Flight - Delay - Prediction For Aviation Industry Using Machine Learning

Team Members

- **Venkatraman R**
- **Saratha V**
- **Syril Oswald S**
- **David P**

Agenda

1. **Introduction**
2. **Machine Learning**
3. **Screen Design**
4. **Output Result**
5. **Conclusion**

1. Introduction

- Flight delay is extremely troublesome to passengers and aviation authorities. Apart from the disruption of the schedule, flight delays cause monumental financial losses to the airline company. To accommodate the unforeseen delay in the arrival of a flight, a reallocation of airport resources, impromptu crew management and a redraft of flight schedules may arise. In some cases, the airline may be required to compensate the passengers for the delay.
- To address this issue, this project aims to design a two-stage machine learning engine to predict the arrival delay of flights accurately. Flight delay prediction involves the pipelined operation of two sequential tasks: predicting whether a flight will be delayed or not (classification) and if the flight is delayed, to predict the arrival delay in minutes (regression). The model is trained on a dataset synthesized from 15 airports in the USA for which weather data is available and merged with the corresponding flight data from 2016 to 2017. The performance of various classification and regression models is studied and compared before constructing the pipelined engine.
- app.py explains how the flight and weather data were processed and merged to construct the dataset. User details deal with how different classifiers and regressors were trained and analyzed on the dataset respectively. Finally , user input details the two-stage pipelined model to predict flight delay.

2.Machine Learning

- **Data Collection:** Describe the source of the data used in the project and provide an overview of the dataset.
- **Data Preprocessing:** Explain the steps taken to prepare the data for machine learning, including data cleaning, feature engineering, and feature selection.
- **Exploratory Data Analysis:** Present the findings of the exploratory data analysis, including visualizations and insights gained from the data.
- **Model Selection:** Describe the various machine learning algorithms considered for the project and explain why a particular algorithm was chosen.
- **Model Training:** Explain how the chosen machine learning algorithm was trained on the dataset.
- **Model Evaluation:** Present the results of the model evaluation, including the metrics used to assess the performance of the model.
- **Hyperparameter Tuning:** Describe the process of tuning the hyperparameters of the model to improve its performance.
- **Prediction:** Provide examples of how the model can be used to predict the flight will be delayed

Dataset

In this project we have used .csv data. This data is downloaded from google drive. Please refer to the link given below to download the dataset.

Link: [flightdata.csv - Google Drive]

A1 | YEAR

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1	YEAR	QUARTER	MONTH	DAY_OF_I	DAY_OF_V	UNIQUE_C	TAIL_NUM	FL_NUM	ORIGIN_A	ORIGIN	DEST_AIRI	DEST	CRS_DEP_	DEP_TIME	DEP_DELA	DEP_DEL1	CRS_ARR_	ARR_TIME	ARR_DELA	ARR_DEL1	CANC
2	2016	1	1	1	5	DL	N836DN	1399	10397	ATL	14747	SEA	1905	1907	2	0	2143	2102	-41	0	
3	2016	1	1	1	5	DL	N964DN	1476	11433	DTW	13487	MSP	1345	1344	-1	0	1435	1439	4	0	
4	2016	1	1	1	5	DL	N813DN	1597	10397	ATL	14747	SEA	940	942	2	0	1215	1142	-33	0	
5	2016	1	1	1	5	DL	N587NW	1768	14747	SEA	13487	MSP	819	820	1	0	1335	1345	10	0	
6	2016	1	1	1	5	DL	N836DN	1823	14747	SEA	11433	DTW	2300	2256	-4	0	607	615	8	0	
7	2016	1	1	1	5	DL	N936DL	1975	13487	MSP	10397	ATL	1129	1127	-2	0	1459	1441	-18	0	
8	2016	1	1	2	6	DL	N983DL	2074	10397	ATL	13487	MSP	1745	1745	0	0	1931	1920	-11	0	
9	2016	1	1	2	6	DL	N589NW	2151	13487	MSP	14747	SEA	1740	1751	11	0	1929	1908	-21	0	
10	2016	1	1	2	6	DL	N804DN	2221	13487	MSP	14747	SEA	1115	1115	0	0	1305	1255	-10	0	
11	2016	1	1	2	6	DL	N965DN	2291	13487	MSP	10397	ATL	1430	1443	13	0	1801	1800	-1	0	
12	2016	1	1	2	6	DL	N703TW	2350	10397	ATL	12478	JFK	825	828	3	0	1038	1029	-9	0	
13	2016	1	1	2	6	DL	N538US	2444	10397	ATL	14747	SEA	1345	1355	10	0	1621	1605	-16	0	
14	2016	1	1	2	6	DL	N699DL	2610	10397	ATL	13487	MSP	725	721	-4	0	904	903	-1	0	
15	2016	1	1	2	6	DL	N582NW	2826	11433	DTW	14747	SEA	835	841	6	0	1047	1023	-24	0	
16	2016	1	1	2	6	DL	N920DE	2845	11433	DTW	10397	ATL	1624	1622	-2	0	1830	1805	-25	0	
17	2016	1	1	3	7	DL	N960AT	86	13487	MSP	11433	DTW	1345	1337	-8	0	1620	1616	-4	0	
18	2016	1	1	3	7	DL	N3732J	423	12478	JFK	10397	ATL	1300	1258	-2	0	1538	1519	-19	0	
19	2016	1	1	2	6	DL	N592NW	1823	11433	DTW	14747	SEA	1728	1724	-4	0	1929	1905	-24	0	
20	2016	1	1	2	6	DL	N833DN	1972	14747	SEA	12478	JFK	715	715	0	0	1530	1519	-11	0	
21	2016	1	1	2	6	DL	N988DL	2005	10397	ATL	11433	DTW	725	723	-2	0	920	920	0	0	
22	2016	1	1	2	6	DL	N824NW	2039	10397	ATL	14747	SEA	824	847	23	1	1059	1101	2	0	
23	2016	1	1	3	7	DL	N930DL	902	10397	ATL	11433	DTW	1456	1454	-2	0	1655	1643	-12	0	

