FINAL PROJECT REPORT
HEALTHCARE PROVIDER FRAUD DETECTION

1.OBJECTIVE:

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.

2.ABSTRACT:

Healthcare provider fraud poses a significant challenge for Medicare, leading to a substantial surge in total Medicare spending as a consequence of fraudulent claims. This form of organized crime involves collaboration among providers, physicians, and beneficiaries to coordinate fraudulent claims. Through thorough examination of Medicare data, numerous physicians engaging in fraudulent activities have been identified. These practitioners employ tactics such as utilizing vague diagnosis codes to endorse costly procedures and drugs. The adverse impact of these deceptive practices is particularly felt by insurance companies, prompting them to raise insurance premiums. Consequently, healthcare costs are escalating, contributing to the increasing financial burden on individuals.Various manifestations of healthcare fraud and abuse are observed, with providers commonly engaging in deceptive practices such as:

a) Submitting claims for services that were never rendered.

b) Repetitively submitting claims for the same service.

c) Falsifying the description of the service provided.

d) Charging for a service of higher complexity or cost than the one actually delivered.

e) Billing for a service covered by insurance when the service provided was not covered.

3.DESIGN & IMPLEMENTATION:

3.1.Dataset Overview:

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

A) Inpatient Data: 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 diagnosis code.

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

C) Beneficiary Details Data: This data contains beneficiary KYC details like health conditions,region they belong to etc.

3.2.Data Wrangling:

First we loaded all the data and took a look into the dataset. Cleaned each dataset and made the data ready for analysis and then merged them into a finalized dataset. For this we first changed the target labels of train_y data into 0 and 1, and checked for null values. Then likewise explored each dataset. Then merged inpatient and outpatient data, then merged the IP_OP dataset with the beneficiary data and finally merged IP_OP_Bene data with the train_y data to get a finalized dataset. Also, we have done some visualizations and plotted features for exploratory analysis.

3.3.Exploratory Data Analysis:

In the Exploratory Data Analysis of Healthcare Provider Fraud Detection dataset, we have seen the relationship between various features by plotting and analyzing the data and visualizations.

