FINAL PROJECT

HEALTHCARE PROVIDER FRAUD DETECTION ANALYSIS

1. Abstract

Healthcare is a basic necessity in people’s lives and it should be affordable. The healthcare industry is an intricate system with numerous moving components and is expanding at an expeditious pace. However, fraud in this industry is also turning into a critical problem. Healthcare fraud is becoming more and more common these days, especially the misuse of the medical insurance systems.Detecting these fraud manually is almost an impossible task, hence we machine Learning algorithms and data mining techniques are being used for automatic detection of these. In this project, we attempt to give a review on frauds in the healthcare industry and the techniques for detecting such frauds. With an emphasis on the techniques used, determining the significant sources and the features of the healthcare data, various available researches were studied in the literature work. From this review it can be concluded that the advanced machine learning techniques and incipiently acquired sources of the healthcare data would be forthcoming subjects of interest in order to make the healthcare affordable, to improve the effectiveness of healthcare fraud detection and to obtain top quality on healthcare systems.

2. Introduction

Healthcare has and perpetuates to be an integral component in people’s lives. The human body is a compound structure. Hence, it is essential to have specialist physicians qualified to diagnose and treat diseases in different parts of the body. This induces several types of treatment procedures that physicians carry out for patients in different specialties. The aim of the health industry is to successfully serve as many patients as possible. But with every treatment there is a price associated with every service provided. Physicians, drug dealers and medical staff have to be paid for their time and prowess including various medical amenities. Oftentimes these prices are not affordable to the patients. Therefore, insurance schemes are used to dispense costs over all patients in the healthcare system and pay for the requisite people and equipment. As with any insurance system, there is a possibility for misuse or fraud activities. Healthcare fraud isincreasingly perceived as one of serious social concerns. Clearly, healthcare fraud is a problem for the government and there is a need for more effective detection methods. To detect healthcare fraud, it requires a great amount of effort with extensive medical knowledge. Traditionally, healthcare fraud detection greatly depends on the experience of domain experts, which is erroneous enough, expensive and time consuming. Manual detection of healthcare fraud involves a few auditors who manually review and identify the suspicious medical insurance claims which requires much effort. But the modern advances of machine learning and data mining techniques led to more efficient and automated detection of healthcare frauds. There has been a growing interest in mining healthcare data for fraud detection in recent years. This paper reviews the various approaches used for detecting the fraudulent activities in Health insurance claim data.

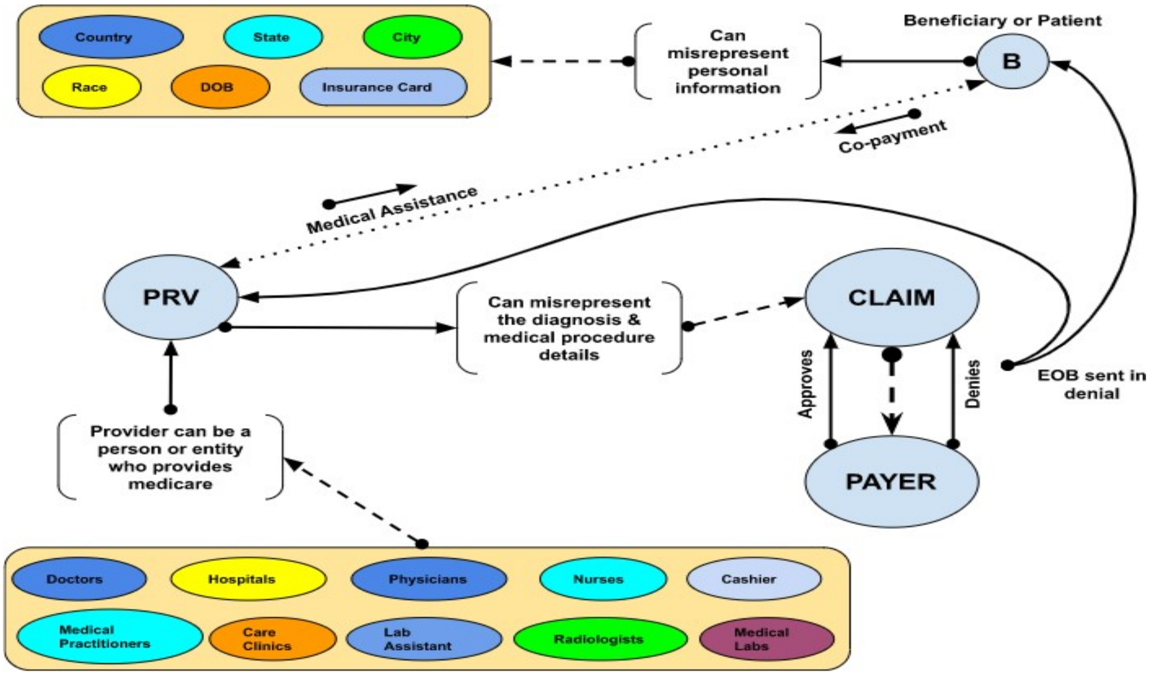
3. Problem Overview

Healthcare fraud is an organized crime that involves peers of providers (hospitals, cashiers, medical labs, nurses, lab assistants, and others), physicians,and beneficiaries acting together to make fraud claims.

Some basic examples of healthcare fraud are:

- Selling drugs, devices, foods, or medical cosmetics that have not been proven effective
- Billing of services or medical procedures which are not performed.

These scams can exist for any disease like weight loss, memory loss, sexual performance, joint pain, and serious illness like cancer, diabetes, heart disease, HIV/AIDS, arthritis, Alzheimer's, and many more.



The above diagram explains the activities performed by different parties involved in the claim filing, approval and rejection process.

