**Predicting Flight Delays with error calculation using Machine Learned Classifiers**

**A Project Report submitted in partial fulfillment of the requirements for the award of the degree of**

**BACHELOR OF TECHNOLOGY IN**

**COMPUTER SCIENCE AND ENGINEERING**

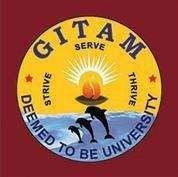
Submitted by:

|  |  |
| --- | --- |
| **M.DHAMU** | **321910302013** |
| **N.SRI MANI SRAVIKA** | **321910302014** |
| **K.SRAVANI** | **321910302022** |
| **P.MAHESWARI** | **321910302034** |
| **V.MAVYA** | **321910302039** |

**Under the esteemed guidance of**

**Dr.M.V.Jaganatha Reddy**

**Assistant Professor**



**Department of Computer Science & Engineering, GITAM SCHOOL OF TECHNOLOGY**

**GANDHI INSTITUTE OF TECHNOLOGY AND MANAGEMENT**

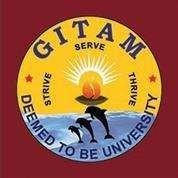
**(Deemed to be University) Bengaluru Campus.**

**April 2023**

**DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING GITAM SCHOOL OF TECHNOLOGY**

**GITAM**

**(Deemed to be University)**



**CERTIFICATE**

This is to certify that the project report entitled **“Predicting flight delays with error calculation using machine learned classifiers”** is a bonafide record of work carried out by **M.DHAMU (321910302013),N.SRI MANI SRAVIKA(321910302014),K.SRAVANI(321910302022),P.M**

**AHESWARI(321910302034),V.MAVYA(321910302039)**submitted in partial fulfillment of requirement for the award of degree of **Bachelors of Technology in Computer Science and Engineering.**

.**Project Guide. Head of the Department.**

**SIGNATURE OF THE GUIDE SIGNATURE OF THE HOD**

**Dr.M.V.Jaganatha Reddy**

Assistant Professor.

Dept of CSE, GST

**Dr. Vamsidhar Yendapalli**

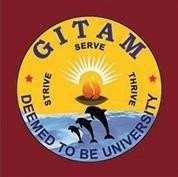
Professor.

Dept of CSE, GST

**DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING GITAM SCHOOL OF TECHNOLOGY**

**GITAM**

**(Deemed to be University)**



**DECLARATION**

We, hereby declare that the project report entitled **“Predicting flight delays with error calculation using machine learned classifiers”** is an original work done in the **Department of Computer Science and Engineering, GITAM Institute of Technology, GITAM (Deemed to be University)** submitted in partial fulfillment of the requirements for the award of the degree of **B.Tech.** in Computer Science and Engineering. The work has not been submitted to any other college or University for the award of any degree.

**Date:**

|  |  |  |
| --- | --- | --- |
| **Registration No(s).** | **Name(s)** | **Signature(s)** |
| **321910302013** | **M.DHAMU** |  |
| **321910302014** | **N.SRI MANI SRAVIKA** |  |
| **321910302022** | **K.SRAVANI** |  |
| **321910302034** | **P.MAHESWARI** |  |
| **321910302039** | **V.MAVYA** |  |

**ACKNOWLEDGEMENT**

The satisfaction and euphoria that accompany the successful completion of any seminar would be incomplete without mentioning the people who made it possible , whose constant guidance and encouragements crowned our efforts with success . I take this opportunity to express the deepest gratitude and appreciation to all those who held us directly or indirectly towards the successful completion of project.

I would like to thank Dr.Nageswara Rao, Director, GITAM School of Technology for facilities provided. I would like to thank Prof. Dr.Vamsidhar Y, HOD, Dept of Computer Science and Engineering for his support and encouragement that went a long way in the successful completion of this seminar report.

I consider this as a privilege to express our heartfelt gratitude and respect to our project guide Dr.M.V.Jaganatha Reddy, Asst. Prof in Computer Science and Engineering department for being our internal guide for her integral and incessant support offered to us throughout the course of this project and for the constant source of inspiration throughout the course of this report.

I would like to thank my team mates, friends for their support and encouragement in the completion of the seminar report in time. Last but not the least I would like to thank all the teaching and non- teaching staff members of the Computer Science and Engineering Department, for their support in completion of the mini project in time.

|  |  |
| --- | --- |
| **Students Name’s** | **Registration number** |
| 321910302013 | M.DHAMU |
| 321910302014 | N.SRI MANI SRAVIKA |
| 321910302022 | K.SRAVANI |
| 321910302034 | P.MAHESWARI |
| 321910302039 | V.MAVYA |

# ABSTRACT

Flight delays hurt airlines, airports, and passengers. Their prediction is crucial during the decision- making process for all players of commercial aviation. Flight delays also result in airline companies operating commercial flights to incur huge losses. so, they are trying their best to prevent or avoid flight delays and cancellations. In recent decades, the growth in airlines sector is immense and is causing traffic congestions. Delays also occurs due to non-suitable weather conditions or any act of god. In this project, we will examine the flight delay data and will use machine learning techniques to predict them. We will compare logistic regression, Decision Tree, Random Forest, Bayesian regression and Gradient Boosting algorithms. To solve this issue, accurately predicting these flight delays allows passengers to be well prepared for the dissuasion caused to their journey and enables airlines to respond to the potential causes of the flight delays in advance to diminish the negative impact. The purpose of this project is to look at the approaches used to build models for predicting flight delays that occur due to bad weather conditions. In the first part of the project, we look at using Python based Logistic Regression. In the second part of the project, breaking the dataset down and identifying relevant attributes. Upon examining the results, we compared to Random Forest Classifier with logistic regression, Decision Tree, Bayesian regression and Gradient Boosting algorithms,Random forest classifier gives more accuracy. We have developed Front -End of our flight prediction using Steam lit.