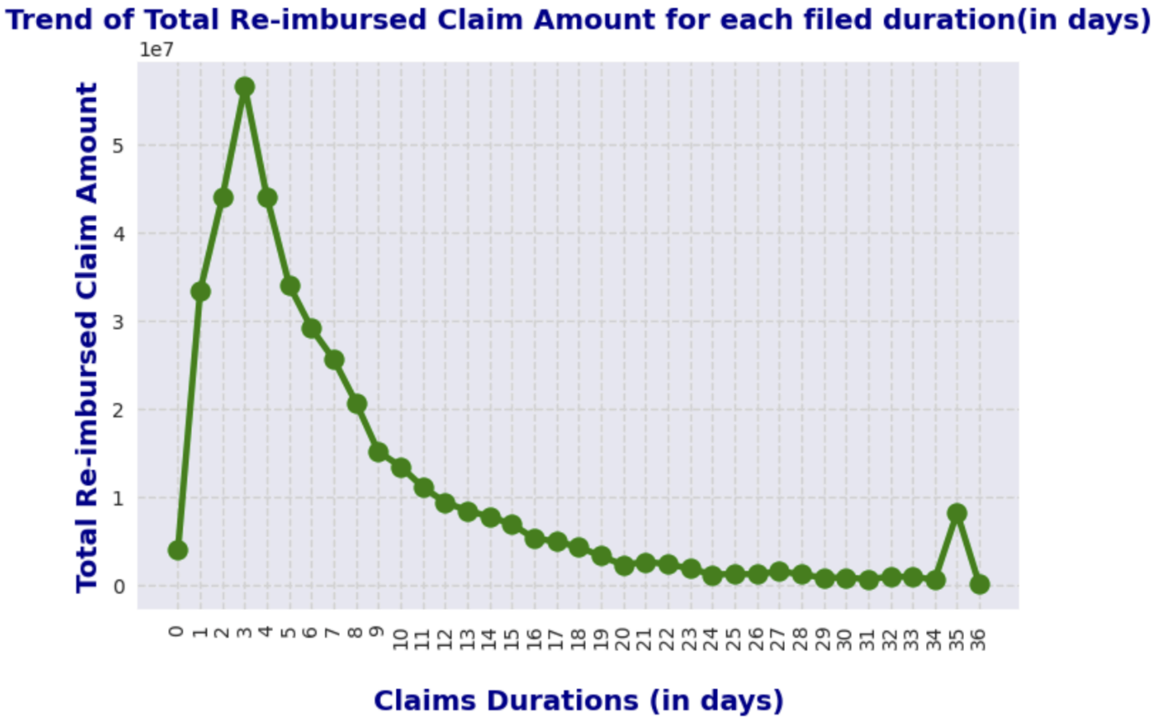

Fig.3.3.1

The graphical representation (Fig. 3.3.1) of Total Reimbursed Claim Amount for each filed duration for inpatients tells us that the Total Reimbursed Amount is the highest for 3 days claims. Also, for claims with durations from 12 to 34 the total reimbursed amount is very less, however, for 35 days duration we can witness a clear spike that can be a potential sign of fraud. We can see a few findings when checked the relationship between Claimed and Admitted Durations with Reimbursed Amount, that there are 49 claims whose Claimed Duration and Admitted Duration are different. The total reimbursed amount is around 0.67 Million. 17 claims out of 49 have Claimed Duration greater than the Admitted Duration. And, for these 17 claims the total reimbursed amount is around 0.27 Million. Also, 32 claims out of 49 have Admitted Duration greater than the claimed Duration. And, for these 32 claims the