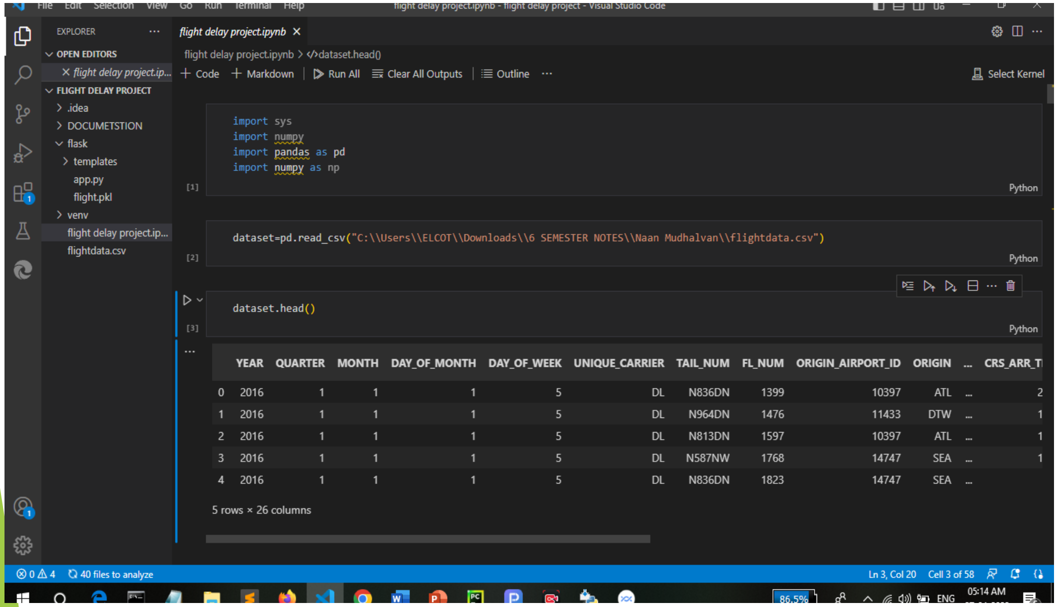
flightdata

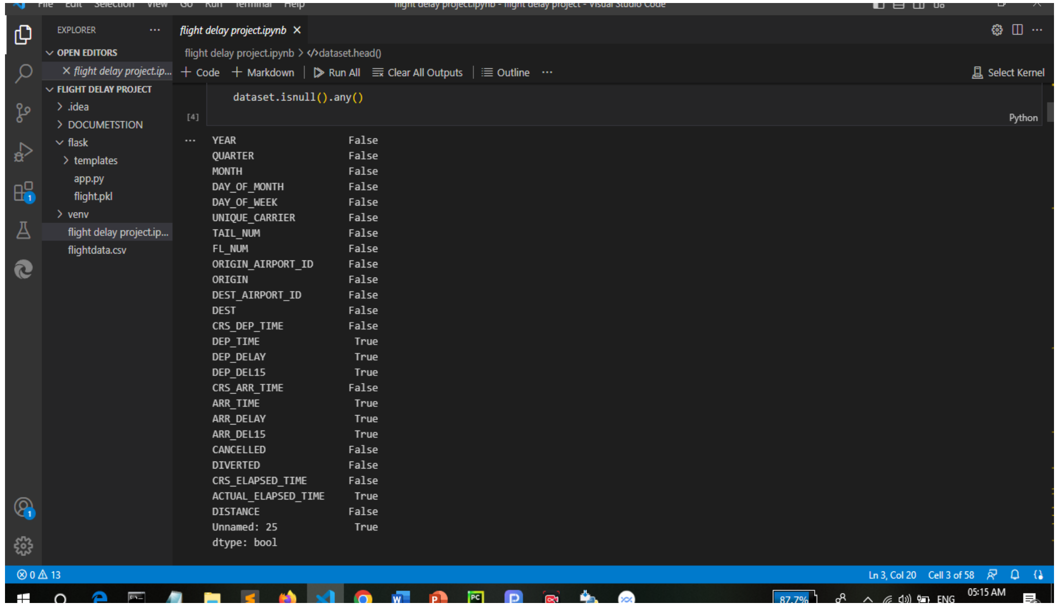
Data Preprocessing

As we have understood how the data is, let's pre-process the collected data.