**TABLE OF CONTENTS**

|  |  |  |
| --- | --- | --- |
| **Chapter No:** | **Title** | **Page No.** |
|  | **Declaration** | I |
|  | **Acknowledgement** | II |
|  | **Abstract** | III |
|  | **Table of Contents** | IV-V |
|  | **List of Figures** | VI |
| **1** | **INTRODUCTION** | 1-3 |
|  | 1.1 Problem Definition | 4 |
|  | 1.2 Project Statement | 4 |
|  | 1.3 Problem Overview | 5 |
|  | 1.4 Objectives | 5 |
|  | **System requirements** |  |
|  | 1.5 Hardware Specifications | 6 |
|  | 1.6 Software Specifications | 6 |
| **2** | **LITERATURE SURVEY** | 7-8 |
|  | **Software requirements** |  |
|  | 2.1 NumPy | 9 |
|  | 2.2 Pandas | 9 |
|  | 2.3 Matplot | 9 |
|  | 2.4 Stream lit | 9-10 |
|  | 2.5 Existing System | 11 |
|  | 2.6 Keyword | 12 |
|  | 2.7 Proposed System | 12 |
|  | 2.8 Applications | 13 |
| **3** | **SYSTEM ANALYSIS & DESIGN** |  |
|  | 3.1 Flowcharts | 13 |
|  | 3.2 System Design | 14 |
| **4** | **Modules** |  |
|  | 4.1 Collecting pre-processing Data | 14 |
|  | 4.2 Feature Extraction | 15 |
|  | 4.3 Training the model | 15 |
| **5** | **Algorithms** |  |
|  | 5.1 Logistic Regression | 16 |
|  | 5.2 Gradient Boosting | 17 |
|  | 5.3 Random Forest Classifier | 18 |
|  | 5.4 Bayesian regression | 19 |
|  | 5.5 Decision Tree | 20 |
| **6** | **Code Implementation** |  |
|  | 6.1 User interface | 21-23 |
|  | 6.2 Data Set | 24-27 |
|  | 6.3 Data pre-processing | 27-29 |
|  | 6.4 Data Cleaning | 30 |
|  | 6.5 Data visualization | 31-32 |
|  | 6.6 Train and test process | 32-33 |
|  | 6.7 Evaluation Metrics | 34-38 |
| **7** | **Experimental Results** | 39-43 |
|  | 7.1 Custom testing | 44 |
|  | 7.2 Accuracy table | 44-46 |
| **8** | **WEBSITE** |  |
|  | 8.1 Front End | 47 |
|  | 8.2 Back End | 48-50 |
| **9** | **CONCLUSION** | 51 |
| **10** | **REFERENCES** | 52 |

## LIST OF FIGURES

### Fig Title Page No.

* 1. Logistic Regression 16
  2. Gradient Boosting 17
  3. Random Forest Classifier 18
  4. Bayesian regression 19
  5. Decision Tree 20

## INTRODUCTION

Nowadays, interactions between people on a global scale become more and more important due to improvements in communication and transportation methods. As a result, the plane has become indispensable for transporting people and goods. However, in many cases, you will suffer delays that are out of your control. According to the U.S. Department of Transportation’s Bureau of Transportation Statistics (BTS), almost 20% of the flights arriving at an airport in the USA in 2015 was delayed, which means that approximately one out of five planes arrived more than 15 minutes later than scheduled.

Important causes include aircraft arriving late from other airports (accounting for 40% of total delay minutes in 2015) and airline delays such as maintenance and cleaning(32%). The National Airspace System also caused delays (due to e.g. non-extreme weather conditions or heavy traffic volume, 23%).

Finally, extreme or hazardous weather conditions also caused delays (5%).According to the bureau of transportation statistics(BTS) of the US ,over 20% of the US flight were delayed During 2018.41 billion us $ Financial losses and increases stress.Air transportation system is one of the crucial modes of modern versatility.

With increasing congestion in air traffic passenger-traffic, it is important to maintain persistence and resilience. Availability of land and resources contribute to the infrastructure of airports. The norms of improving technology and procedure are to maintain safety, efficiency, capacity, etc., Therefore, the National Airspace System (NAS) focuses on minimizing the environmental effects as a result of improvisation.With the current technology in hand, passengers can visualize their flightpath, altitude, heading and other related parameters during their journey. However, air-traffic authorities continuously try to depreciate the delay in departure and arrival of flights. Though their efforts were in phase, the outcome is undesirable as the delays are in terms of hours sometimes causing chaos. Some important parameters that cause delay include weather, maintenance, security, and carrier. Corporate travel and tourism are the two major contributors to flight transportation system which is expected to be doubled by 2030. As a result of this increase, the air traffic is also expected to increase in the same multiple. To minimize the air-traffic congestion new airports can be constructed. But, the complexity still grows exponentially. Hence, the only possible way of minimizing the delay is to improvise the existing airports. Considering the limited availability of land resources, the latter is more of a logical Delay basically represents the period by which the aircraft is late or cancelled. Commercial aviation is likely to be affected if there is a delay in their mobility. This delay results in the dissatisfaction of trusted customers and sometimes even marketing strategies. Between 2011 and 2017, the total number of commercial flights operating in a year decreased by around 6.7%, while the total number of delayed minutes increased by the same percentage (Figure 1.1.1). Although the number of flights1has declined since 2013, the ratio between total delay minutes per year to the number of flights has increased. The Federal Aviation Agency (FAA) forecast predicts that flight operations will increase by nearly 35% in the next 20 years, i.e., around 1.5% each year for all commercial airlines (Aviation Administration, n.d.), which is likely to increase the total delays even further.According to the National Center of Excellence for Aviation Operations Research (NEXTOR) Report published in 2010, the total cost (direct and indirect cost) of domestic airline delays to the US economy was around $32 Billion in 2007 (Ball et al., 2010). The total cost to the US economy due to the airline delay was around $26.6 Billion in 2017(Airlines For America, 2018). The cost to airlines fell from $82.2/minute to $68.48/minute, while the cost to passengers increased from $37.6/hour to $49/hour between 2007 and 2017 (Airlines For America, 2018; Ball et al., 2010). Over the next 20 years, the number of passengers traveling via commercial airlines is expected to grow from 840.8 Million in 2017 to 1.Billion (Bureau of Transportation Statistics, 2017).Flight delays could always be annoying, especially in the case when the period of delay was so long that there was even a danger to miss the next flight. However, if there was a way to predict whether there would be a delay or even better – how long the delay could be, then people could make earlier preparation to reschedule following flights in an earlier manner.For that consideration, we adopted a dataset containing airline delayed time and other air liner information provided by Kaggle to building a model, mainly aiming to solve the following questions. Whether there would be a delay with certain publicly reachable resources; and 2. How long delayed timeone could expect with the same information given. We deployed python s k learn and pandas library to build our model, and evaluate our model based on R-Square for linear regression and accuracy rate for logistic regression.As a brief result of our project, we found, it would be helpful to use the following factors in evaluating our model: week, month, airline carrier reference, planned elapsed time (in air time), distance between two departure and destinations, flight planned departure time, departure airport code, and taxi-in and taxi out time.The continuous increase of storage capacities and computational power is currently pulling the development of data analytics.Indeed, companies are collecting massive volume of data , such as web logs, customer information, production and sales tracking, etc. Analyzing these datasets, withdata mining algorithms for example, allows the extraction of information that can help a company to gain knowledge or to use the information as a basis for new products or services.