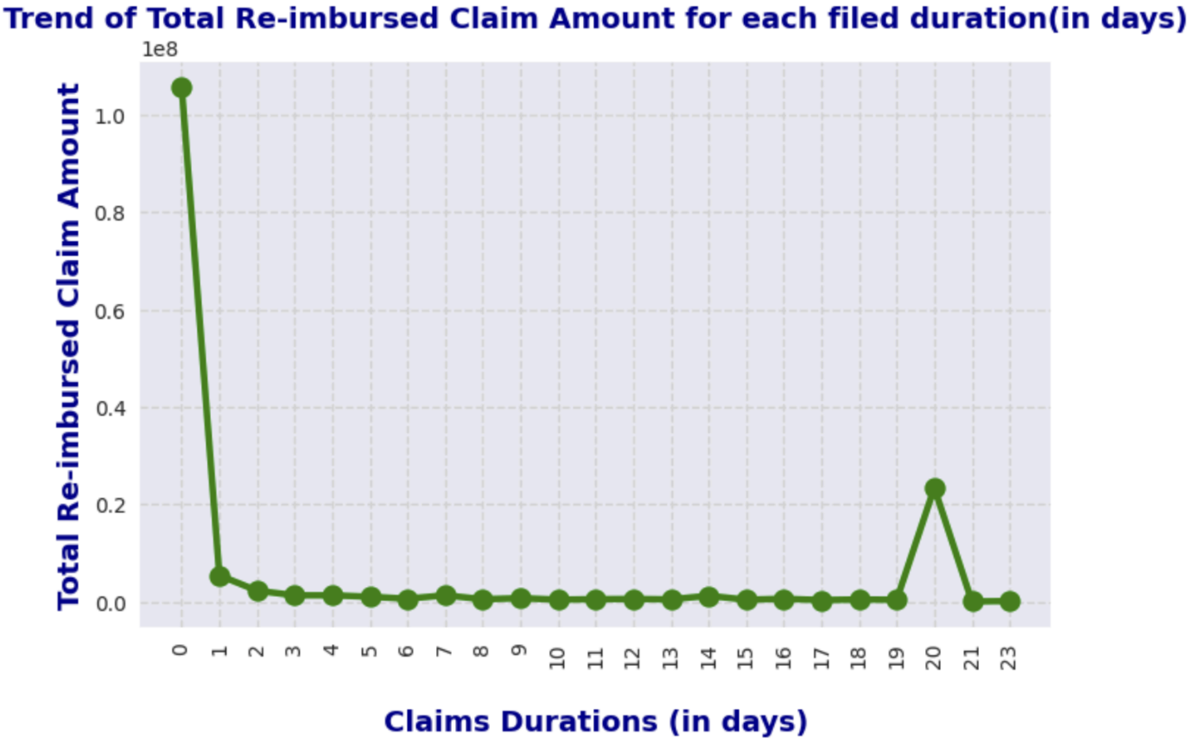

Fig.3.3.2

total reimbursed amount is around 0.39 Million. When checked the relationship between DeductibleAmtPaid and Reimbursed Amount, we found that there are a 2% of total claims for which there is no co-payment.And, for these 2% (or 899) of total claims the total reimbursed amount is 10.6 Million that is 2.6% of the total reimbursed amount. Also, analyzed that the Provider Ids who filed 1 or 2 claims got the entire amount reimbursed by checking the relationship of Providers with Total number of claims filed & Reimbursed Amount. This can be a potential sign of fraud because many small-small hospitals in rural areas, who don't have much facilities or equipment, made fraud for benefits.

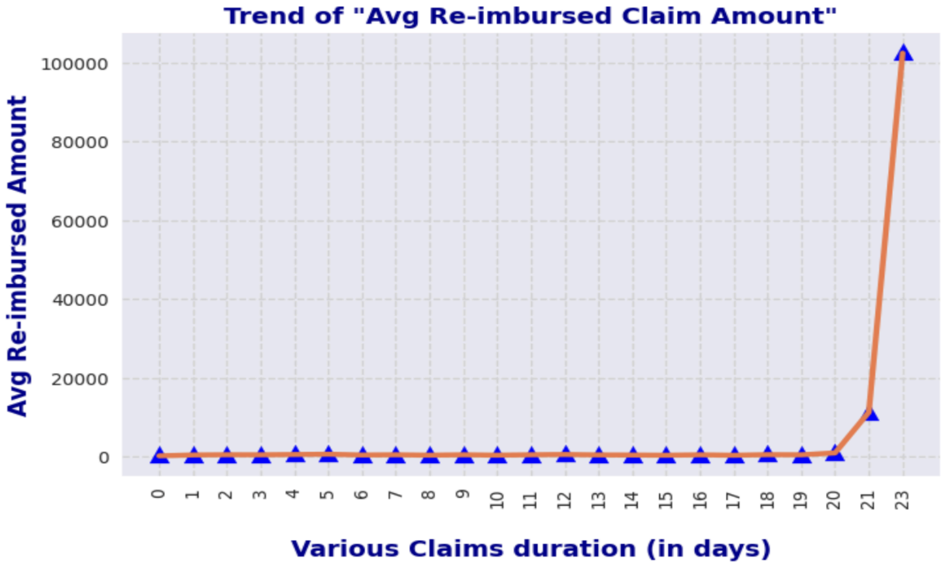

Fig.3.3.3

The graphical representation (Fig.3.3.2) of Total Reimbursed Claim Amount for each filed duration for outpatients tells us that the majority of the claims filed for less than or equal to 2 days and the Total Reimbursed Amount is the highest for 0 days claims. Also, for claims with durations from 2 to 19 the total reimbursed amount is very less or similar, however, for a 20 days duration we can witness a clear spike that can be a potential sign of fraudulent. In (Fig.3.3.3) the Average Reimbursed Amount is the same throughout the various durations except for 21 and 23 days.

