**IMPLEMENTATION**

**MODULES:**

* Data Collection
* Dataset
* Data Preparation
* Model Selection
* Analyze and Prediction
* Accuracy on test set
* Saving the Trained Model

**MODULES DESCSRIPTION:**

**Data Collection:**

This is the first real step towards the real development of a machine learning model, collecting data. This is a critical step that will cascade in how good the model will be, the more and better data that we get, the better our model will perform.

There are several techniques to collect the data, like web scraping, manual interventions and etc.

The dataset used in this Flight Delay dataset taken from kaggle

Link: https://www.kaggle.com/divyansh22/flight-delay-prediction

**Dataset:**

The dataset consists of 583985 individual data. There are 21 columns in the dataset, which are described below.

YEAR — Year in which flight took place  
QUARTER — Quarter in which flight took place (1–4)  
MONTH — Month in which flight took place (1–12)  
DAY\_OF\_MONTH — Day of the month in which flight took place (1–31)  
DAY\_OF\_WEEK — 1 for Monday, 2 for Tuesday,etc. in which flight took place  
UNIQUE\_CARRIER — Airline carrier code  
TAIL\_NUM — Aircraft tail number  
FL\_NUM — Flight number  
ORIGIN\_AIRPORT\_ID — ID of origin airport  
ORIGIN — Code of origin airport(ATL, DFW, SEA, etc.)  
DEST\_AIRPORT\_ID — ID of destination airport  
DEST — Code of destination airport (ATL, DFW, SEA, etc.)  
CRS\_DEP\_TIME — Scheduled departure time  
DEP\_TIME — Actual departure time  
DEP\_DELAY — Departure Delay in minutes  
DEP\_DEL15 — 1 if departure is delayed by 15 minutes or more else 0  
CRS\_ARR\_TIME — Scheduled arrival time  
ARR\_TIME — Actual arrival time  
ARR\_DELAY — Arrival Delay in minutes  
ARR\_DEL15 — 1 if arrived late by 15 minutes or more else 0  
CANCELLED — 1 if Flight was cancelled else 0  
DIVERTED — 1 if Flight was diverted else 0  
CRS\_ELAPSED\_TIME — Scheduled flight time in minutes  
ACTUAL\_ELAPSED\_TIME — Actual flight time in minutes  
DISTANCE — Distance traveled in miles

**Data Preparation:**

we will transform the data. By getting rid of missing data and removing some columns. First we will create a list of column names that we want to keep or retain.

Next we drop or remove all columns except for the columns that we want to retain.

Finally we drop or remove the rows that have missing values from the data set.

**Model Selection:**

While creating a machine learning model, we need two dataset, one for training and other for testing. But now we have only one. So lets split this in two with a ratio of 80:20. We will also divide the dataframe into feature column and label column.

Here we imported train\_test\_split function of sklearn. Then use it to split the dataset. Also, test\_size = 0.2, it makes the split with 80% as train dataset and 20% as test dataset.

The random\_state parameter seeds random number generator that helps to split the dataset.

The function returns four datasets. Labelled them as train\_x, train\_y, test\_x, test\_y*.* If we see shape of this datasets we can see the split of dataset.

We will use Random Forest Classifier, which fits multiple decision tree to the data. Finally I train the model by passing train\_x, train\_y to the fit method.

Once the model is trained, we need to Test the model. For that we will pass test\_x to the predict method.

Random Forest is one of the most powerful methods that is used in machine learning for classification problems. The random forest comes in the category of the supervised classification algorithm. This algorithm is carried out in two different stages the first one deals with the creation of the forest of the given dataset, and the other one deals with the prediction from the classifier.

**Analyze and Prediction:**

In the actual dataset, we chose only 10 features :

1. DAY\_OF\_MONTH — Day of the month in which flight took place (1–31)
2. DAY\_OF\_WEEK — 1 for Monday, 2 for Tuesday,etc. in which flight took place
3. OP\_CARRIER \_AIRLINE\_ID — ID of origin airline
4. ORIGIN\_AIRPORT\_ID — ID of origin airport
5. DEST\_AIRPORT\_ID — ID of destination airport
6. DEP\_TIME — Actual departure time
7. ARR\_TIME — Actual arrival time
8. DEP\_DEL15 — 1 if departure is delayed by 15 minutes or more else 0
9. DIVERTED — 1 if Flight was diverted else 0
10. DISTANCE — Distance traveled in miles
11. ARR\_DEL15 — 1 if arrived late by 15 minutes or more else 0

**Accuracy on test set:**

We got a accuracy of 92.1% on test set.

**Saving the Trained Model:**

Once you’re confident enough to take your trained and tested model into the production-ready environment, the first step is to save it into a .h5 or . pkl file using a library like pickle .

Make sure you have pickle installed in your environment.

Next, let’s import the module and dump the model into . pkl file