**IS597MLC: Final Project Proposal**

**NetID: abulatov**

**Student Name: Adilbek Bulatov**

**Title**

***Healthcare provider fraud detection analysis***

**Motivation & Objective**

**Background information**

Provider Fraud is one of the biggest problems facing Medicare. According to the government, the total Medicare spending increased exponentially due to frauds in Medicare claims. Healthcare fraud is an organized crime which involves peers of providers, physicians, beneficiaries acting together to make fraud claims. Fraudulent activities can take many forms, including billing for services not rendered, upcoding (charging for a more expensive service than was provided), and misrepresenting dates or providers of service. The detection and prevention of such fraud require sophisticated analytical techniques to identify anomalies and suspicious patterns within large datasets. Machine learning (ML) techniques are increasingly being applied to tackle this challenge due to their ability to recognize complex, non-linear patterns and relationships in data that may not be apparent through traditional rule-based systems.

**Motivation and objective**

The main objective of this project is to build a machine learning model capable of identifying healthcare providers that are likely to be involved in fraudulent activities. This objective will be achieved by analyzing the relationships between claim details, provider behaviors, and patient characteristics. The model will help insurers prioritize investigative efforts by flagging high-risk providers for further scrutiny. Additionally, insights derived from the analysis may offer strategic guidance to healthcare policymakers and insurance companies on how to strengthen fraud prevention measures.

**Research Questions**

1. **What features within healthcare insurance claims data are most indicative of fraudulent activity?** This question aims to identify which factors, such as diagnosis codes, procedure codes, or claim amounts, have the strongest influence on the likelihood of a provider being flagged as fraudulent.
2. **Can machine learning models effectively predict fraudulent providers, and which model performs best in terms of accuracy, precision, and recall?** This question will evaluate and compare the performance of different machine learning models, such as decision trees, random forests, and gradient boosting models, in detecting fraudulent claims.

**Related Articles**

1. "Predicting health insurance claim frauds using supervised machine learning" (2023).

<https://ieeexplore.ieee.org/document/10142604>

This paper proposes a machine learning model that utilizes appropriate methods to address issues in health insurance claims, focusing on supervised learning techniques.

1. "Machine Learning for Health Insurance Fraud Detection: Techniques, Insights, and Implementation Strategies" by Harish Narne (2024).

<https://www.researchgate.net/publication/386384259_Machine_Learning_for_Health_Insurance_Fraud_Detection_Techniques_Insights_and_Implementation_Strategies>

This paper explores various machine learning techniques for fraud detection, including supervised learning algorithms such as decision trees and neural networks. It discusses implementation strategies, challenges, and future directions in the field.

1. "Health insurance fraud detection based on multi-channel heterogeneous graph structured learning" (2024).

<https://pmc.ncbi.nlm.nih.gov/articles/PMC11061682/>

This study proposes a novel approach called MHGSL, which combines supervised learning with graph neural networks to capture complex relationships in healthcare data for improved fraud detection.

1. "Healthcare insurance fraud detection using data mining" (2024).

