
| DeductibleAmtPaid | 9.11 |
| ClaimDuration | 9.25 |
| ChronicCond_stroke | 4.94 |
| IPAnnualReimbursementAmt | 23.04 |
| IPAnnualDeductibleAmt | 158.20 |
| OPAnnualReimbursementAmt | 51.02 |
| OPAnnualDeductibleAmt | 26.17 |
| ClaimDurationSquared | 26.98 |
| ClaimDurationStandardized | 9.25 |
| MissingDiagnosisCode | 48.42 |

---

### 7. Export Data to CSV

You can request the generated CSV file by emailing us at the university email address.

```r
# Export the cleaned dataset to csv
write.csv(Train_claim_cleaned, "Train_claim_cleaned.csv", row.names = FALSE)
```

---

### References:

[1] A. Bhardwaj, S. Kumar and A. Naidu, "Predictive analysis and supervised detection for fraudulent cases in healthcare," 2022 12th International Conference on Cloud Computing, Data Science & Engineering (Confluence), Noida, India, 2022, pp. 416-421, doi: 10.1109/Confluence52989.2022.9734195.

[2] V. Rawte and G. Anuradha, "Fraud detection in health insurance using data mining techniques," 2015 International Conference on Communication, Information & Computing Technology (ICCICT), Mumbai, India, 2015, pp. 1-5, doi: 10.1109/ICCICT.2015.7045689.

[3] J. M. Johnson and T. M. Khoshgoftaar, "Healthcare Provider Summary Data for Fraud Classification" 2022 IEEE 23rd International Conference on Information Reuse and Integration for Data Science (IRI), San Diego, CA, USA, 2022, pp. 236-242, doi: 10.1109/IRI54793.2022.00060.

[4] N. Agrawal and S. Panigrahi, "A Comparative Analysis of Fraud Detection in Healthcare using Data Balancing & Machine Learning Techniques" 2023 International Conference on Communication, Circuits, and Systems (IC3S), BHUBANESWAR, India, 2023, pp. 1-4, doi: 10.1109/IC3S57698.2023.10169634.

[5] V. K et al., "Predicting health insurance claim frauds using supervised machine learning technique" 2023 Eighth International Conference on Science Technology Engineering and Mathematics (ICONSTEM), Chennai, India, 2023, pp. 1-7, doi: 10.1109/ICONSTEM56934.2023.10142604.

[6] Ganjeer, P. R. (2023). Healthcare Fraud Detection Using Machine Learning. Retrieved from https://pulkitratnaganjeer.medium.com/healthcare-fraud-detection-using-machine-learning-5996d63bd3c7