The primary goal of this paper is to predict the airline delays caused by various factors such as technical issue with aircraft, bad weather, air traffic control problems, bird strikes. Major problem in the current aviation system due to flight delay. Some methods are required to introduce how delay propagates in the airport networks.

Flight delays becomes an important subject and problem for air transportation system all over the world. The aviation industry is continuing to suffer from economic losses associated with the flight delays all the time.

# Problem Definition:

Flight Planning is one of the challenges in industrial world which faces many uncertain conditions. One such condition is delay occurrence, which stems from various factors and imposes considerable costs on airlines, operators, and travelers.

Delays in departure can occur due to bad weather conditions, seasonal and holiday demands, airline policies, technical issue such as problems in airport facilities, luggage handling and mechanical apparatus, and accumulation of delays from preceding flights. Here in-flight delay prediction system based on the weather parameters which can result in delays.

The system considers the temperature, humidity, rain in mm, visibility and month number as important parameters for prediction of delay. One of the key business issues that airlines face is that the vital prices that are related to flights being delayed because of natural occurrences and operational shortcomings that is an upscale affair for the airlines, making issues in scheduling and operations for the end users therefore inflicting unhealthy name and client discontent.

As we all know that we have a tendency to not get the flight delay before departure as customers of the Airline Company neither the airline companies ground staff gets the airline delay prediction supported varied conditions.

However, we all know that one in all the most reasons for delay in flights is that the weather.

We have calculate error calculation results and developed Front-End.

# Problem statement

Flight delay is a major hurdle for airlines which depends on razor thin profit margins and for customers or passengers as well. In this project, we are comparing different regression models and visualizing them based on multiple regression metrics.

# Project Overview

Attribute selection takes the attributes which are useful for the prediction of the heart disease. After identifying the data from the available resources, they are further selected for processing which includes data cleaning. Different classification algorithms are performed on the pre- processed data to find the chance of getting heart disease. It also finds the accuracy of the algorithms and compares the accuracy among the algorithms. In addition to the prediction, it also produces the various reasons for the disease or the diagnosis report from which treatment can be suggested easily

# Objectives

The objective of the project is to perform analysis of the historic flight data to gain valuable insights and build a predictive model to predict whether a flight will be delayed or not given a set of flight characteristics. Questions to be answered post analysis:

* + - Does the month of flight have any impact on flight delays?
    - Flights to which destination have seen the most delays?
    - Which day of the week sees the least and most flight delays?
    - Which time of day is most suitable for preventing flight delays?
    - Which airline has the greatest number of flights delayed?
    - What are the primary causes for flight delays?

The objective of the predictive model is to build a model to predict whether a flight will be delayed or not based on certain characteristics of the flight. Such a model may help both passengers as well as airline companies to predict future delays and minimize them.

**System requirements**

# Hardware Specifications

* + - CPU: 2 x 64-bit 2.8 GHz 8.00 GT/s CPUs
    - RAM: 2 GB (or 4 GB of 1600 MHz DDR3 RAM)
    - Storage: 300 GB. (600 GB for air-gapped deployments.) Additional space recommended if the repository will be used to store packages built by the customer. With an empty repository, base install requires 2 GB.
    - Internet access to download the files from Anaconda Cloud or a USB drive containing all of the files you need with alternate instructions for air gapped installations.

# Software Specifications

* + - Anaconda Prompt Software
    - Jupyter notebook
    - Python 3.7.1 and
    - Vs Code

# LITERATURE SURVEY

|  |  |  |  |
| --- | --- | --- | --- |
| **Authors Name** | **Title of the Paper** | **Methodology** | **Future scope** |
| Y. Jiang, J. Miao, X. Zhang and N. Le | A multi-index prediction method for flight delay based on long short-term memory network model | The model takes into account a wide variety of variables, including flight volume, scheduled departure and arrival times, actual departure and arrival times, flight features, airport and city weather, etc. | The model is able to accurately forecast outcomes and follow the trends of several delay indicators. The technique can predict shifts in delay trends, providing aircraft scheduling managers with actionable data. |
| J. Wang and W. Pan | Flight delay prediction based on ARIMA | Air travel has expanded rapidly, making flight delays a major issue for civil aviation. There are a number of causes for flight delays, all of which have a negative impact on airlines' ability to run smoothly and provide excellent service. | The findings demonstrate that the ARIMA model is a strong fit for the data on flight delays, that it can guarantee the precision of flight delay prediction, and that it can provide useful guidance for addressing flight delays in advance. |
| J. Tao, H. Man and L. Yanling | Flight delay prediction based on LightGBM | Despite the progress made in the field of civil aviation, the incidence of aircraft delays remains high. That's why it's crucial for several airlines to be able to anticipate when flights will be delayed. U.S. Bureau of Transportation Statistics data was used for the analysis (BST). | To anticipate flight delays, the authors of this work propose using the LightGBM algorithm and optimizing its parameters via grid search and cross-validation. Results from experiments reveal that compared to XGBoost and GBDT, LightGBM achieves better prediction accuracy and requires less time to train than its rivals. |
| I. M. Almaameri and A. Mohammed | Predicting Airplane Flight Delays Using Neural Networks | Predicting delays in the aviation network is a critical part of the flight planning process. Knowing where and how often delays occur in the airline network is essential for making it more resilient to disruptions. When bottlenecks in the network are identified, measures might be considered to absorb the extra time. | Predicting delays in the aviation network is a critical part of the flight planning process. Knowing where and how often delays occur in the airline network is essential for making it more resilient to disruptions. When bottlenecks in the network are identified, measures might be considered to absorb the extra time. |
| S. Harinandene and S. Muthukumar | Flight Delay Forecasting Based on Airport Networks | Constant growth in demand for air travel has outpaced the capacity of the current infrastructure, leading to dubious flight plans such as lengthy flight delays and unpredictability in landing, takeoff, and taxi pickup and drop-off times. To solve this problem, experts devised a model that accounts for airport throughputs while also anticipating flight delays. | Predictions and schedules are analyzed using Big Data technologies to cut down on wait times as much as possible. Using Spark technology, which is being proposed for the future, the datasets can be made real-time, which significantly cuts down on flight delays. |

# Software requirements

* 1. **NumPy**

NumPy is a Python library used for working with arrays. It also has functions for working in domain of linear algebra, Fourier transform, and matrices. NumPy was created in 2005 by Travis Oliphant. It is an open-source project and you can use it freely. NumPy stands for Numerical Python. In Python we have lists that serve the purpose of arrays, but they are slow to process. NumPy aims to provide an array object that is up to 50x faster than traditional Python lists. The array object in NumPy is called np array, it provides a lot of supporting functions that make working with np array very easy. Arrays are very frequently used in data science, where speed and resources are very important. NumPy arrays are stored at one continuous place in memory unlike lists, so processes can access and manipulate them very efficiently. This is the main reason why NumPy is faster than lists.