<https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-024-02512-4>

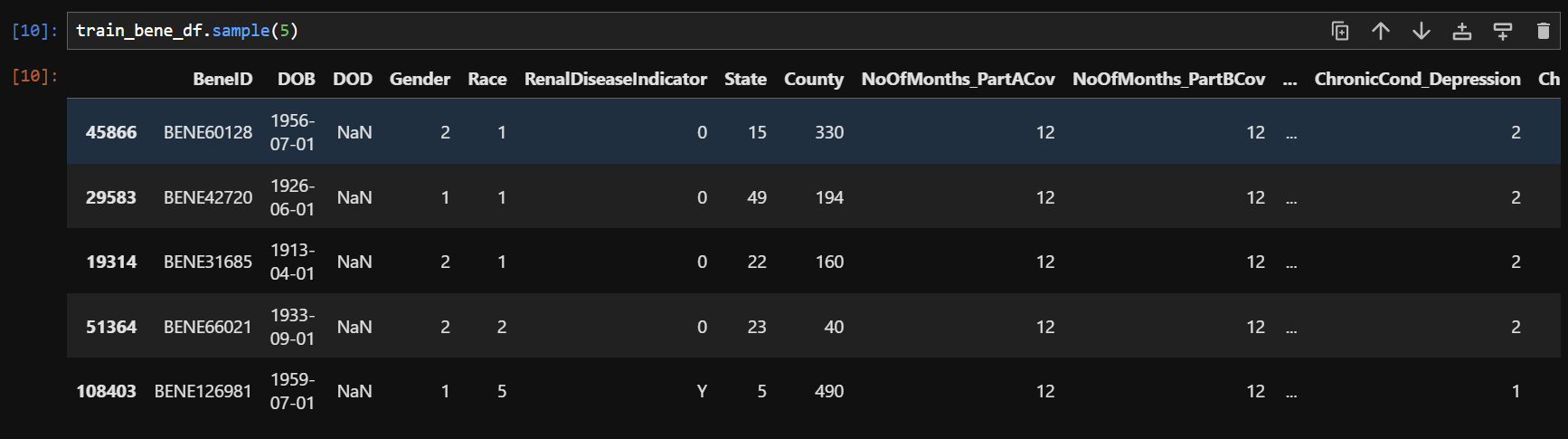
While this article focuses on unsupervised learning, it mentions the use of supervised learning methods like naive Bayes and random forest for identifying fraudulent behavior in healthcare insurance claims.

**Data**

1. **Data Collection**

The dataset for this project was sourced from Kaggle (<https://www.kaggle.com/datasets/rohitrox/healthcare-provider-fraud-detection-analysis?resource=download>), specifically the Healthcare Provider Fraud Detection Analysis dataset. It contains over 30,000 records across multiple related files, including beneficiary, inpatient, outpatient claim data, and provider information. To create a unified dataset, the files— *Train\_Beneficiarydata-1542865627584.csv*, *Train\_Inpatientdata-1542865627584.csv*, *Train\_Outpatientdata-1542865627584.csv*, and *Train-1542865627584.csv* —were merged using common keys such as BeneID, ClaimID, and Provider. The final combined train dataset contains 558,211 instances (rows) and 57 attributes (columns), while the test dataset contains 135,392 instances and 56 attributes. The dataset includes details on patient demographics (age, gender, race, and health conditions), claim information (claim dates, diagnosis and procedure codes), provider data, and financial metrics (reimbursements and deductibles). The target variable is 'PotentialFraud', a binary label indicating if a healthcare provider is suspected of fraudulent activity ("Yes" or "No"). Data preprocessing steps included merging datasets, handling missing values, encoding categorical features, and normalizing continuous variables to ensure compatibility with machine learning algorithms. This comprehensive dataset facilitates the development of a robust fraud detection model for healthcare claims.

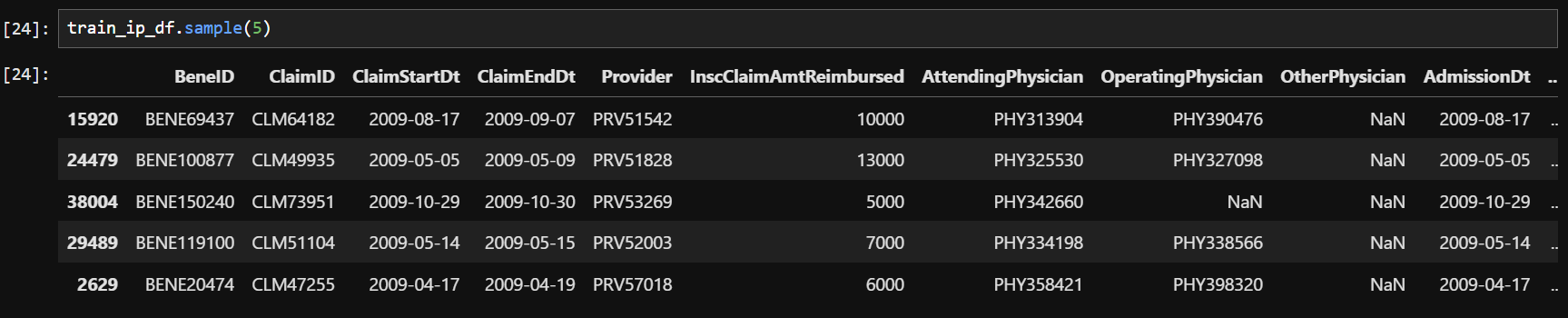
**Train\_Beneficiarydata-1542865627584.csv (Test\_Beneficiarydata-1542969243754.csv)**