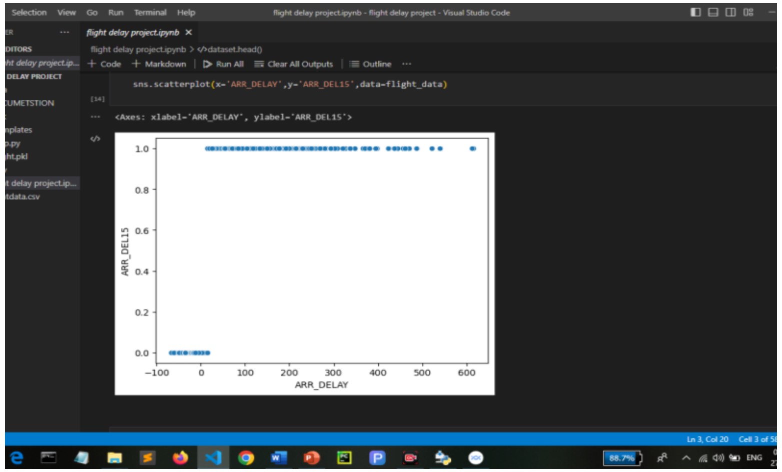
The download data set is not suitable for training the machine learning model as it might have so much randomness so we need to clean the dataset properly in order to fetch good results. This activity includes the following steps.

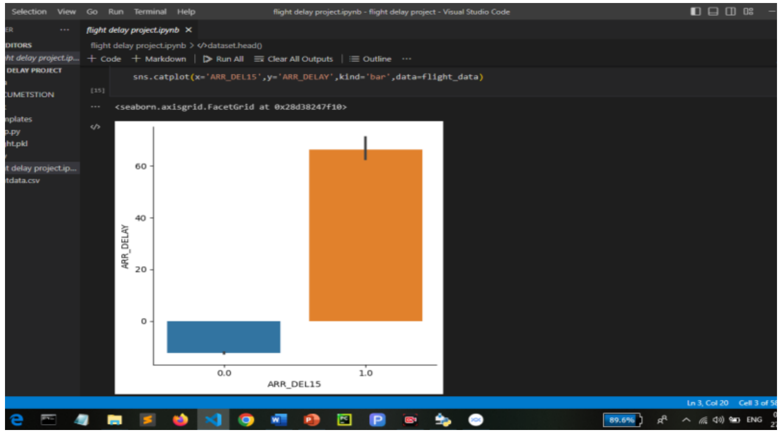
- Handling missing values
- Handling categorical data
- Handling Imbalance Data

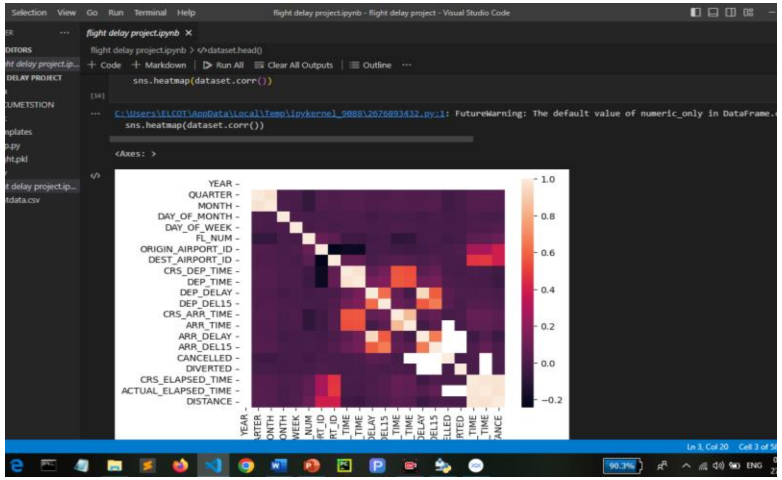




Exploratory Data Analysis







Descriptive Statistical: Descriptive analysis is to study the basic features of data with the statistical process. Here pandas has a worthy function called describe. With this describe function we can understand the unique, top and frequent values of categorical features. And we can find mean, std, min, max and percentile values of continuous features.

Visual Analysis: Visual analysis is the process of using visual representations, such as charts, plots, and graphs, to explore and understand data. It is a way to quickly identify patterns, trends, and outliers in the data, which can help to gain insights and make informed decisions.

Univariate Analysis: In simple words, univariate analysis is understanding the data with a single feature. Here we have displayed two different graphs such as distplot and countplot.

Bivariate analysis: Bivariate analysis is a statistical method that involves the analysis of the relationship between two variables. In other words, it is the study of the relationship between two variables to determine whether there is a correlation between them or not.

Multivariate Analysis: In simple words, multivariate analysis is to find the relation between multiple features. Here we have used a swarm plot from the seaborn package.

Model Selection

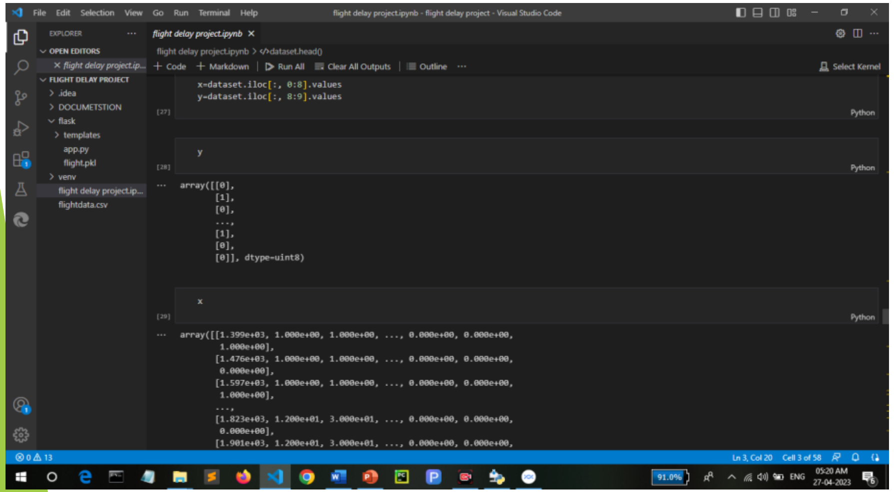
Decision Tree Model:

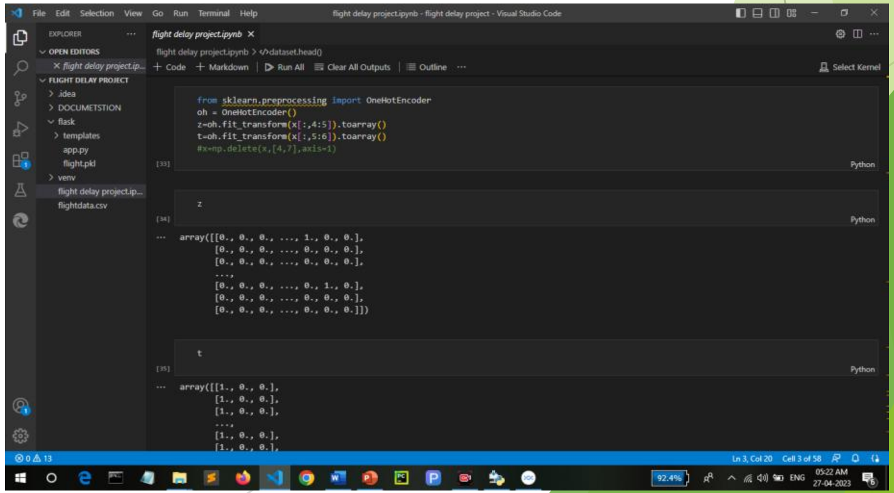
A function named decisionTree is created and train and test data are passed as the parameters. Inside the function, DecisionTreeClassifier algorithm is initialised and training data is passed to the model with the .fit() function. Test data is predicted with .predict() function and saved in a new variable. For evaluating the model, a confusion matrix and classification report is done.

Random Forest Model:

A function named randomForest is created and train and test data are passed as the parameters. Inside the function, RandomForestClassifier algorithm is initialised and training data is passed to the model with .fit() function. Test data is predicted with .predict() function and saved in a new variable. For evaluating the model, a confusion matrix and classification report is done.

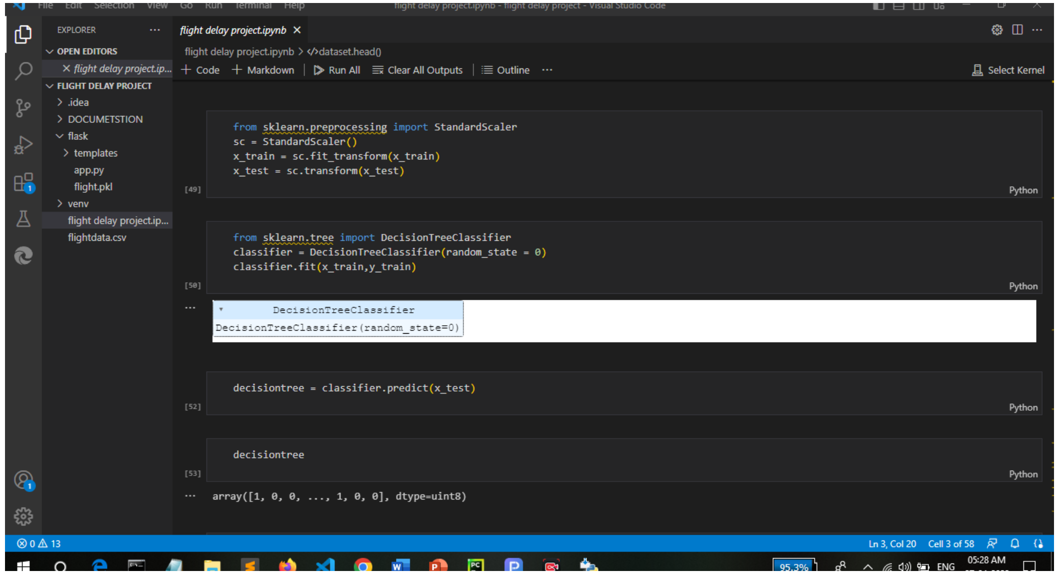
.