# Pandas

Pandas is a Python library used for working with data sets. It has functions for analyzing, cleaning, exploring, and manipulating data. The name "Pandas" has a reference to both "Panel Data", and "Python Data Analysis" and was created by Wes McKinney in 2008. Pandas allows us to analyze big data and make conclusions based on statistical theories. Pandas can clean messy data sets, and make them readable and relevant. Relevant data is very important in data science. Pandas are also able to delete rows that are not relevant, or contain wrong values, like empty or NULL values. This is called cleaning the data

## Matplot

Matplotlib is a plotting library for the Python programming language and its numerical mathematics extension NumPy. It provides an object-oriented API for embedding plots into applications using general-purpose GUI toolkits like Tkinter, wxPython, Qt, or GTK. There is also a procedural "pylab" interface based on a state machine (like OpenGL), designed to closely resemble that of MATLAB, though its use is discouraged. SciPy makes use of Matplotlib. Matplotlib was originally written by John D. Hunter. Since then it has had an active development community and is distributed under a BSD-style license.

**2.4 Streamlit**

Streamlit is a free and open-source framework to rapidly build and share beautiful machine learning its

data science web apps. It is a Python-based library specifically designed for machine learning engineers.

With Streamlit, the developer is able to focus on implementing their back end logic while leaving the and

Frontend implementation largely to the framework itself. In addition, using Streamlit you can render respo-

nsive interfaces for PC, tablet, and mobile platforms at your fingertips with no added overhead.Streamlit is

an open source app framework in Python language. It helps us create web apps for the data science and ma-

chine learning in a short time.

# 2.5 Existing system

* It considers the flight dataset with and data cleaning is done.
* Data cleaning is the process of detecting and correcting corrupt or inaccurate records from a record set.
* Fitting this data with some machine learning models and finding out the most accurate one.
* Training and testing the data with the best model and thus predicting the disease.It also find the accuracy of the algorithms and compares the accuracy among all the algorithms.
* The prediction, analysis and cause of flight delays have been a major problem for air traffic control, decision-making by airlines and ground delay response programs. Studies are conducted on the delay propagation of the sequence. In the past, researchers have tried to predict flight delays with Machine Learning algorithms like logistic regression, decision tree etc all are used for classification purposes only.

# Advantages

* High accuracy.
* Low complexities.
* Low cost.
* Can predict actual delay figures.

# Disadvantages:

* High complexities.
* High cost.
* Inability to predict actual delay figures.

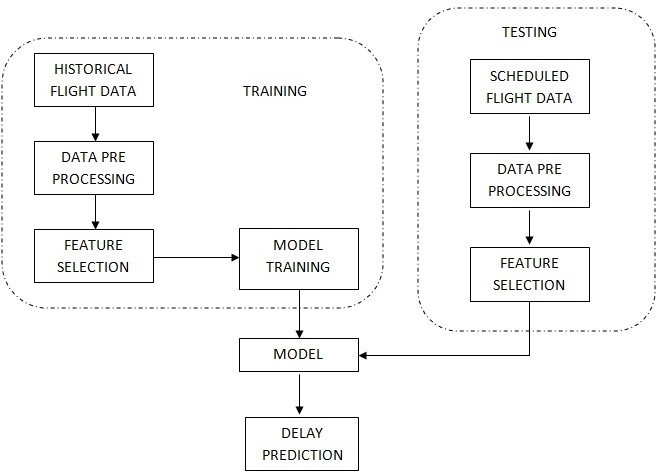
## 2.6 KEYWORDS:

Flight Prediction, Machine Learning, Logistic Regression, Decision Tree, Bayesian Ridge, Random Forest, Gradient Boosting, U.S. Flight Data.

# 2.7 Proposed System

Most of the works mentioned before uses classification or unsupervised learning techniques for flight delay prediction. In this project, we will compare Decision Tree, Bayesian ridge, Random forest and Gradient boosting algorithms all of them are of supervised learning methods. We use Mean squared error, mean absolute error, explained variance, Median absolute error and R2 score to evaluate these models instead of accuracy which is used in unsupervised learning.