****

*This data contains beneficiary KYC details like health conditions, region they belong to etc.*

|  |  |
| --- | --- |
| **Attribute Name** | **Description** |
| BeneID | Unique identifier for the beneficiary (patient) in the healthcare system. |
| DOB | Date of birth of the beneficiary, used to calculate the patient\u2019s age at the time of a claim. |
| DOD | Date of death of the beneficiary, if applicable. Identifies claims made after the patient's death. |
| Gender | Gender of the beneficiary, typically 'M' (male) or 'F' (female). |
| Race | Race or ethnicity of the beneficiary, represented by a categorical or numerical code. |
| RenalDiseaseIndicator | Binary indicator ('Yes' or 'No') specifying if the beneficiary has renal disease. |
| State | State of residence of the beneficiary, often represented by a state code or name. |
| County | County of residence of the beneficiary, providing a localized geographical indicator. |
| NoOfMonths\_PartACov | Number of months the beneficiary was covered under Medicare Part A (hospital insurance). |
| NoOfMonths\_PartBCov | Number of months the beneficiary was covered under Medicare Part B (outpatient medical insurance). |
| ChronicCond\_Alzheimer | Binary indicator (0 or 1) indicating if the beneficiary has a chronic condition related to Alzheimer\u2019s disease. |
| ChronicCond\_Heartfailure | Binary indicator showing if the beneficiary has been diagnosed with chronic heart failure. |
| ChronicCond\_KidneyDisease | Binary indicator reflecting if the beneficiary has chronic kidney disease. |
| ChronicCond\_Cancer | Binary indicator showing if the beneficiary has a chronic cancer diagnosis. |
| ChronicCond\_ObstrPulmonary | Binary indicator for the presence of obstructive pulmonary disease in the beneficiary. |
| ChronicCond\_Depression | Binary indicator identifying if the beneficiary has been diagnosed with depression. |
| ChronicCond\_Diabetes | Binary indicator indicating if the beneficiary has been diagnosed with diabetes. |
| ChronicCond\_IschemicHeart | Binary indicator that shows if the beneficiary has ischemic heart disease. |
| ChronicCond\_Osteoporasis | Binary indicator representing whether the beneficiary has been diagnosed with osteoporosis. |
| ChronicCond\_rheumatoidarthritis | Binary indicator for the presence of rheumatoid arthritis in the beneficiary. |
| ChronicCond\_stroke | Binary indicator that denotes if the beneficiary has experienced a stroke. |
| IPAnnualReimbursementAmt | Total annual reimbursement amount for inpatient services for the beneficiary. |
| IPAnnualDeductibleAmt | Total annual deductible amount for inpatient services. Amount the beneficiary must pay before coverage begins. |
| OPAnnualReimbursementAmt | Total annual reimbursement amount for outpatient services. |
| OPAnnualDeductibleAmt | Total annual deductible amount for outpatient services. Amount the beneficiary must pay before coverage begins. |