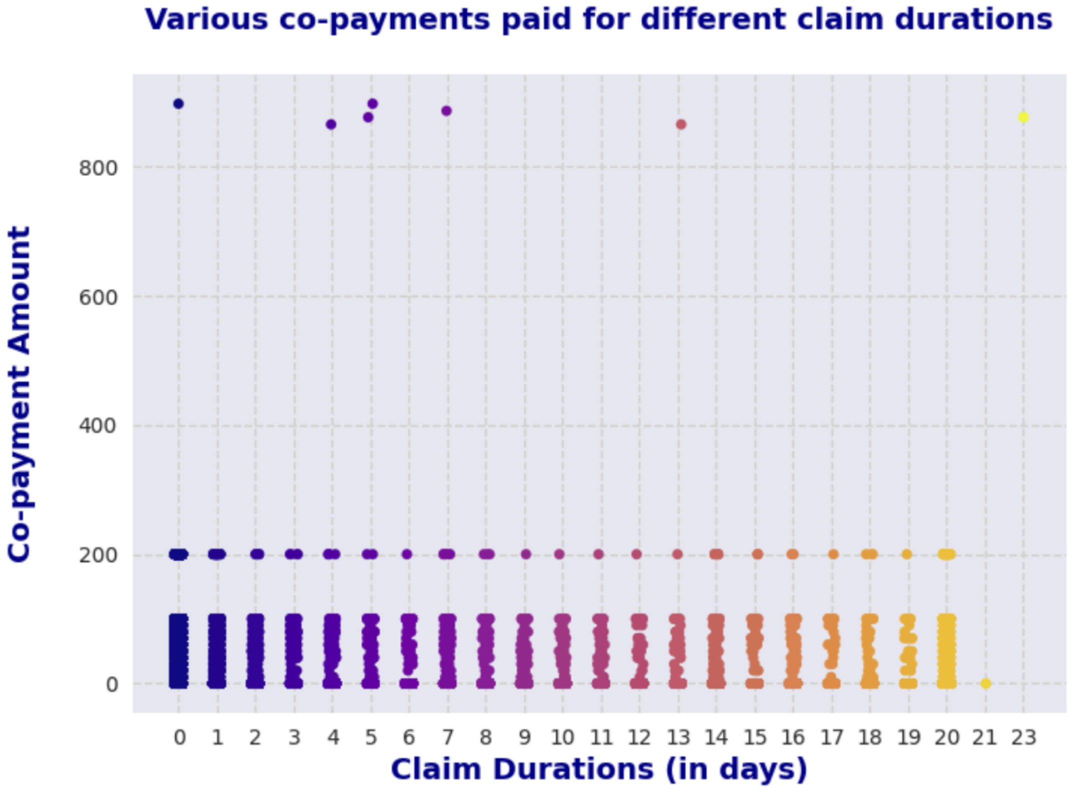

Fig.3.3.4

When we plot (Fig.3.3.4) for various co-payments paid for different claim durations we can deduce that the co-payment is similar across the various durations, most of them with 200, however, there are few co-payment which are very high or more than 800. Also checked the relationship between DeductibleAmtPaid and Reimbursed Amount, the findings tells us that there are 95% of total claims for which there is no co-payment. And, for these 95% of total claims the total reimbursed amount is 142.3 Million. Similar to inpatient data, the Provider Ids who filed 1 or 2 claims got the entire amount reimbursed by checking the relationship of Providers with Total number of claims filed & Reimbursed Amount for the outpatients data.

The below box plot (Fig.3.3.5) suggest that if PRV_Insc_Clm_ReImb_Amt is high then it increases the chances of fraud. After merging the datasets we checked the Distribution of Fraud & Non-fraud claims and found that 62% of claims are Non-Fraud and 32% of them are Fraudulent (Fig.3.3.6). If the Insurance Claim ReImb_Amt for provider is high then it increases the chances of fraud. Added aggregated features at different levels (Provider, Beneficiary, Attending Physician, Operating Physician, Other Physician, etc.) assuming that fraud can be done by an individual or group of individuals or entities involved in the claim process. The