# BLOCK DIAGRAM:



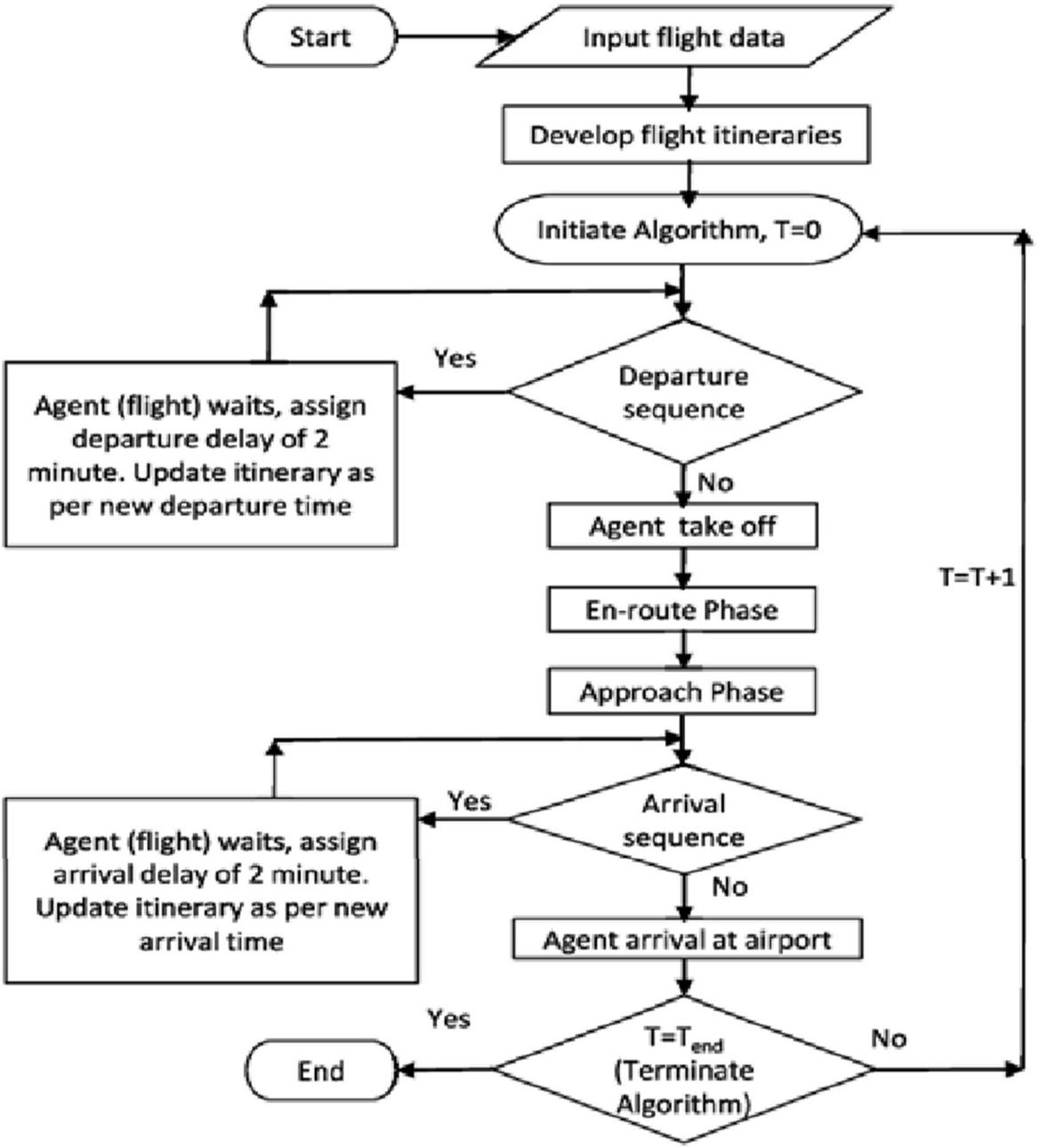
The above diagram shows the training and testing of the model with feature selection by isolating the most consistent,non-redundant,and relevant features to use in model construction.

## 2.8 APPLICATIONS:

* + - Used by airline industry to better predict airline delays and manage their services based on the predictions.

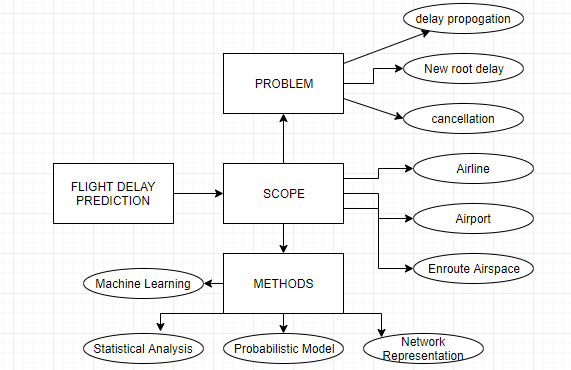
**3 System Analysis and Design**

**3.1 Flow chart:**



The above flowchart considers the input flight data and other features given and shows the conditions in case of arrival time and departure time. After entering the flight data, the flight schedules the flight ticket in an actual document that allows to board the plane. Then initial algorithm is sent to the departure sequence, if the departure sequence is true then it happens as shown in flowchart or else it undergoes the above steps and then goes to arrival sequence. If the arrival sequence is true then it happens as shown in flowchart or else it goes to agent arrival at airport and then terminate algorithm is produced.

# System design:



The above figure shows the data pre-processing, cleaning, data visualization, modeling, and the delay prediction. After unprocessed data set is imported to the pre processing then it is send to cleaning ,after cleaning it under goes data visualization. Next train and test the model with the given data.

**4 Modules**

**4.1: Collecting and preprocessing data**

* We used data collected by the U.S. Department of Transportation's Bureau of Transportation Statistics on all domestic flights taken in 2015 to train models for predicting flight delays. There are 25 columns and 59986 rows total in the dataset.
* A large number of lines had blank or null values. It is necessary to preprocess the data before using it.
* Since real-world data is often unreliable due to its incompleteness, noise, and inconsistency, it is necessary to perform data preprocessing before conducting any analysis.
* There are 25 columns and 59986 rows total in the dataset.
* Numerous rows had blank or empty cells. In order to clean up the data set, the null values were removed using the dropna() function in pandas.
* The number of rows was cut down to 54486 after some preprocessing.

**4.2 : Feature extraction**

* Feature extraction is a process of dimensionality reduction by which an initial set of raw data is reduced to more manageable groups for processing,A characteristic of these large data sets is a large number of variables that require a lot of computing resources to process.
* Feature extraction is the name for methods that select and /or combine variables into features, effectively reducing the amount of data that must be processed, while still accurately and completely describing the original data set.
* The process of feature extraction is useful when you need to reduce the number of resources needed for processing without losing important or relevant information.

**4.3 : Training the model**

* Model training is the primary step in machine learning, resulting in a working model that can then be validated, tested and deployed. The model’s performance during training will eventually determine how well it will work when it is eventually put into an application for the end-users.
* Both the quality of the training data and the choice of the algorithm are central to the model training phase. In most cases, training data is split into two sets for training and then validation and testing.
* The selection of the algorithm is primarily determined by the end-use case. However, there are always additional factors that need to be considered, such as algorithm-model complexity, performance, interpretability, computer resource requirements, and speed. Balancing out these various requirements can make selecting algorithms an involved and complicated process.