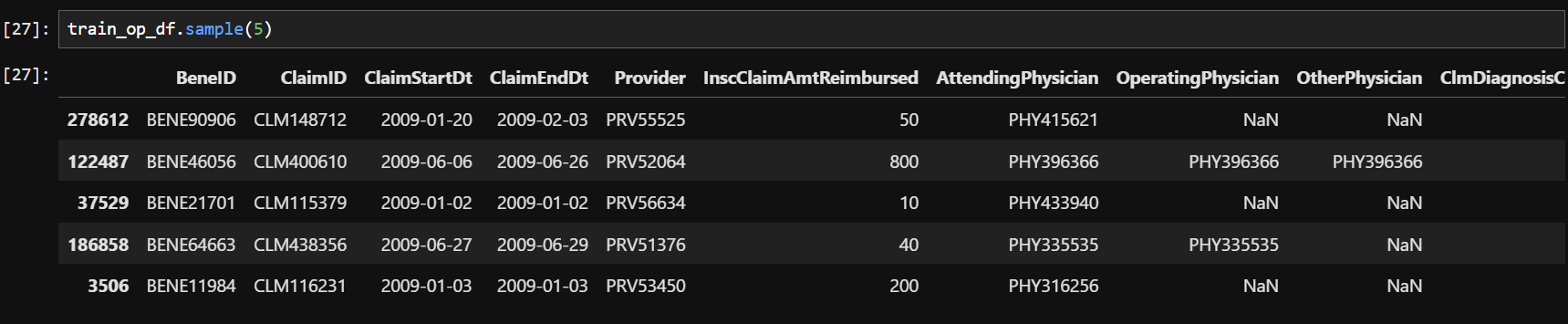
**Train\_Inpatientdata-1542865627584.csv (Test\_Inpatientdata-1542969243754.csv)**

****

*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*

|  |  |
| --- | --- |
| **Attribute Name** | **Description** |
| BeneID | Unique identifier for the beneficiary (patient) associated with the claim. |
| ClaimID | Unique identifier for the medical claim. |
| ClaimStartDt | The start date of the medical claim. |
| ClaimEndDt | The end date of the medical claim. |
| Provider | Identifier for the healthcare provider associated with the claim. |
| InscClaimAmtReimbursed | The amount reimbursed for the claim by the insurance provider. |
| AttendingPhysician | Identifier for the primary physician who attended to the patient during the claim. |
| OperatingPhysician | Identifier for the physician who performed any procedures during the claim period. |
| OtherPhysician | Identifier for any other physician involved in the patient's treatment. |
| AdmissionDt | The date the patient was admitted to the hospital. |
| ClmAdmitDiagnosisCode | Diagnosis code at the time of hospital admission. |
| DeductibleAmtPaid | The amount paid by the patient as part of their deductible for the claim. |
| DischargeDt | The date the patient was discharged from the hospital. |
| DiagnosisGroupCode | Grouping code for related diagnoses for the claim. |
| ClmDiagnosisCode\_1 | Diagnosis code (e.g., ICD code) for the patient's medical condition related to the claim. |
| ClmDiagnosisCode\_2 | Secondary diagnosis code for the patient's medical condition. |
| ClmDiagnosisCode\_3 | Tertiary diagnosis code for the patient's medical condition. |
| ClmDiagnosisCode\_4 | Fourth diagnosis code related to the claim. |
| ClmDiagnosisCode\_5 | Fifth diagnosis code related to the claim. |
| ClmDiagnosisCode\_6 | Sixth diagnosis code related to the claim. |
| ClmDiagnosisCode\_7 | Seventh diagnosis code related to the claim. |
| ClmDiagnosisCode\_8 | Eighth diagnosis code related to the claim. |
| ClmDiagnosisCode\_9 | Ninth diagnosis code related to the claim. |
| ClmDiagnosisCode\_10 | Tenth diagnosis code related to the claim. |
| ClmProcedureCode\_1 | Procedure code (e.g., CPT or ICD procedure code) for the first medical procedure performed. |
| ClmProcedureCode\_2 | Procedure code for the second medical procedure performed. |
| ClmProcedureCode\_3 | Procedure code for the third medical procedure performed. |
| ClmProcedureCode\_4 | Procedure code for the fourth medical procedure performed. |
| ClmProcedureCode\_5 | Procedure code for the fifth medical procedure performed. |
| ClmProcedureCode\_6 | Procedure code for the sixth medical procedure performed. |

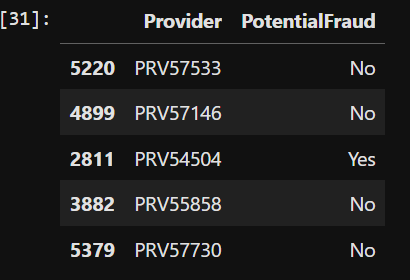
**Train\_Outpatientdata-1542865627584.csv (Test\_Outpatientdata-1542969243754.csv)**

