**HEALTHCARE PROVIDER FRAUD DETECTION ANALYSIS**

Machine Learning (ML) offers a scalable solution, with its inherent ability for pattern recognition, consistency, and efficiency, as well as continuous improvement in fraudulent claim identification. Bayesian hyperparameter optimization algorithm can be used for predicting fraudulent health insurance claims. The primary focus of this methodology is the classification of authentic and fraudulent transactions. The results derived from the proposed model can be evaluated and contrasted with those obtained from other Machine Learning techniques, utilizing various performance metrics. By utilizing Bayesian optimization , we can efficiently tune the hyperparameters of machine learning models used to predict health insurance frauds, leading to better performance with fewer evaluations. This method is particularly useful in handling the complexity and high dimensionality often present in fraud detection tasks.

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 behaviour of potentially fraud providers. Further, we will study fraudulent patterns in the provider's claims to understand the future behaviour of providers.

**Dataset size:**

* Training data: 32958 records which was split into training, validation and test set
* Features: 54 columns (e.g., total\_claims, unique\_benefeciaries, inpatient\_days)
* Target variable: ‘PotentialFraud’ (binary: ’Yes’ or ‘No’)

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

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 admited 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.

The dataset includes following key features:

* TotalClaims: Number of claims submitted by the provider.
* InpatientDays: Total number of inpatient days claimed.
* UniqueBeneficiaries: Number of unique patients for the provider
* OutpatientDays: Total number of outpatient days claimed.
* TotalReimbursetmentAmount: Total money reimbursed by insurance.

The dataset is available to everyone on Kaggle under the name ‘HEALTHCARE PROVIDER FRAUD DETECTION ANALYSIS’. Its original contributor is Rohit Anand Gupta.

[HEALTHCARE PROVIDER FRAUD DETECTION ANALYSIS (kaggle.com)](https://www.kaggle.com/datasets/rohitrox/healthcare-provider-fraud-detection-analysis/code)

**Model:**

**Model Architecture:**

Model Type: Supervised Classification

Preprocessing Steps

1. Missing value Handling
2. Feature encoding
3. Scaling

Baseline Model

* Model: Random Forest Classifier
* Metrics:

Accuracy: 0.85

Precision: 0.88

Recall: 0.82

F1-Score: 0.85

Bayesian Optimization for Hyperparameter Tuning

To improve the model’s performance, we applied Bayesian Optimization for hyperparameter tuning in order to find the optimal values.

* Hyperparameters Tuned:

N\_estimators: Number of trees in Random Forest

Max\_depth: Maximum depth of each tree

Min\_samples\_split: Minimum number of samples required to split a node.

Min\_samples\_leaf: Minimum number of samples required at each leaf node.

* Bayesian Optimization Process:

Acquisition Function: Expected Improvement (EI)

Number of iterations: 50 iterations to maximize performance

Libraries: Scikit-optimize (skopt)

**Performance**

Metrics

After applying Bayesian optimization, the tuned Random Forest model demonstrated improved performance:

* Accuracy: 0.88
* Precision: 0.90
* Recall: 0.87
* F1-Score: 0.89
* ROC AUC Score: 0.91

Feature importance

The following features were found to be the most important in determining healthcare fraud:

1. TotalClaims
2. UniqueBeneficiaries
3. TotalReimbursementAmount
4. InpatientDays
5. OutpatientDays