# 5 Algorithms Used

## 5.1 Logistic Regression:

A large part of data analysis boils down to a simple question: is something “A” or “B?” Is it “positive” or “negative?” Is this person a “potential customer” or “not a potential customer?” Machine learning accommodates such questions through logistic equations, and specifically through what is known as the sigmoid function. The sigmoid function produces an S-shaped curve that can convert any number and map it into a numerical value between 0 and 1, but it does so without ever reaching those exact limits. A common application of the sigmoid function is found in logistic regression. Logistic regression adopts the sigmoid function to analyze data and predict discrete classes that exist in a dataset. Although logistic regression shares a visual resemblance to linear regression, it is technically a classification technique. Whereas linear regression addresses numerical equations and forms numerical predictions to discern relationships between variables, logistic regression predicts discrete classes.

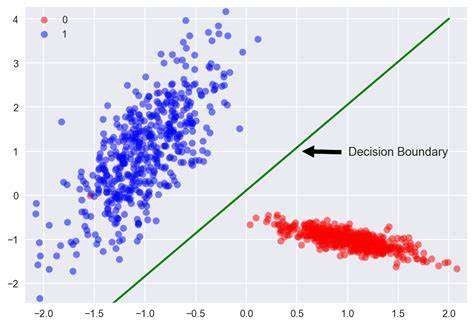


Fig 5.1

In the above figure,Logistic regression is shown by decision boundary line.

## 5.2 Gradient boosting

The idea of gradient boosting originated in the observation by Leo Breiman that boosting can be interpr-

eted as an optimization algorithm on a suitable cost function. Explicit regression gradient boosting algori

thms were subsequently developed, by Jerome H. Friedman simultaneously the more general functional

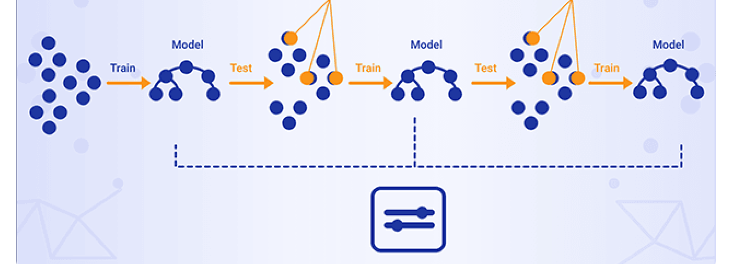
gradient boosting perspective of Llew Mason, Jonathan Baxter, Peter Bartlett and Marcus Frean.The latt-

er two papers introduced the view of boosting algorithms iterative functional gradient descent algorithms.

That is,algorithms that optimize a cost function over function space by iteratively choosing a function that

points in the negative gradient direction. This functional gradient view of boosting has led to the developm

ent of boosting algorithms in many areas of the machine learning and statistics beyond regression and classifications.Gradient Boosting is a powerful boosting algorithm that combines several learners into strong learners, in which each new model is trained to minimize the loss function such as mean squared error or cross-entropy of the previous model using gradient descent. In each iteration, the algorithm computes the gradient of the loss function with respect to the predictions of the current ensemble and then trains a new weak model to minimize this gradient. The predictions of the new model are then added to the ensemble, and the process is repeated until a stopping criterion is met.Generally, most supervised learning algorithms are based on a single predictive model such as linear regression, penalized regression model, decision trees, etc. But there are some supervised algorithms in ML that depend on a combination of various models together through the ensemble. In other words, when multiple base models contribute their predictions, an average of all predictions is adapted by boosting algorithms.



## The above figure shows the implementation of Gradient boosting

.

# 5.3 Random Forest Classifier

Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks, that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. Random decision forests correct for decision trees’ habit of over fitting to their training set. A decision tree is a Machine Learning algorithm capable of fitting complex datasets and performing both classification and regression tasks. The idea behind a tree is to search for a pair of variable-value within the training set and split it in such a way that will generate the "best" two child subsets. The goal is to create branches and leaves based on an optimal splitting criteria, a process called tree growing. Specifically, at every branch or node, a conditional statement classifies the data point based on a fixed threshold in a specific variable, therefore splitting the data.The algorithm used to train a tree is called CART(®) (Classification And Regression Tree). As we already mentioned, the algorithm seeks the best feature–value pair to create nodes and branches. After each split, this task is performed recursively until the maximum depth of the tree is reached or an optimal tree is found. Trees have a high risk of overfitting the training data as well as becoming computationally complex if they are not constrained and regularized properly during the growing stage. This overfitting implies a low bias, high variance trade-off in the model. Therefore, in order to deal with this problem, we use Ensemble Learning, an approach that allows us to correct this overlearning habit and hopefully, arrive at better, stronger results.

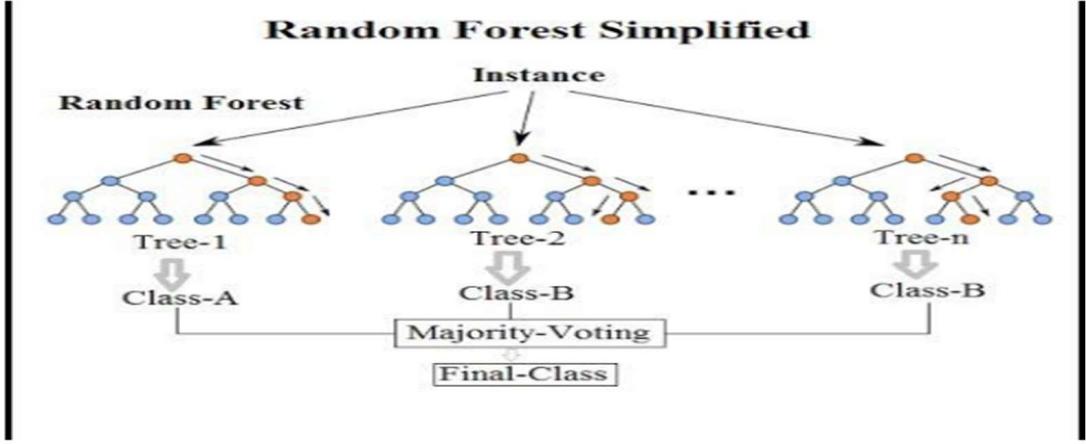


Fig 5.3

The above figure shows the implementation of Random Forest classifier.