****

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

|  |  |
| --- | --- |
| **Attribute Name** | **Description** |
| BeneID | Unique identifier for the beneficiary (patient) associated with the claim. |
| ClaimID | Unique identifier for the outpatient medical claim. |
| ClaimStartDt | The start date of the outpatient medical claim. |
| ClaimEndDt | The end date of the outpatient medical claim. |
| Provider | Identifier for the healthcare provider associated with the claim. |
| InscClaimAmtReimbursed | The amount reimbursed for the claim by the insurance provider. |
| AttendingPhysician | Identifier for the primary physician who attended to the patient during the claim. |
| OperatingPhysician | Identifier for the physician who performed any procedures during the claim period. |
| OtherPhysician | Identifier for any other physician involved in the patient's treatment. |
| ClmDiagnosisCode\_1 | Primary diagnosis code (e.g., ICD code) for the patient's medical condition related to the claim. |
| ClmDiagnosisCode\_2 | Secondary diagnosis code for the patient's medical condition. |
| ClmDiagnosisCode\_3 | Tertiary diagnosis code for the patient's medical condition. |
| ClmDiagnosisCode\_4 | Fourth diagnosis code related to the claim. |
| ClmDiagnosisCode\_5 | Fifth diagnosis code related to the claim. |
| ClmDiagnosisCode\_6 | Sixth diagnosis code related to the claim. |
| ClmDiagnosisCode\_7 | Seventh diagnosis code related to the claim. |
| ClmDiagnosisCode\_8 | Eighth diagnosis code related to the claim. |
| ClmDiagnosisCode\_9 | Ninth diagnosis code related to the claim. |
| ClmDiagnosisCode\_10 | Tenth diagnosis code related to the claim. |
| ClmProcedureCode\_1 | Procedure code (e.g., CPT or ICD procedure code) for the first medical procedure performed. |
| ClmProcedureCode\_2 | Procedure code for the second medical procedure performed. |
| ClmProcedureCode\_3 | Procedure code for the third medical procedure performed. |
| ClmProcedureCode\_4 | Procedure code for the fourth medical procedure performed. |
| ClmProcedureCode\_5 | Procedure code for the fifth medical procedure performed. |
| ClmProcedureCode\_6 | Procedure code for the sixth medical procedure performed. |
| DeductibleAmtPaid | The amount paid by the patient as part of their deductible for the claim. |
| ClmAdmitDiagnosisCode | Diagnosis code at the time of outpatient service admission. |

**Train-1542865627584.csv (Test-1542969243754.csv)**

****

*This data provides information about providers*

|  |  |
| --- | --- |
| **Attribute Name** | **Description** |
| Provider | Unique identifier for the healthcare provider associated with the claim. |
| PotentialFraud | Indicator of whether the provider is suspected of fraudulent activity (e.g., "Yes" for suspected fraud, "No" otherwise). |

1. **Data Pre-processing**

To clean the healthcare fraud detection dataset, a systematic data preprocessing approach will be employed to ensure data quality and compatibility with machine learning models. The process will involve the use of Python libraries such as Pandas, NumPy, and Scikit-learn. First, duplicate records will be identified and removed using drop\_duplicates() to avoid redundancy. Missing values will be handled based on the nature of the attribute—columns with a high percentage of missing values may be dropped, while missing numerical values may be imputed using statistical methods (like the median or mean).

For data type conversions, certain columns will be transformed into numeric format. Specifically, the PotentialFraud and RenalDiseaseIndicator columns, originally categorical, will be converted to integer format (int64). Additionally, the Provider column will be transformed by extracting the numeric portion (excluding the first three characters) and converting it to an integer (int64).

Date columns, such as DOB, ClaimStartDt, and ClaimEndDt, will be converted to datetime format, allowing the calculation of derived features like patient age, claim duration or treatment duration. Non-numeric categorical attributes (e.g., Gender, Race, and RenalDiseaseIndicator columns) will be numerically encoded, rather than treated as categorical features. This encoding will ensure that the data is in a format suitable for machine learning models.

The target variable for the project is PotentialFraud, a binary classification label that identifies providers as either fraudulent (1) or not fraudulent (0). Since this column is already converted to integer format, no further transformation is required.

