**Project Documentation**

**Problem Statement**

In this classification problem, we will employ data science techniques to address the following objectives:

1. Address the class imbalance issue using data resampling techniques.

2. Create machine learning models to predict "Coverage Code" and "Accident Source" from the given dataset.

3. Design a user-friendly GUI, which takes an Excel dataset file as input, runs the model, and provides evaluation results on the screen. The model results will be stored in an Excel file.

4. The system should take an Excel file as input and add two new columns for predictions, and the output must be an Excel file.

**Dataset Description**

The dataset contains over 190,000 claim records, each with a single feature, "Claim Description." The target columns are "Coverage Code" and "Accident Source."

**Evaluation**

The project's evaluation will be based on precision and recall scores for the "Coverage Code" and "Accident Source" predictions. The evaluation results are as follows:

- Coverage Code Precision: 64.76

- Coverage Code Recall: 65.62

- Accident Source Precision: 60.93

- Accident Source Recall: 56.82

**Libraries and Algorithms Used**

We will employ the following libraries and algorithms for this project:

**Libraries:**

- pandas: For data manipulation and analysis.

- numpy: For numerical computations.

- scikit-learn (sklearn): For machine learning tasks.

- xgboost: For implementing the XGBoost classifier.

- imbalanced-learn (imblearn): For addressing class imbalance using resampling techniques.

**Algorithms:**

- RandomForestClassifier + MultiOutputClassifier: For creating a multi-output classification model.

- XGBoostClassifier: For building an XGBoost-based classifier.

**Project Components**

**Addressing Class Imbalance**

To handle the class imbalance issue, we will employ resampling techniques from the imbalanced-learn library, such as oversampling, undersampling, or synthetic data generation methods like SMOTE. The choice of technique will be based on the dataset's characteristics and experimental results.

**Model Creation**

We will create machine learning models to predict "Coverage Code" and "Accident Source" using the given dataset. Two primary algorithms are chosen for this purpose:

1. RandomForestClassifier + MultiOutputClassifier: To build a multi-output classification model.

2. XGBoostClassifier: For implementing the XGBoost-based classifier.

The performance of these models will be evaluated using precision and recall scores.

**GUI Design**

We will design a user-friendly GUI using Streamlit, which provides an easy way for users to interact with the model. The GUI components will include:

- An input mechanism to select the dataset file, enabling users to upload their data.

- A "Run" button to initiate the model's predictions.

- A display area for showing the evaluation results, including precision and recall scores.

- An option to specify the output folder for storing the Excel file with predictions.

**Excel Output**

After running the model, the system will generate an Excel file that includes the original data along with two new columns: "Predicted Coverage Code" and "Predicted Accident Source." Users will have the flexibility to choose the location where the output Excel file will be saved.

**Conclusion**

This project leverages various libraries and algorithms to address a classification problem with class imbalance. By designing a Streamlit-based GUI, users can easily execute the model and obtain predictions for "Coverage Code" and "Accident Source." The model's performance will be assessed based on precision and recall scores, and the results will be displayed in the GUI and stored in an output Excel file. This project aims to provide an efficient tool for automating predictions and enhancing the management of insurance claim data.

Project Structure:

1. App.py : Main file to run project ( use streamlit run app.py command in terminal ).
2. Dataset\_Public.xlsx : Given dataset.
3. Input\_file.xlsx : Input dataset.
4. Model.ipynb : Model building file.
5. xgbModel\_as.pkl : Model for predict values for Coverage Code.
6. xgbModel\_cc.pkl : Model for predict values for Accident Source.