**5.4 Bayesian regression**

Bayesian regression is a type of conditional modeling in which the mean of one variable is described by a

linear combination of other variables, with the goal of obtaining the posterior probability of the

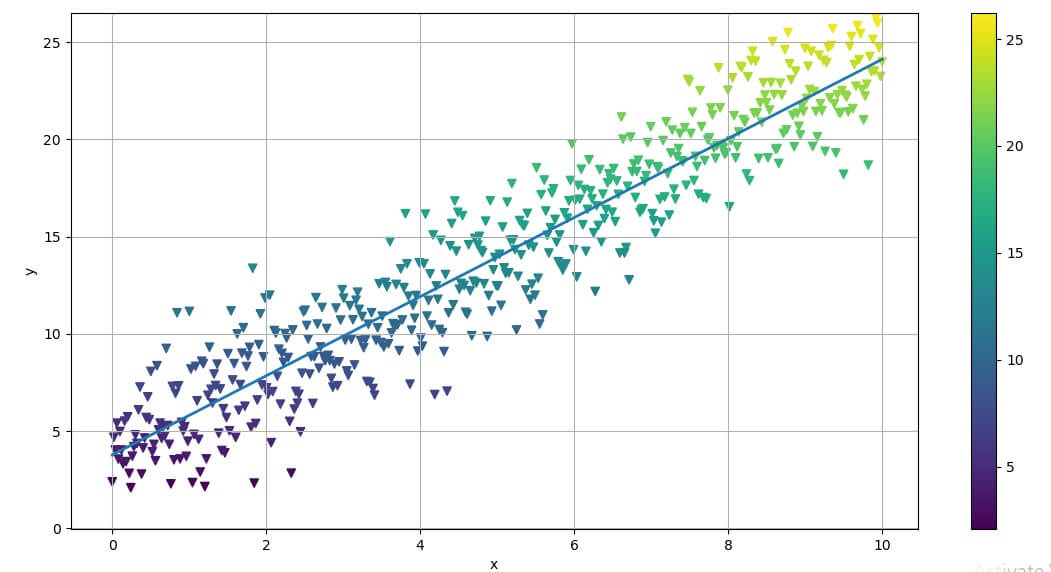
regression coefficients (as well as other parameters describing the distribution of the regressand) and

ultimately allowing the out-of-sample prediction of the regressand (often labelled y) conditional on observed

values of the regressors (usually X). The simplest and most widely used version of this model is the normal

linear model, in which y given X is distributed Gaussian. In this model, and under a particular choice

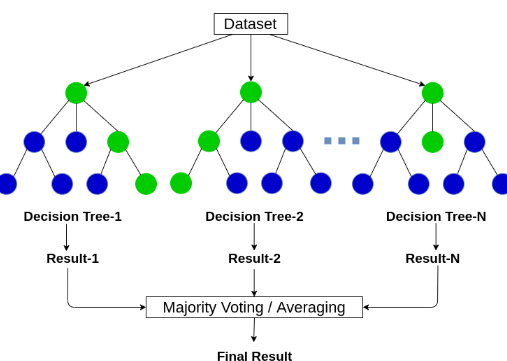
of prior probabilities for the parameters—so-called conjugate priors—the posterior can be found analytically. With more arbitrarily chosen priors, the posteriors generally have to be approximated. When the regression model has errors that have a normal distribution, and if a particular form of the prior distribution is assumed, explicit results are available for the posterior probability distributions of the model’s parameters.The most common interpretation of Bayes’ formula in [finance](https://laconicml.com/finance-with-python-convex-optimization/) is the diachronic interpretation. This mainly states that over time we learn new information about certain variables or parameters of interest, like the mean return of a time series.Equation 1 states the theorem formally. Here, H stands for an event, the hypothesis, and D represents the data an experiment or the real world might presentDoing Bayesian regression is not an algorithm but a different approach to statistical inference. The major advantage is that, by this Bayesian processing, you recover the whole range of inferential solutions, rather than a point estimate and a confidence interval as in classical regression.Bayesian analysis is a statistical method that allows researchers (decision makers) to take into account data as well as prior beliefs to calculate the probability that an alternative (decision, treatment) is superior.



The above figure shows the implementation of Bayesian regression

**5.5 Decision Tree**

As the name suggest the main idea behind decision tree algorithm is to make a tree like structure and get the answers in form of true or false. The model begins from a root node and ends on the decision. Each node receives a Yes or No question and answer is passed on to the next node. Root node gets all the input of the training dataset.Decision Tree is a **Supervised learning technique**that can be used for both classification and Regression problems, but mostly it is preferred for solving Classification problems. It is a tree-structured classifier, where**internal nodes represent the features of a dataset, branches represent the decision rules** and **each leaf node represents the outcome.**In a Decision tree, there are two nodes, which are the **Decision Node** and**Leaf Node.** Decision nodes are used to make any decision and have multiple branches, whereas Leaf nodes are the output of those decisions and do not contain any further branches.The decisions or the test are performed on the basis of features of the given dataset.It is a graphical representation for getting all the possible solutions to a problem/decision based on given conditions.It is called a decision tree because, similar to a tree, it starts with the root node, which expands on further branches and constructs a tree-like structure.Decision Trees usually mimic human thinking ability while making a decision, so it is easy to understand.The logic behind the decision tree can be easily understood because it shows a tree-like structure.In a decision tree, for predicting the class of the given dataset, the algorithm starts from the root node of the tree. This algorithm compares the values of root attribute with the record (real dataset) attribute and, based on the comparison, follows the branch and jumps to the next node.For the next node, the algorithm again compares the attribute value with the other sub-nodes and move further. It continues the process until it reaches the leaf node of the tree.