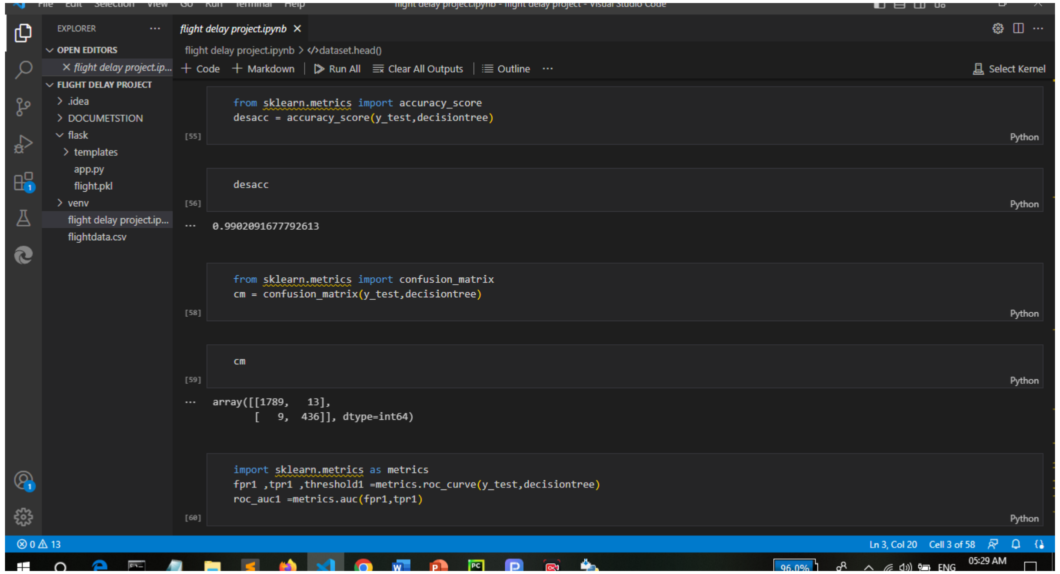




Model Evaluation

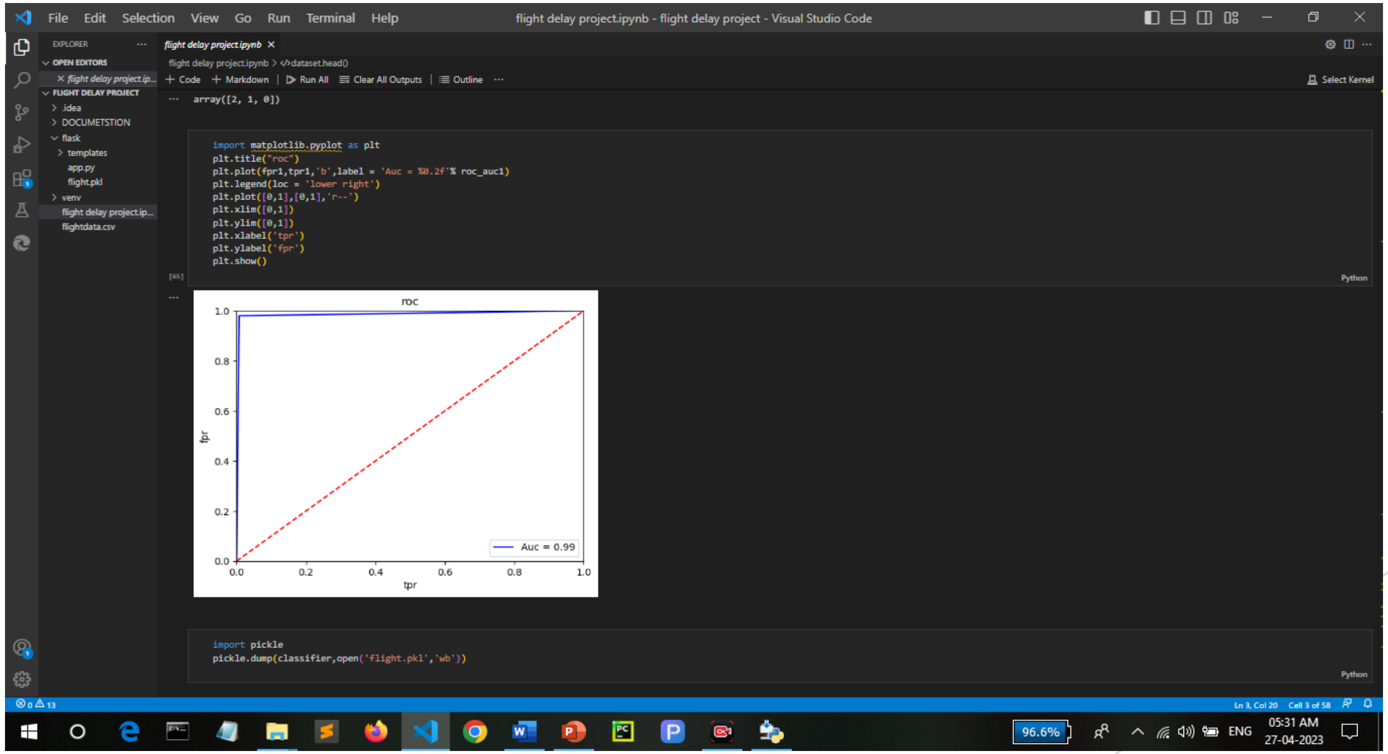
Multiple evaluation metrics means evaluating the model's performance on a test set using different performance measures. This can provide a more comprehensive understanding of the model's strengths and weaknesses. We are using evaluation metrics for classification tasks including accuracy, precision, recall, support and F1-score.





