Title: Predicting Flight Delays

***Abstract*—This report presents a case study analyzing the factors contributing to flight delays. The study analyzes data from multiple sources, including flight records, weather reports, and airline operations data, to identify patterns and potential causes of delays. The report also discusses potential solutions and recommendations to minimize the impact of delays on passengers and airlines.**

I. Introduction:

The following code is a Python program that uses various Machine Learning algorithms to predict flight delays. It uses a dataset called "FlightDelays.csv" which contains information about various flights such as the carrier, origin, destination, flight date, tail number, and flight status (on-time or delayed). The program uses various Python libraries such as Pandas, Matplotlib, Seaborn, Scikit-learn, and Random Forest Classifier.

II. Methodology:

The program begins by importing various libraries such as Pandas, Matplotlib, Seaborn, Scikit-learn, and Random Forest Classifier. It then reads in the "FlightDelays.csv" file and performs basic exploratory data analysis by printing the data types of various columns using df.info() function. It also checks for missing values in the dataset by printing the sum of null values for each column using df.isnull().sum() function.

The program then proceeds to visualize the categorical variables by plotting count plots for each variable using the Seaborn library. It also creates a correlation matrix using the Pandas library and visualizes it using the Seaborn library's heatmap function. The program then drops the FL_DATE and TAIL_NUM columns from the dataset as they are not relevant for the analysis.

Next, the program performs one-hot encoding on the categorical variables using the Scikit-learn library's OneHotEncoder function. It then concatenates the encoded columns with the original dataset and maps the Flight Status column to a binary format where 'ontime' is mapped to 0 and 'delayed' is mapped to 1. It then splits the dataset into training and testing sets using the Scikit-learn library's train_test_split function.

The program then trains three different Machine Learning models, namely K-Nearest Neighbors, Naive Bayes, and Random Forest Classifier, using the Scikit-learn library. It evaluates each model's performance using the Scikit-learn library's score function and stores the score for each model in a dictionary. It then selects the best performing model by selecting the one with the highest score and prints the results.

III. Results:

The program uses three different Machine Learning models, namely K-Nearest Neighbors, Naive Bayes, and Random Forest Classifier, to predict flight delays. The results show that K-Nearest Neighbors performed the best with a score of 0.8775. Random Forest Classifier performed second best with a score of 0.8752, and Naive Bayes performed the worst with a score of 0.7709.

IV. Conclusion:

In conclusion, the program successfully predicts flight delays using Machine Learning algorithms. It uses various Python libraries such as Pandas, Matplotlib, Seaborn, Scikit-learn, and Random Forest Classifier to perform exploratory data analysis, data visualization, one-hot encoding, and model training. The results show that the that K-Nearest Neighbors performs the best, and the program selects it as the best performing model for predicting flight delays.