4. Project Goal

4.1 The potential impacts of Healthcare Fraud on society

- Waste of funds that would have been otherwise used for providing better medical treatments or services to actual patients.
- It is a criminal act as patients were either given false medicines or procedures which were not required.
- It increases the overall expenditure on Healthcare and it returns as a burden on the insured user because private prayers increase their premiums.
- The false money earned by doing such frauds has also been used for carrying out various illegal activities that can be potentially harmful either to the nation or the entire world. Healthcare fraud detection involves account auditing and detective investigation. And, billing for healthcare is the complex part because of multiple aspects with vast amounts of data. \newline

The goal of this project is to "predict the potentially fraudulent providers" based on the claims filed by them. Along with this, we will also discover important variables helpful in detecting the behavior of potentially fraud providers. further,we will study fraudulent patterns in the provider's claims to understand the future behavior of providers.
Rigorous analysis of the Medicare data has yielded many physicians who indulge in fraud. They adopt ways in which an ambitious diagnosis code is used to bill the costliest procedures and drugs.Insurance companies are the most vulnerable institutions impacted due to these bad practices.

5. Data Description and Model

For the purpose of this project, we are considering Inpatient claims, Outpatient claims and Beneficiary details of each provider. Lets see their details :

- Inpatient Data
    1. Claim ID
    2. Beneficiary ID
    3. Provider ID
    4. Admission Date
    5. Discharge Date
    6. Attending Physician
    7. Operating Physician

This data provides insights about the claims filed for those patients who are admitted in the hospitals. It also provides additional details like their admission and discharge dates and admit d diagnosis code.

- Outpatient Data
    1. Claim ID
    2. Beneficiary ID
    3. Provider ID
    4. Attending Physician
    5. Operating Physician

This data provides details about the claims filed for those patients who visit hospitals and not admitted in it.

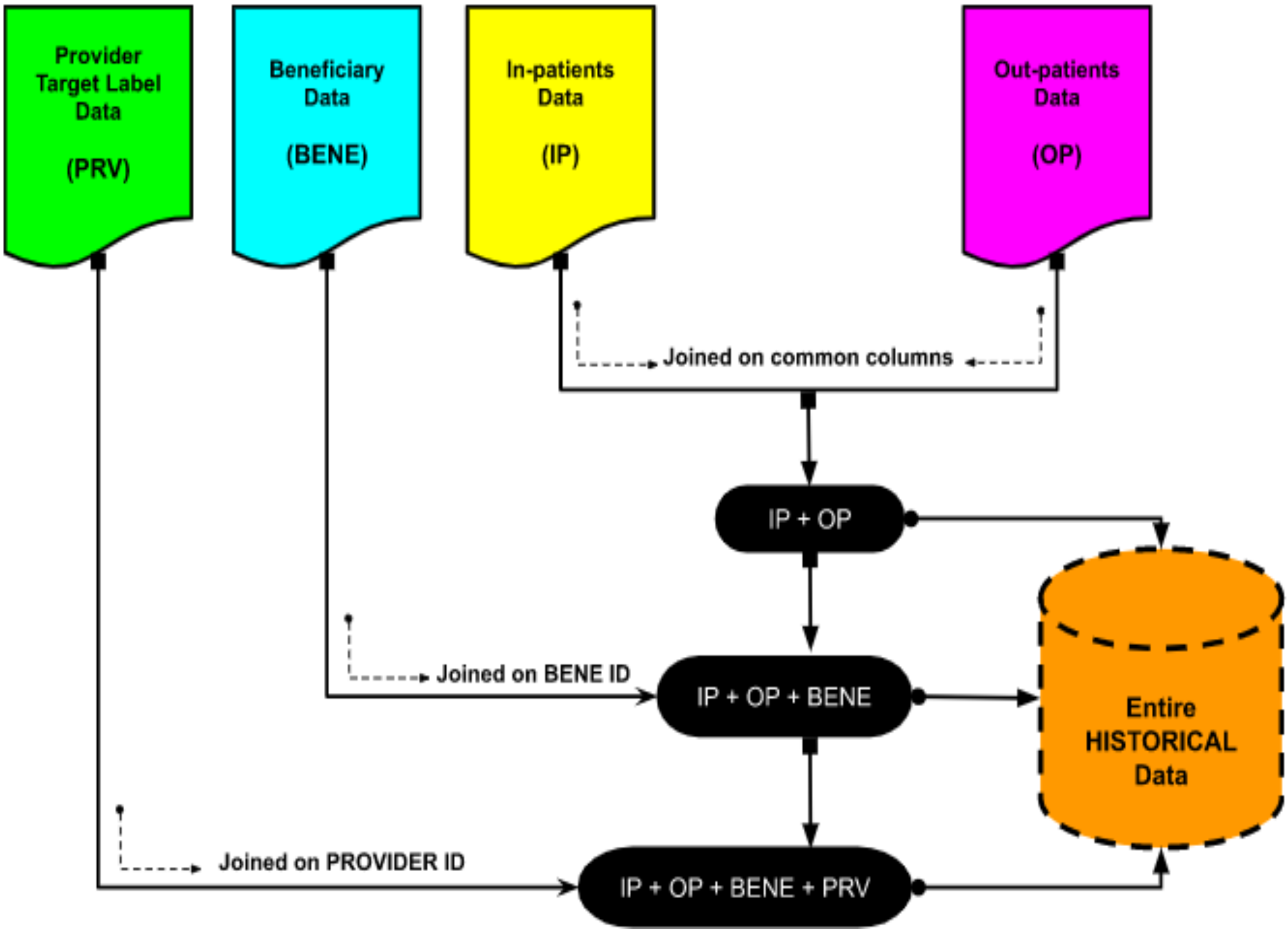
- Beneficiary Details Data
    1. Beneficiary ID
    2. Date of Birth
    3. Date of Death
    4. Gender
    5. Race
    6. Chronic Disease
    7. Reimbursed Amount
    8. Deductible Amount

This data contains beneficiary KYC details like health conditions, region they belong to, chronic condition indicators, gender, etc.