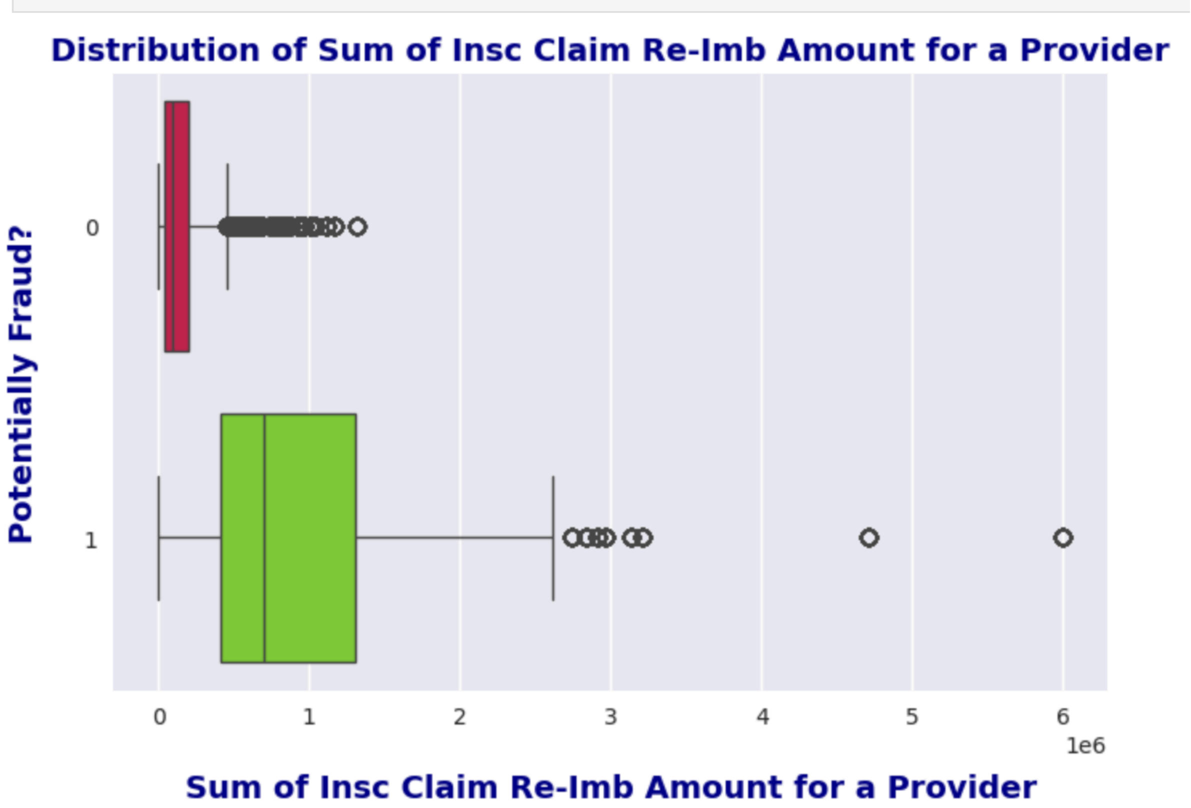

Fig.3.3.5

merged dataset is ready for preprocessing after removing unwanted columns and assuming that the features total reimbursed amount, claim duration, admitted duration, deductible amount paid, along with aggregated features will contribute to the prediction of fraud and non-fraud providers in the healthcare industry.

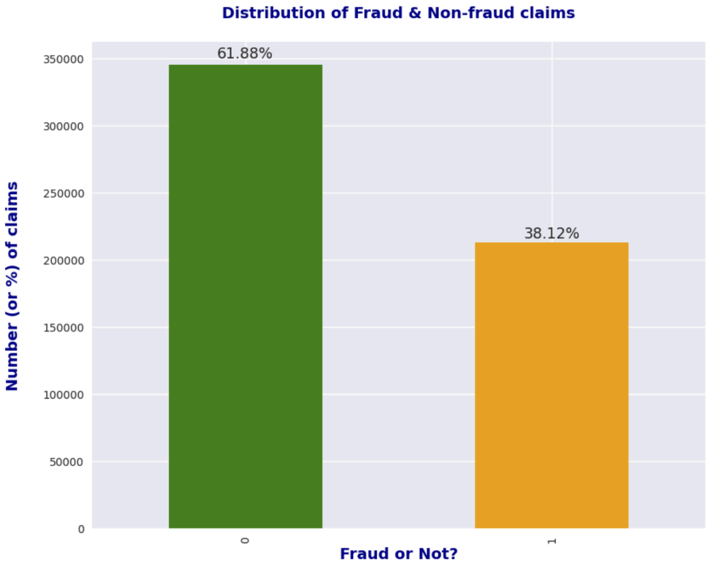

Fig.3.3.6

3.4.Preprocessing:

The preprocessing phase is an important step in developing an effective healthcare provider fraud detection model. This report focuses on the preprocessing steps related to Train-Test Split and Standardization, using the RobustScaler from the scikit-learn library. These preprocessing steps set the foundation for subsequent stages in the development of an accurate and reliable fraud detection system.

3.4.1.Train-Test Split

In order to evaluate the performance of the fraud detection model, it is essential to split the dataset into training and validation sets. This separation allows us to train the model on one portion of the data and validate its performance on another. Here, 80% of the data is used for training, and 20% is reserved for testing.

3.4.2. Robust Scaling

To address potential outliers and variations in the data, the RobustScaler from scikit-learn's preprocessing module is employed. Robust scaling is particularly useful in the context of healthcare data where outliers may exist due to variations in medical conditions and billing practices. Robust scaling is applied to the features to ensure that the model is less sensitive to outliers and can better handle variations in the data.