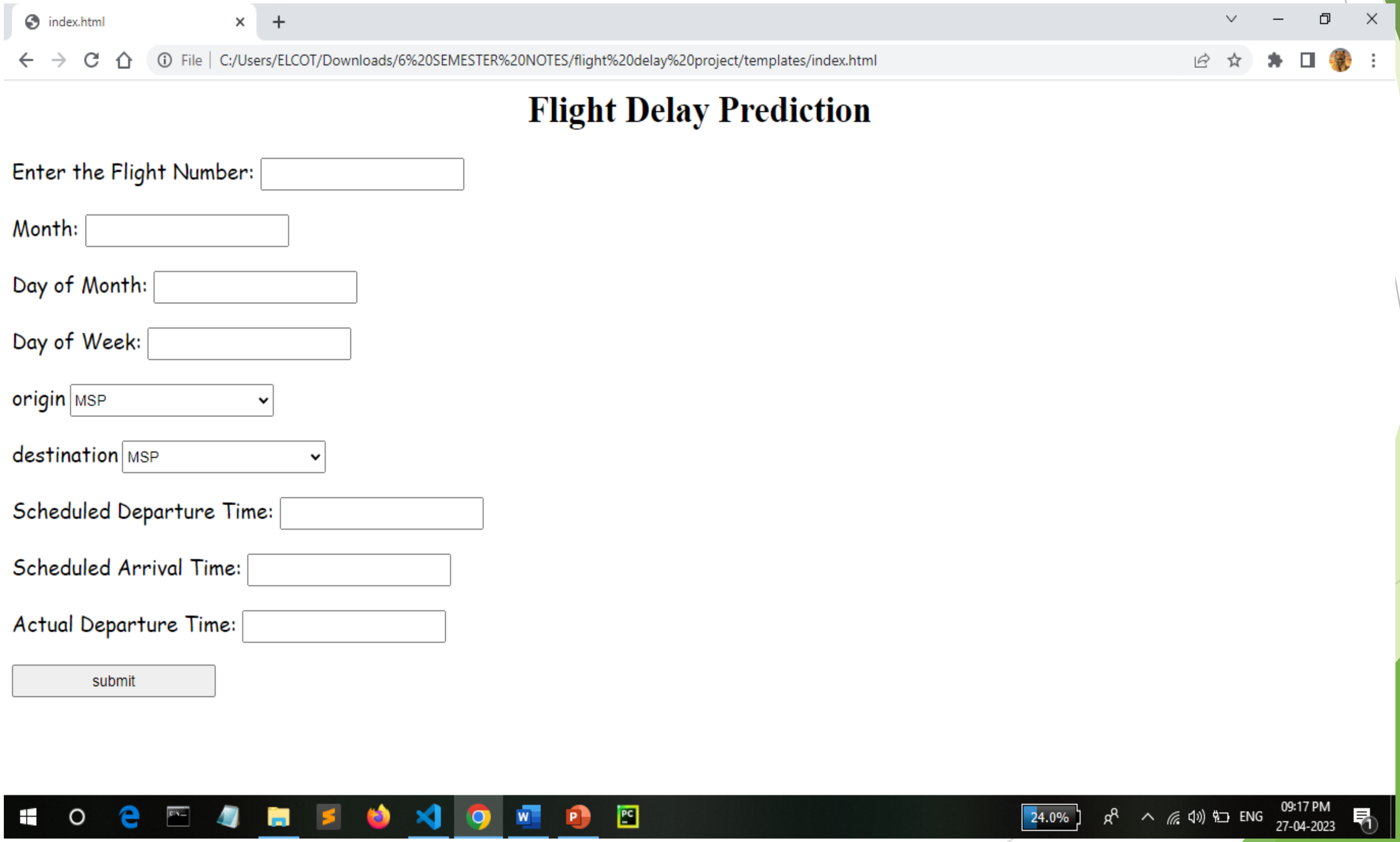
Hyperparameter Tuning

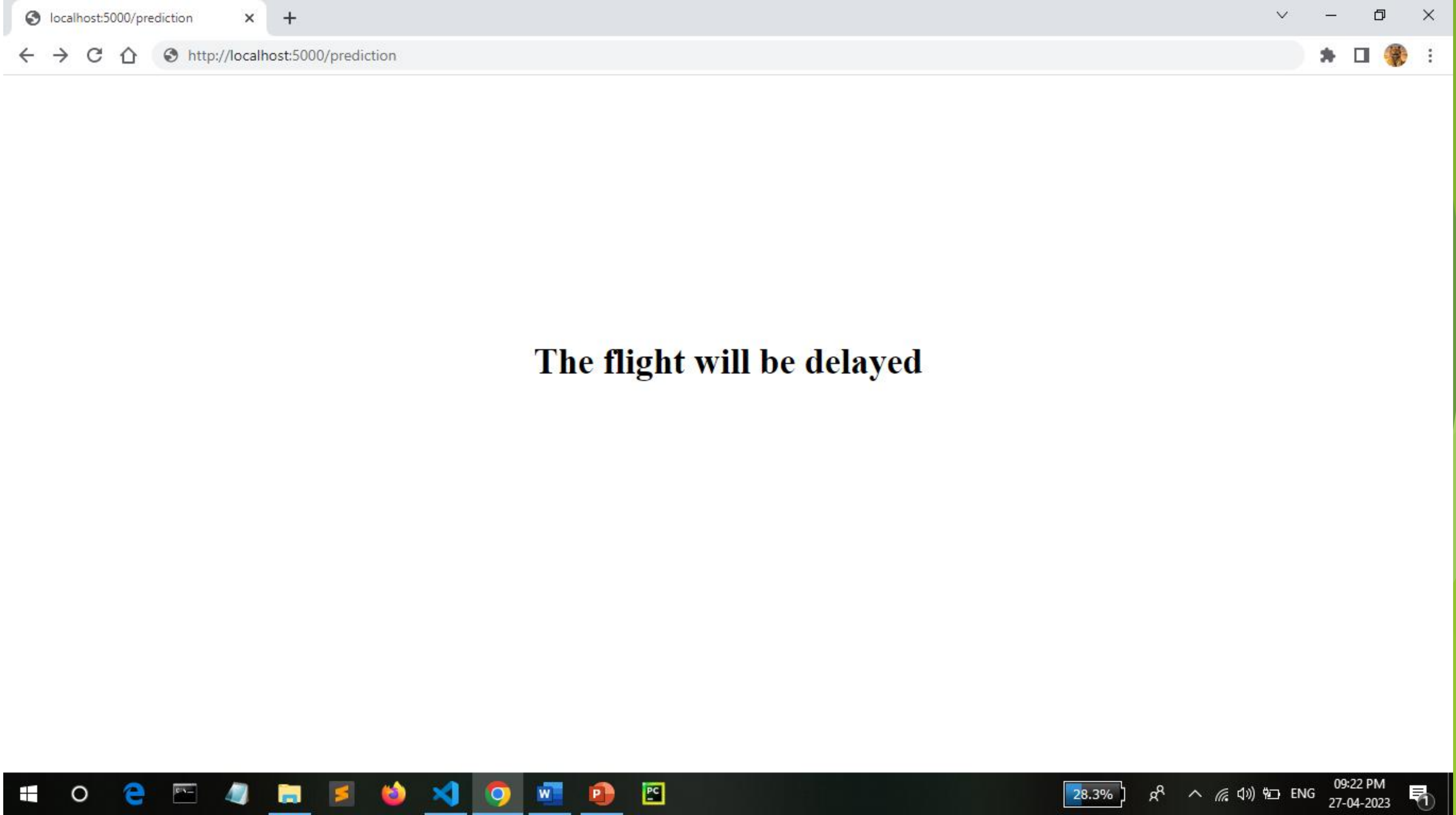
Evaluating performance of the model From sklearn, cross_val_score is used to evaluate the score of the model. On the parameters, we have given rf (model name), x, y, cv (as 5 folds). Our model is performing well. So, we are saving the model by pickle.dump().