- Provider Data
    1. Provider ID
    2. Potential fraud?

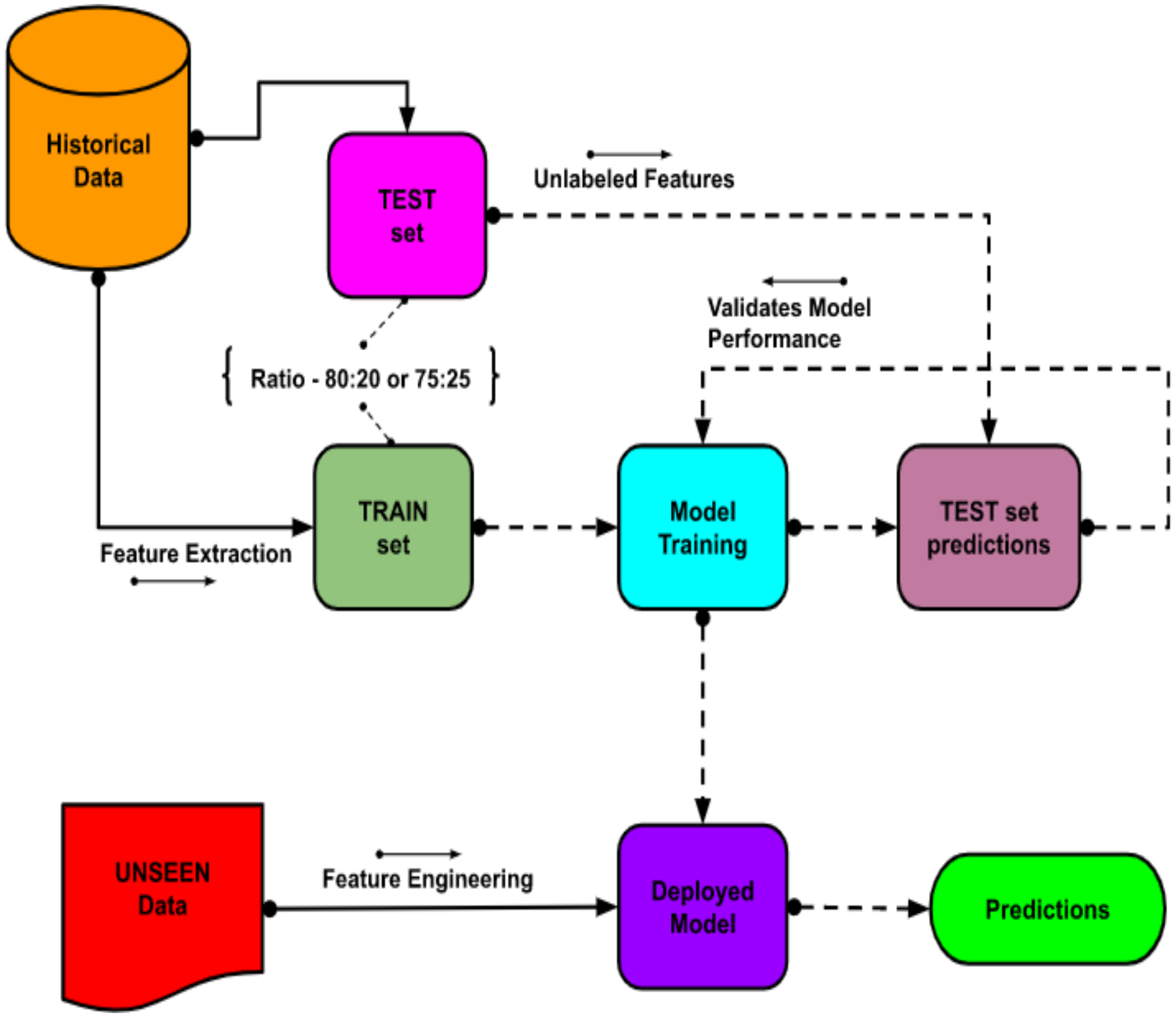
- For the 80:20 train and test ratio, class weighing scheme is used to deal with the imbalance problem.
- For 75:25 train and test ratio, synthetic minority class oversampling is used to deal with imbalance problem.

The below diagram illustrates how the information provided at different levels gets joined:



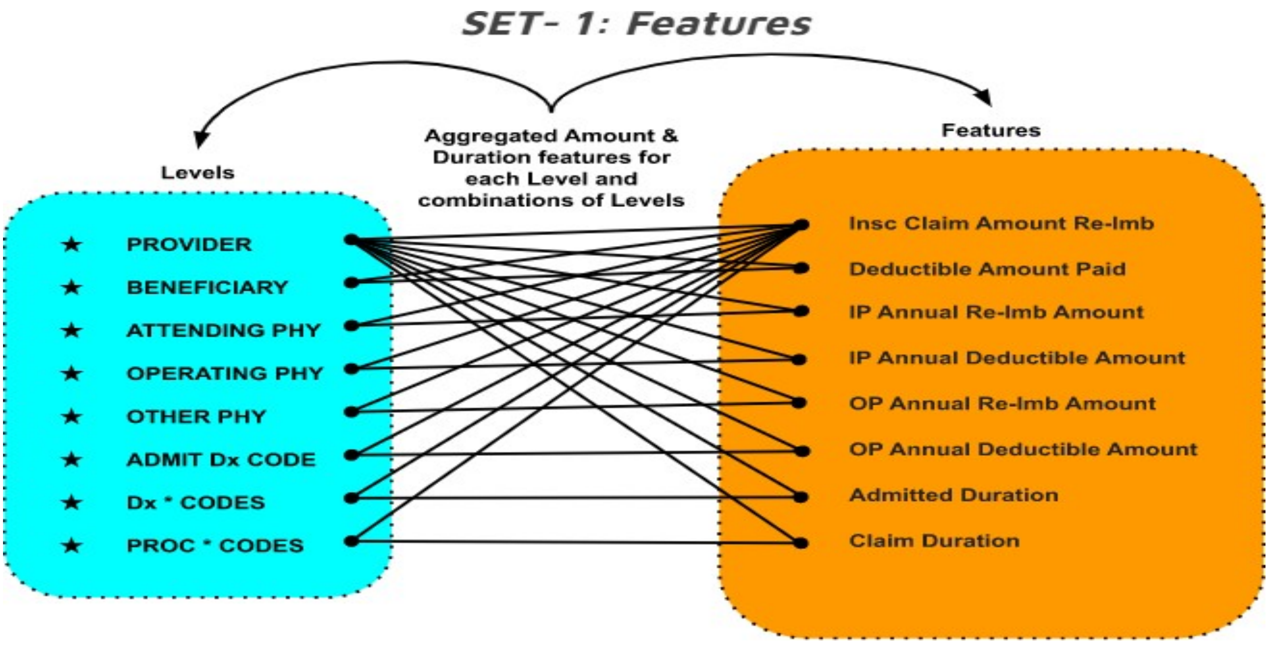
Inpatient and Outpatient data are joined based on common columns, then the resultant matrix is joined with beneficiaries' details based on Bene ID. Finally, it gets the provider's potentially fraudulent information using Provider ID.

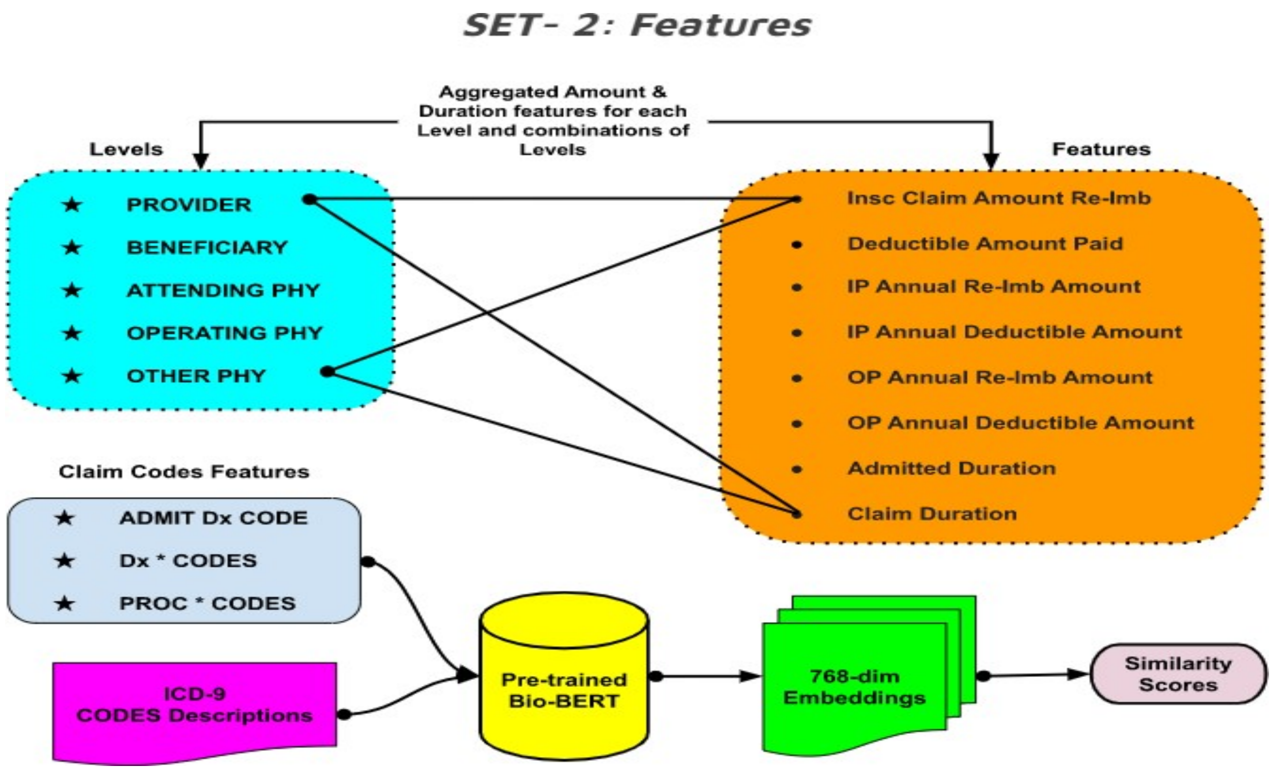
5.1 Model training and evaluation design



