Task Performed: Week-2

In second week, the dataset analysis was performed for finalizing the attributes to be considered for flight fare prediction and possible machine learning approaches.

Dataset 1- https://www.kaggle.com/code/varunsaikanuri/flight-fare-prediction-10-ml-models/data
Dataset 2: https://www.kaggle.com/datasets/thedevastator/airlines-traffic-passenger-statistics

Analysis of Dataset-1:

The dataset used in the Kaggle project "Flight Fare Prediction - 10 ML Models" contains information about various flights from different airlines in India. The dataset includes 10,683 records of flight data and has 11 columns with the following attributes:

Airline: The name of the airline.

Date_of_Journey: The date of the journey.

Source: The source from which the flight departs.

Destination: The destination where the flight lands.

Route: The route taken by the flight to reach the destination.

Dep_Time: The departure time of the flight.

Arrival_Time: The arrival time of the flight.

Duration: The duration of the flight.

Total_Stops: The total number of stops between the source and destination.

Additional_Info: Additional information about the flight.

Price: The price of the ticket (Target variable).

The dataset is provided in a CSV file, and it contains no missing values, which makes it easier to work with.

In this project, the aim is to predict the flight fare for a given flight based on various features mentioned above. For this purpose, the dataset is divided into training and testing sets. The training set is used to train the machine learning models, while the testing set is used to evaluate the performance of the models.

There are a total of 10 machine learning models used in this project, including linear regression, decision tree, random forest, and XGBoost. The performance of each model is evaluated based on the mean squared error (MSE) and the R-squared score (R2).

Overall, the dataset used in this project is well-suited for the task of flight fare prediction, and the various features included in the dataset provide a good basis for the machine learning models to learn from.

Analysis of Dataset-2:

The dataset available at https://www.kaggle.com/datasets/thedevastator/airlines-traffic-passenger-statistics contains information on airline passenger traffic from various airlines operating across the world. The dataset provides a comprehensive overview of various key indicators related to airline passenger traffic, such as the number of passengers carried, the number of flights operated, the revenue generated, and the market share of various airlines.

The dataset is in a CSV (Comma Separated Values) format and contains 9 columns:

1. Year: The year for which the data has been recorded.
2. Month: The month for which the data has been recorded.
3. Airline: The name of the airline for which the data has been recorded.
4. Country: The country where the airline is headquartered.
5. Revenue: The revenue generated by the airline in the given month and year.
6. Passengers: The number of passengers carried by the airline in the given month and year.
7. Freight: The amount of freight carried by the airline in the given month and year.
8. Mail: The amount of mail carried by the airline in the given month and year.
9. Flights: The number of flights operated by the airline in the given month and year.

The dataset provides information for the years 2015-2019 and contains data for 33 airlines across the world. The dataset contains a total of 12,027 rows, each corresponding to a specific month and airline. The dataset is relatively small in size, with each row containing only 9 variables, making it easy to handle and analyze using various data analysis tools.

The dataset can be used for various types of analysis, such as identifying trends in airline passenger traffic over the years, comparing the performance of different airlines, identifying the market share of various airlines, and predicting future passenger traffic for specific airlines.

However, it is important to note that the dataset only contains information for a limited number of airlines, and may not be representative of the entire aviation industry. Additionally, the dataset does not provide information on factors that may impact airline passenger traffic, such as macroeconomic indicators, industry regulations, and geopolitical events. Therefore, any analysis conducted using this dataset should be supplemented with additional information and data sources to provide a more comprehensive understanding of the aviation industry.