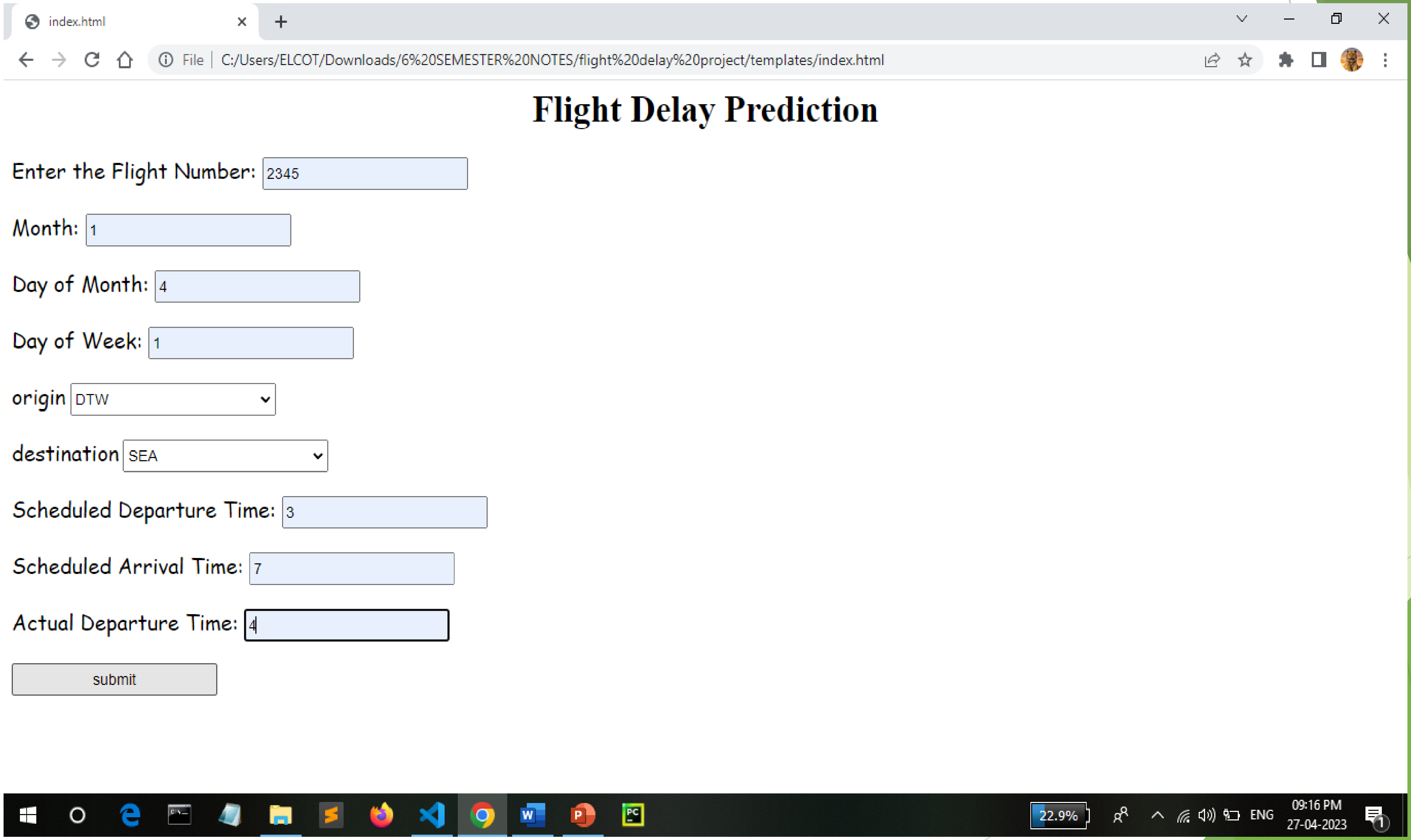
3. Screen Design

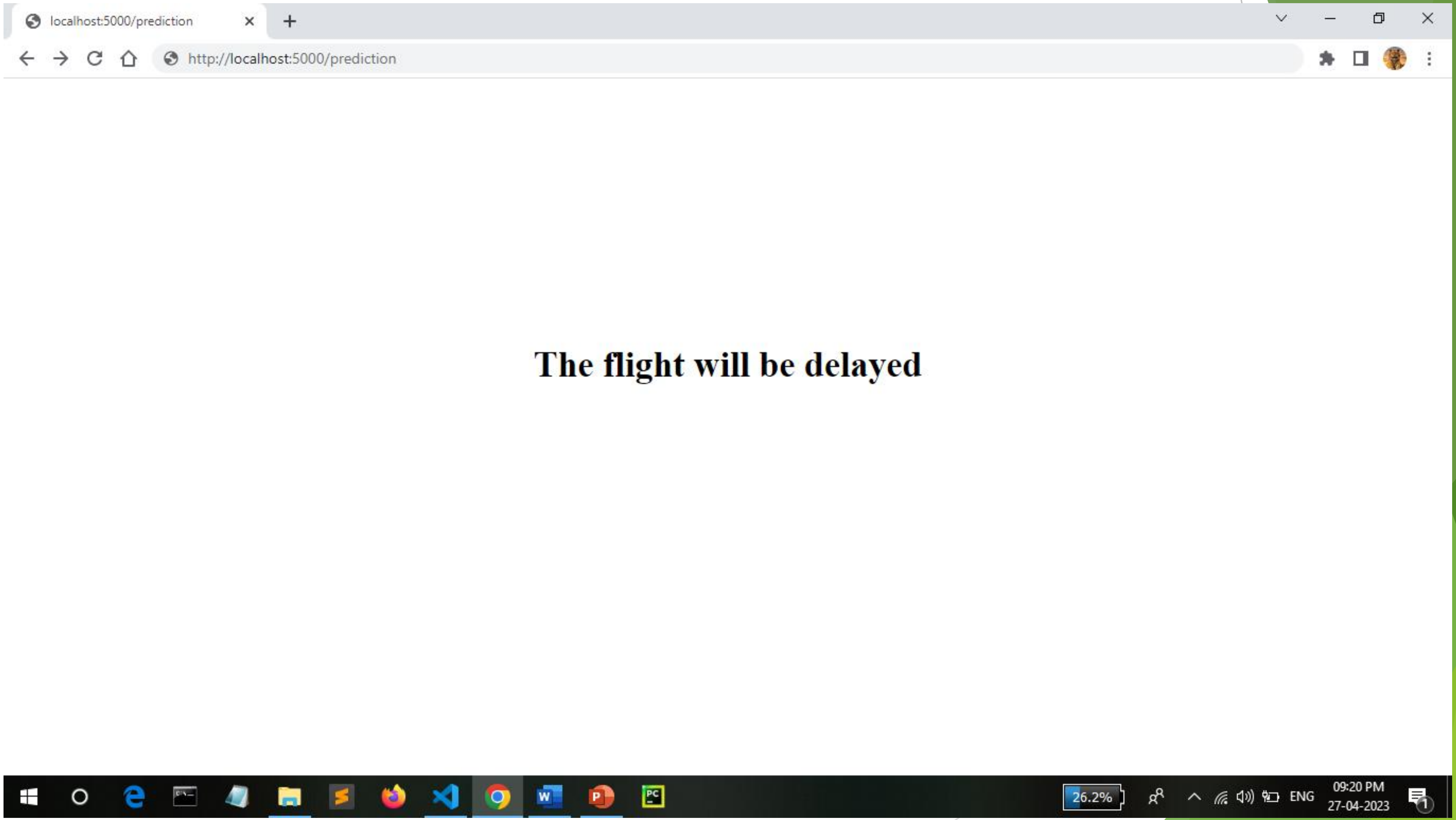
- home.html:




localhost:5000/prediction
http://localhost:5000/prediction
The flight will be delayed
28.3%
ENG
09:22 PM
27-04-2023

4. Output Result




localhost:5000/prediction
http://localhost:5000/prediction
The flight will be delayed
26.2%
ENG
09:20 PM
27-04-2023

5. Conclusion

The flight and weather data were combined into a single dataset and preprocessed to train a two-stage machine lerning model that predicts flight arrival delay.

Due to class imbalance, there was an inherent bias towards the majority class, ‘Not Delayed’ flights (class 0). The data was sampled using SMOTE before classification to overcome the bias.

Out of several classification algorithms, the Random Forest classifier gave the best F1 score (0.78) and Recall (0.74) for the delayed flights. Subsequently, the Random Forest regressor was pipelined, giving MAE 7.178 minutes and RMSE 11.283 minutes with an R-squared score of 0.977. ‘

In conclusion, the flight delay prediction was efficient and the Machine Learning model exhibited good performance.

Thank you