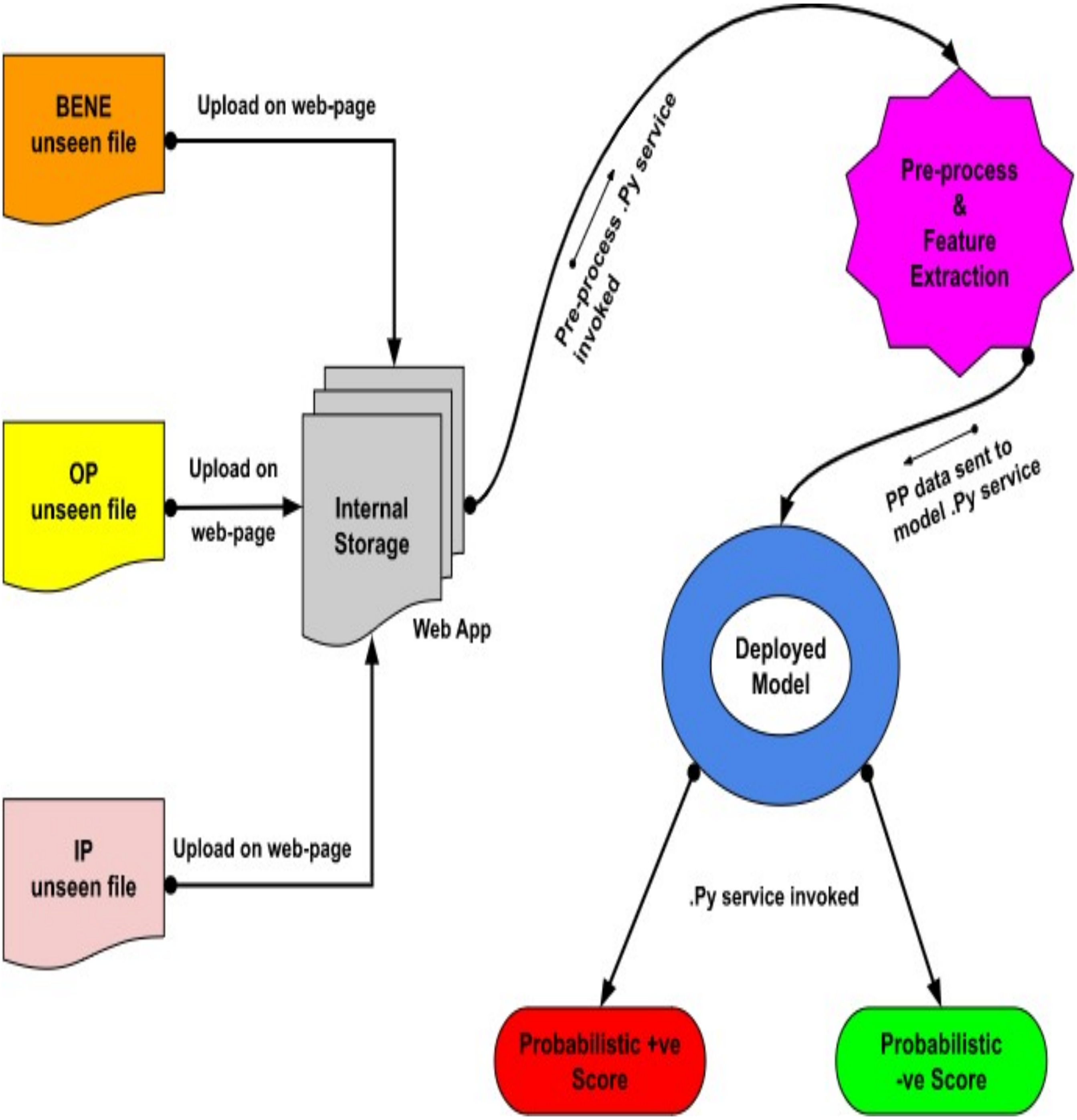
6. Experiment Setting and Description

6.1.Feature Engineering Design





6.2.Deployed Model Pipeline Design



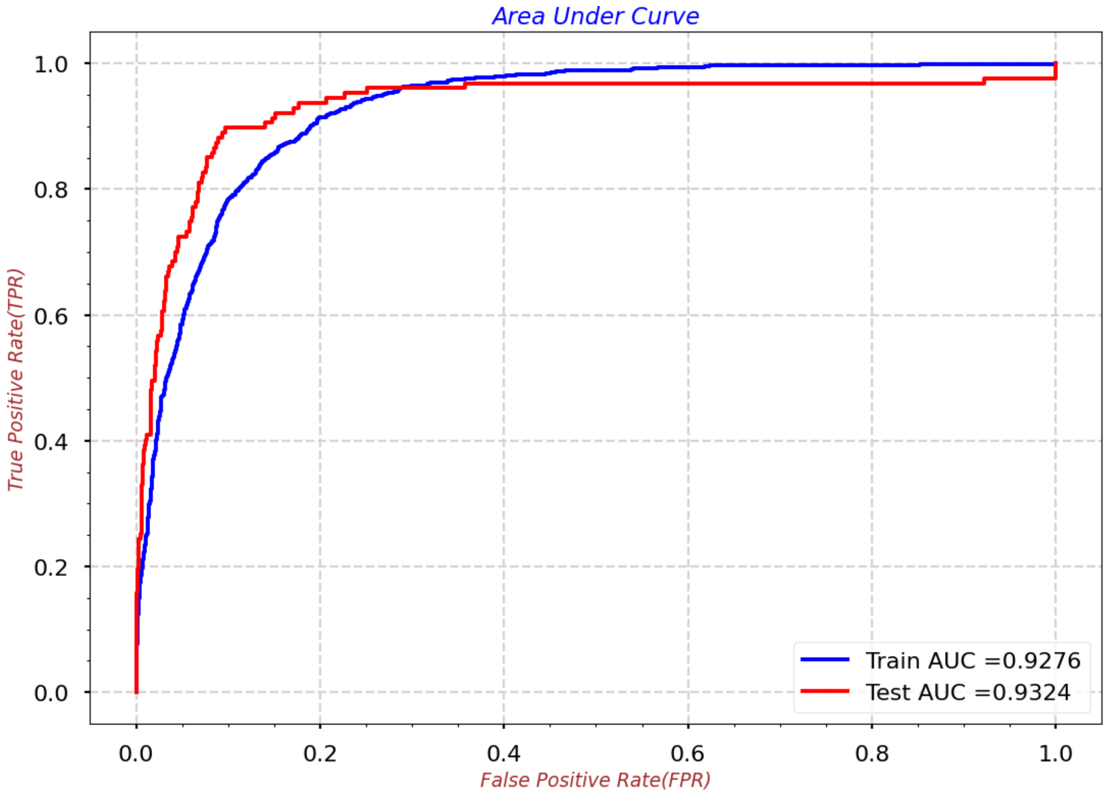
The flowchart shows how a deployed model pipeline will work on a future unseen data set