3.5.Modeling:

The modeling phase is pivotal in the development of a healthcare provider fraud detection system. This section of the documentation focuses on the implementation and evaluation of three distinct models: Logistic Regression, Decision Tree, and Random Forest. This includes model selection, hyperparameter tuning, cross-validation, training & evaluation and finally model comparison.

3.5.1.Model Selection & Hyperparameter tuning

Model selection is a critical step that involves choosing the most appropriate machine learning algorithm for the specific task. In this project, three models: Logistic Regression, Decision Tree, and Random Forest are considered.

3.5.1.1.Logistic Regression

Logistic Regression is chosen as it is well-suited for binary classification problems, which is often the case in fraud detection. Models the probability of a binary outcome based on one or more predictor variables. It is a linear model with a logistic function that transforms the linear combination of predictors into a probability.

3.5.1.2.Decision Tree

Decision Trees are selected for their ability to capture complex decision-making processes. Partition the feature space into regions based on feature values. The tree structure represents a series of decisions, leading to a final outcome.

3.5.1.3.Random Forest

Random Forest is an ensemble method that combines multiple decision trees to improve predictive accuracy and control overfitting. Builds a multitude of decision trees and merges their predictions. It introduces randomness in the tree-building process to enhance diversity among the trees.

3.5.2.Hyperparameter Tuning

Hyperparameter tuning involves selecting the optimal values for the hyperparameters of a model to improve its performance.

3.5.2.1.Logistic Regression

The ‘`solver`’ and ‘`max_iter`’ are tuned. The 'liblinear' solver is chosen for logistic regression, and ‘`max_iter`’ is increased to ensure convergence.

3.5.2.2.Decision Tree and Random Forest

Parameters such as tree depth, minimum samples per leaf, and the number of trees in the forest are tuned.Cross-validation is employed to assess the model's performance under various hyperparameter configurations.

3.5.3.Cross-validation

Cross-validation assesses the model's generalization performance by splitting the training data into multiple folds. Each model is cross-validated using 5-fold cross-validation. It involves partitioning the training data into folds, training the models on subsets, and validating on the remaining data to ensure robustness.

3.5.4.Training and Evaluation

The model is trained on the transformed training data, and predictions are made on the transformed test data. Evaluation metrics such as accuracy, precision, recall, and F1-score are calculated to assess the performance of each model individually.

Evaluation Metrics:

- **Accuracy:** The proportion of correctly classified instances out of the total instances.
- **Precision:** The ratio of true positive predictions to the total predicted positives.

- **Recall):** The ratio of true positive predictions to the total actual positives.
- **F1-Score:** The harmonic mean of precision and recall, providing a balance between the two metrics.

Model Metrics

Models	Accuracy	Precision	Recall	F1-Score
Logistic Regression	0.926	0.676	0.457	0.545
Decision Tree	0.889	0.429	0.438	0.433
Random Forest	0.926	0.704	0.409	0.518

Table 3.5.4.1

Model performances are compared based on evaluation metrics. Each model's strengths and weaknesses are identified, providing insights into their suitability for fraud detection in healthcare provider data.

The accuracy, recall and f1 score of Logistic Regression is greater than all other models (Table 3.5.4.1). Hence we can conclude that the best performing model among these models is Logistic Regression with an accuracy of 0.926 and F1-Score: 0.545.

3.5.5.Model Comparison

We conducted a feature importance analysis on the best-performing model, Logistic Regression. This analysis allows us to identify the features that have the most significant impact on the model's predictions. The feature importances were calculated and visualized to provide insights into the relative importance of each feature. The analysis identified the top features that exert the most influence on the model's predictions. These features include InscClaimAmtReimbursed,ClmCount_Provider_BeneID_OperatingPhysician_ClmDiagnosisCode_4,ChronicCond_Heartfailure,ClmCount_Provider_BeneID_ClmDiagnosisCode_2, ChronicCond_Alzheimer,ClmCount_Provider_BeneID_ClmProcedureCode_1, ClmCount_Provider_DiagnosisGroupCode are the first seven features with highest feature importance value greater than 1, which consistently contributed to the accuracy and robustness of the model.