To ensure model readiness, feature scaling will be applied to numerical columns (such as reimbursement amounts) using normalization or standardization, which can help improve model convergence and performance. Outliers will be identified and handled as needed to prevent skewed model training.

This comprehensive, multi-step preprocessing approach ensures the data is clean, consistent, and fully prepared for training machine learning models.

**Analysis & Methodology**

To achieve the objective of detecting healthcare insurance fraud, a systematic and data-driven approach is adopted. The process involves a series of analytical and machine learning techniques to ensure comprehensive feature engineering, robust model training, and thorough evaluation. The following outlines the key stages of analysis and the methodologies employed.

**1. Data Preparation and Feature Engineering** The first step in the analysis involves preparing the data by creating separate sets for independent features and the target column. This step ensures clarity and precision in the subsequent model training phase. The dataset is then split into training and validation sets, enabling the evaluation of model performance on unseen data.

To identify the most significant predictors, feature selection is carried out using Recursive Feature Elimination (RFE) in conjunction with a RandomForestClassifier model. This method determines the top 10 most impactful features, reducing dimensionality and improving computational efficiency.

**2. Baseline Model Development** A baseline model is created using a Gaussian Naive Bayes (GaussianNB) classifier. This model serves as a performance benchmark for subsequent, more sophisticated models. The confusion matrix and validation report reveal the following key metrics:

* **Accuracy:** 62%
* **Precision (class 0):** 63%, **Recall (class 0):** 91%, **F1-Score (class 0):** 75%
* **Precision (class 1):** 49%, **Recall (class 1):** 14%, **F1-Score (class 1):** 22% These results highlight areas for improvement, particularly in the classification of fraudulent claims (class 1).

**3. Model Comparison and Selection** To enhance predictive performance, five classification models are tested: DecisionTreeClassifier, LogisticRegression, KNeighborsClassifier, RandomForestClassifier, and XGBClassifier. Each model’s performance is evaluated using confusion matrices and classification reports.

* **DecisionTreeClassifier**: Accuracy of 68% with better recall for class 1 (28%) compared to the baseline.
* **LogisticRegression**: Accuracy of 55%, with improved F1-scores for both classes but suboptimal classification for fraudulent claims.
* **KNeighborsClassifier**: Accuracy of 63%, achieving a more balanced classification performance with F1-scores of 70% (class 0) and 49% (class 1).
* **RandomForestClassifier**: Accuracy of 89%, with precision and recall of over 89% for both classes, representing a significant improvement.
* **XGBClassifier**: Accuracy of 81%, with strong precision and recall metrics for class 1, achieving better balance across classes than some other models.

Based on these results, RandomForestClassifier and XGBClassifier demonstrate superior performance, with RandomForestClassifier emerging as the leading candidate for further optimization.

**4. Model Tuning** To further improve the performance of the RandomForestClassifier, hyperparameter tuning is conducted using RandomizedSearchCV. This process optimizes parameters such as max\_depth, max\_features, min\_samples\_leaf, min\_samples\_split, and n\_estimators. The best parameters identified are:

* **max\_depth**: 40
* **max\_features**: 'log2'
* **min\_samples\_leaf**: 1
* **min\_samples\_split**: 3
* **n\_estimators**: 152

The tuned RandomForest model achieves an accuracy of 89%, maintaining robust precision, recall, and F1-scores across both classes.

**5. Evaluation Metrics** To assess the effectiveness of the models, key evaluation metrics are used, including accuracy, precision, recall, and F1-score. These metrics provide a comprehensive view of the model's ability to classify both fraudulent and non-fraudulent claims accurately. Precision is particularly crucial in fraud detection, as it measures the proportion of actual fraud cases correctly identified from all predicted fraud cases.

In summary, this methodology involves data preparation, feature selection, baseline model development, model comparison, hyperparameter tuning, and model evaluation. The RandomForestClassifier emerges as the most effective model, demonstrating high accuracy and balanced performance for both classes. The proposed methods and evaluation metrics ensure a rigorous approach to healthcare fraud detection, supporting the development of a robust and reliable predictive system.

**References**

Please insert citations of any scientific articles you include in this proposal to support your research plan. Use standard citation styles such as APA, MLA, Chicago, etc.