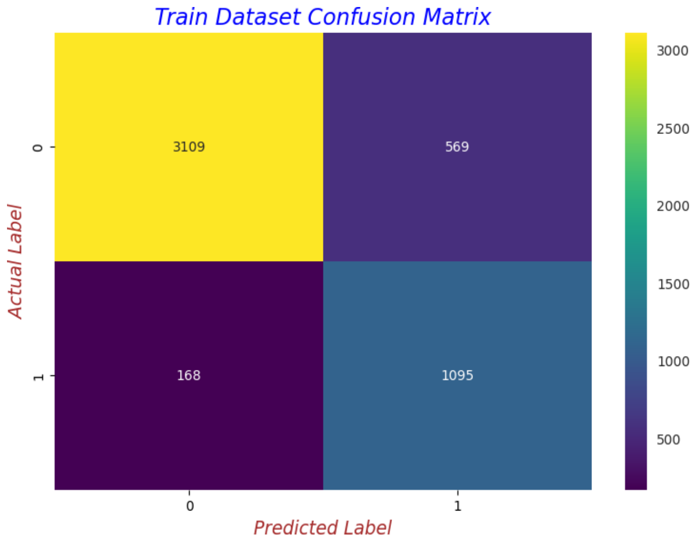
6.3.Model Results and Observations

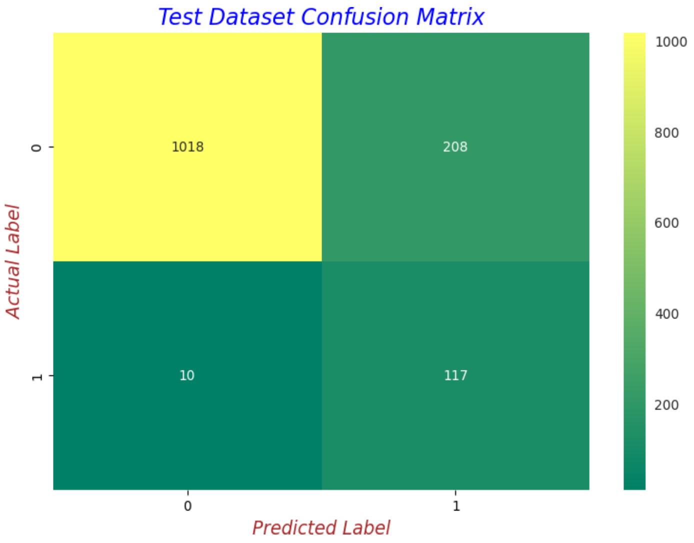
6.3.1.Logistic Regression

Logistic regression is a process of modeling the probability of a discrete outcome given an input variable. The most common logistic regression has a binary outcome, in our scenario, we are getting the outputs regarding if the particular patient/hospital is a fraud or not.

The graphs and confusion matrix below shows the accuracy of the model used on our dataset and the numbers of false positives or negatives.