A bar (Fig.3.5.5.1) chart was generated to visually represent the importance scores of the top features. This visualization aids in the quick interpretation of which features play a pivotal role in the model's decision-making process.

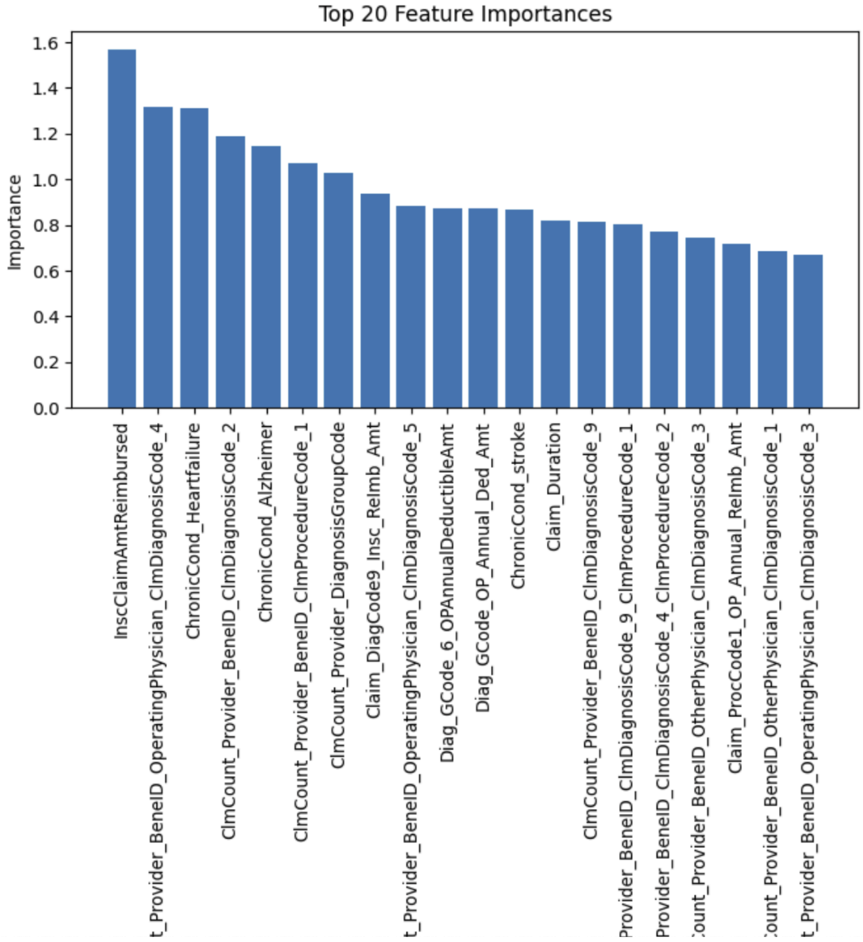

Fig.3.5.5.1

The feature importances offers valuable insights into the underlying patterns within the dataset. Features with higher importances suggest a stronger association with the target variable and contribute significantly to the model's ability to distinguish between classes.

4.CONCLUSION:

In conclusion, the healthcare provider fraud detection project has undergone a comprehensive modeling and evaluation process, aiming to develop a robust predictive model for identifying fraudulent activities within healthcare claims. We employed various machine learning algorithms, including Logistic Regression, Decision Tree, and Random Forest, and evaluated their performance based on key metrics such as accuracy, precision, recall, and F1-score.After thorough analysis, it is evident that the Logistic Regression model performed best, demonstrating superior predictive capabilities. Feature importance analysis highlighted critical factors contributing to the model's decision-making process, providing valuable insights into the intricate patterns within the dataset.

5.FUTURE WORK:

In Future, we can explore advanced feature engineering techniques to extract more meaningful insights from the data. Incorporating domain-specific knowledge and creating new features may contribute to further model refinement.Investigate the potential benefits of ensemble models and model stacking. Combining the strengths of multiple models could lead to improved predictive performance and robustness against diverse fraudulent patterns.Explore the feasibility of implementing real-time fraud detection systems. Investigate technologies and architectures that enable timely detection and response to fraudulent activities as they occur. Establish a framework for continuous model monitoring and updating. Fraud patterns may evolve over time, and a proactive approach to model maintenance ensures its relevance and effectiveness in detecting emerging threats.