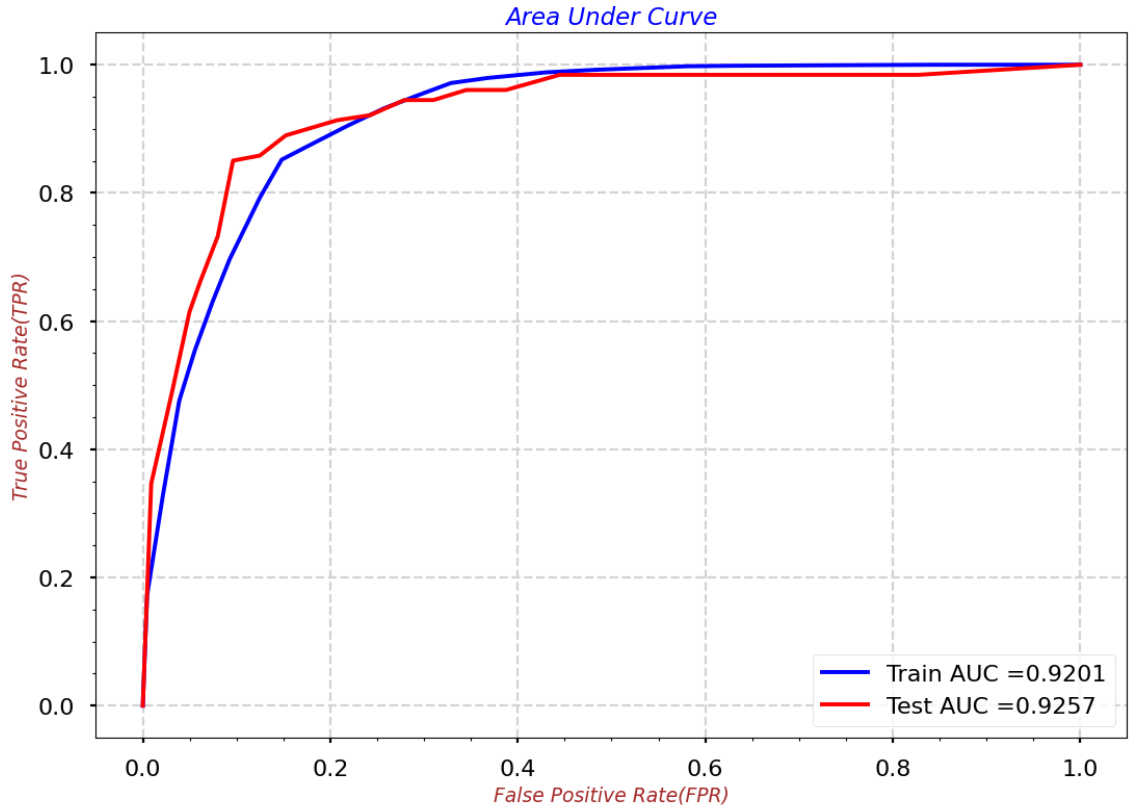


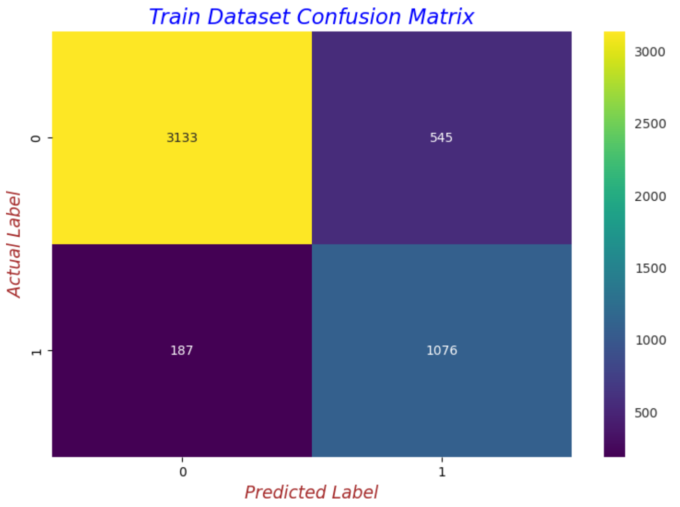


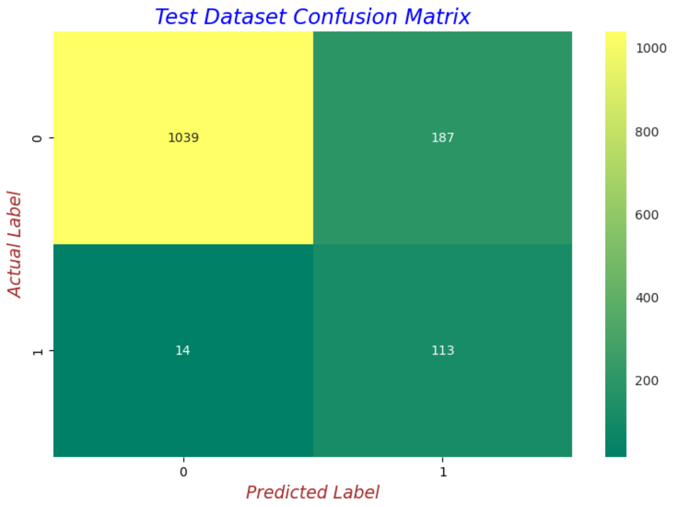


6.3.2.Decision Tree

Decision Trees are a non-parametric supervised learning method used for classification and regression. The goal is to create a model that predicts the value of a target variable by learning simple decision rules by taking the input features. The graphs for our Decision tree model is given below:



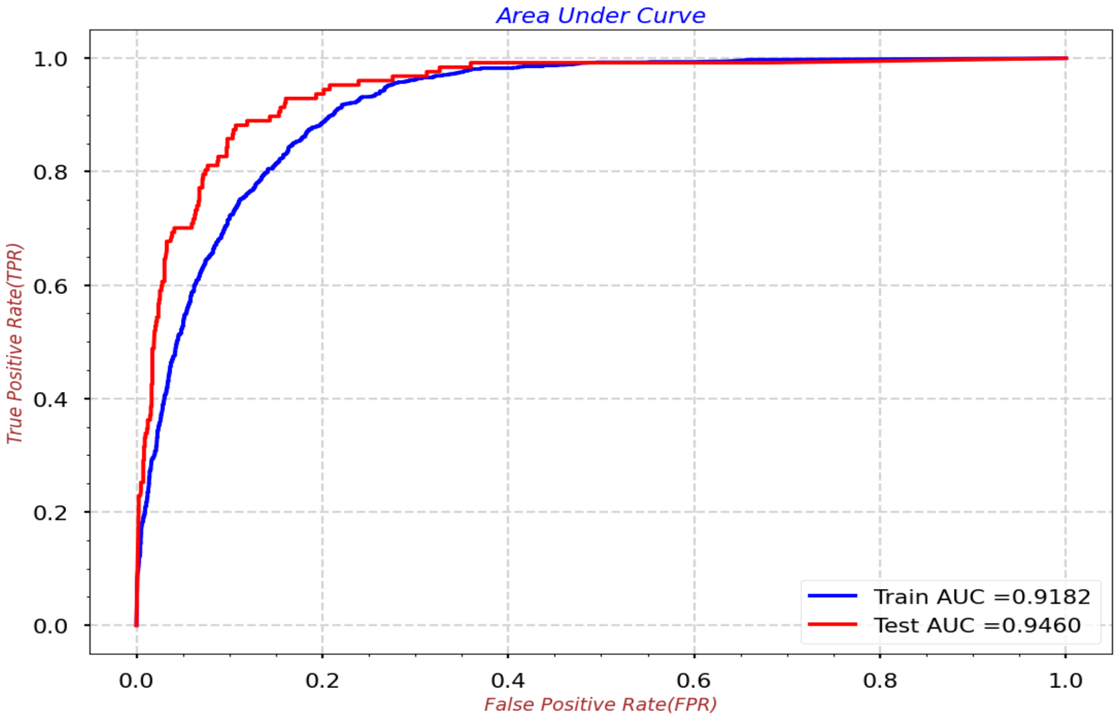


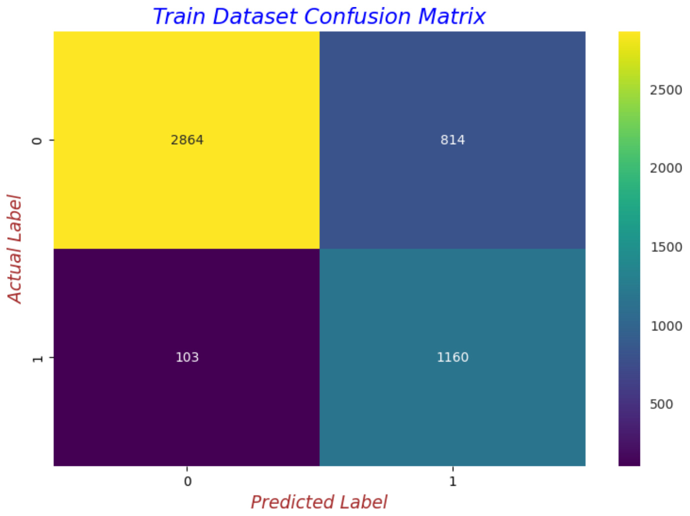


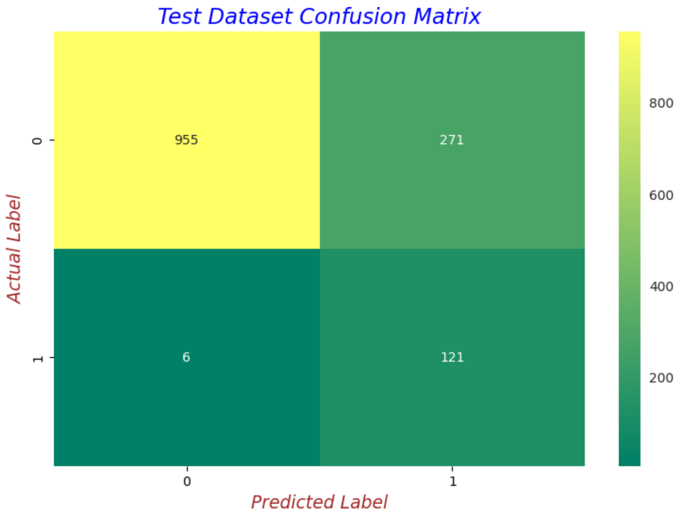
6.3.3.Random Forest Classifier

A random forest classification takes multiple decision trees on various sub samples and uses average to improve the accuracy of the model and control over-fitting.

The graphical representation for the accuracy of this model is given below:







6.3.4.Observations

- Adding the Similarity Score features of embeddings between below mentioned doesn’t really help in improving the model's performance.
- Doing the synthetic oversampling of the minority class doesn’t provide a gain in the model’s performance whereas we can see a noticeable drop in the performance.

6.4.Description (code working)

- First all the data is loaded starting from the dataset of beneficiaries with diseases and then segregate the number of beneficiaries with chronic and non chronic diseases and recognize how many are fraud and how many are not.
- The average amount reimbursed for various people with different diseases is checked and then the in-patient and out-patient data is classified accordingly. The data is cleaned and processed. Percentiles for pre-diseases are calculated and generated.
- New columns are created, data is analyzed and feature engineering is done i.e.,
    1. The impact of Amount(Insc, IP, OP, and Deductible), Discharge period, Claim settlement period features on the fraudulent claim ratio is evaluated.
    2. Impact of treatment days, gender, race, state, country, renal disease and chronic condition on fraudulent ratio are checked.
    3. Also, the relationship between diagnosis group code, claim diagnosis and procedure code is checked as there might be a case that a patient might be billed for the services which were not even provided.
- Once all the data is collected, it is mapped in different ways to get the best way to find fraud cases or an idea on how to classify the data and which field has relatively more weight than the other.

- All the individual tables are merged into one so that the data can be trained and tested using machine learning algorithms.

  Models used: Logistic Regression
  Decision Tree
  Random Forest Classifier
- Confusion matrix (it is the table where TP, FP, TN, and FN counts will be plotted. From this table, we can visualize and track the number of mistakes made by the model). is created for each model

7. Conclusion

In this project, information regarding healthcare frauds, types and sources of health care fraud data is collected. The major part of the data comes from governmental resources and private insurance companies. Mainly, we came to know how machine learning and data mining are used for Healthcare fraud detection. We also learn about Supervised, unsupervised and semi-supervised learning approaches in Machine learning. In most of the cases, semi-supervised learning approaches are used by many researchers. But, to detect frauds in the healthcare system more efficiently, new semi-supervised learning approaches can be proposed in a few cases. But, to conceal all the instances of healthcare fraud, there doesn’t exist any particular standard approach or patterns. It can be concluded from this review that the advanced machine learning techniques and newly acquired sources of the healthcare data would be forthcoming subjects of interest in order to make the healthcare affordable, to improve the effectiveness of healthcare fraud detection and to bestow top quality on healthcare systems.

7.1.Future Scope

After reviewing different studies on healthcare fraud detection, it can be concluded that frauds or abuse that occur in health insurance systems can be of different unusual patterns. To detect such suspicious patterns more research work is needed by using advanced data mining and machine learning techniques. There is a need to propose new methods while considering minute details of healthcare data. To achieve this, correlations between different entities of healthcare data can be taken into account.

8. Code Link

https://colab.research.google.com/drive/11LQ0oR1yH9sVWr9k_UwTtguX0sdl7jDc?usp=sharing#scrollTo=b2d12961

(since the size of code is large sometimes it might take a while to load)

9. References

- https://www.kaggle.com/datasets/rohitrox/healthcare-provider-fraud-detection-analysis

- https://towardsdatascience.com/building-a-logistic-regression-in-python-step-by-step-becd 4d56c9c8

- https://www.dataquest.io/blog/sci-kit-learn-tutorial/

- https://imbalanced-learn.org/stable/references/generated/imblearn.over_sampling.ADASY N.html#imblearn.over_sampling.ADASYN