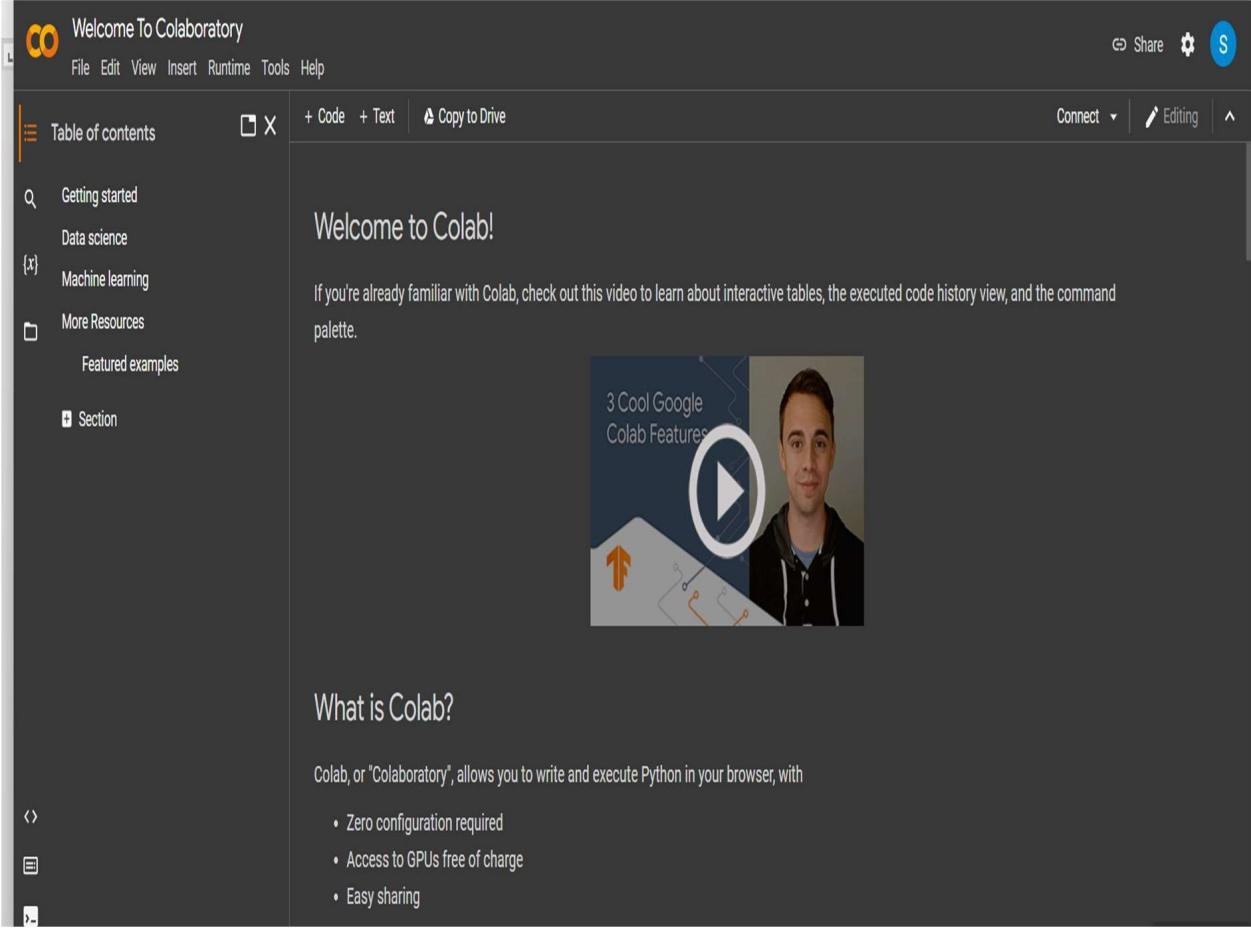
The above figure shows the implementation of Decision Tree

# 6 Code Implementation

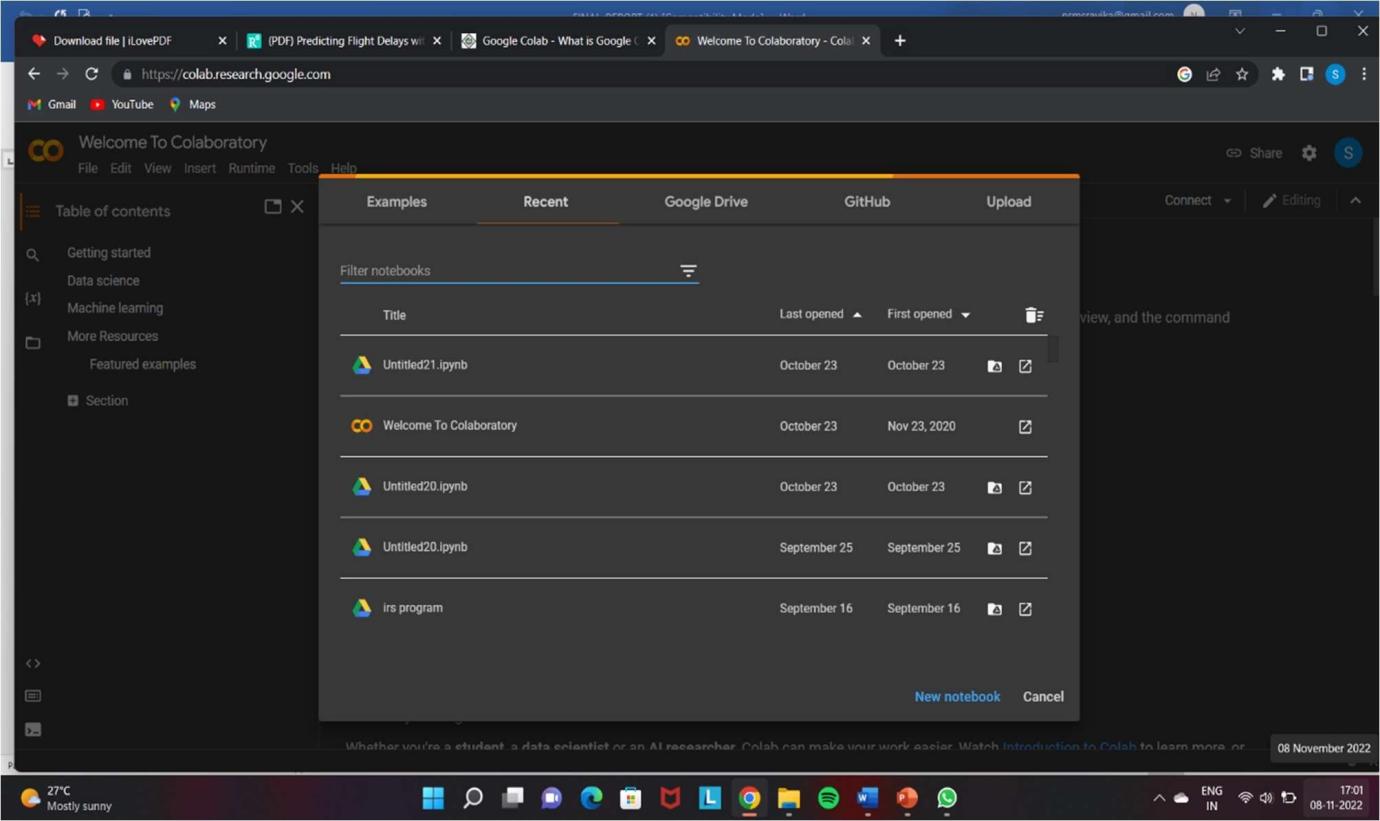
## 6.1 User Interface

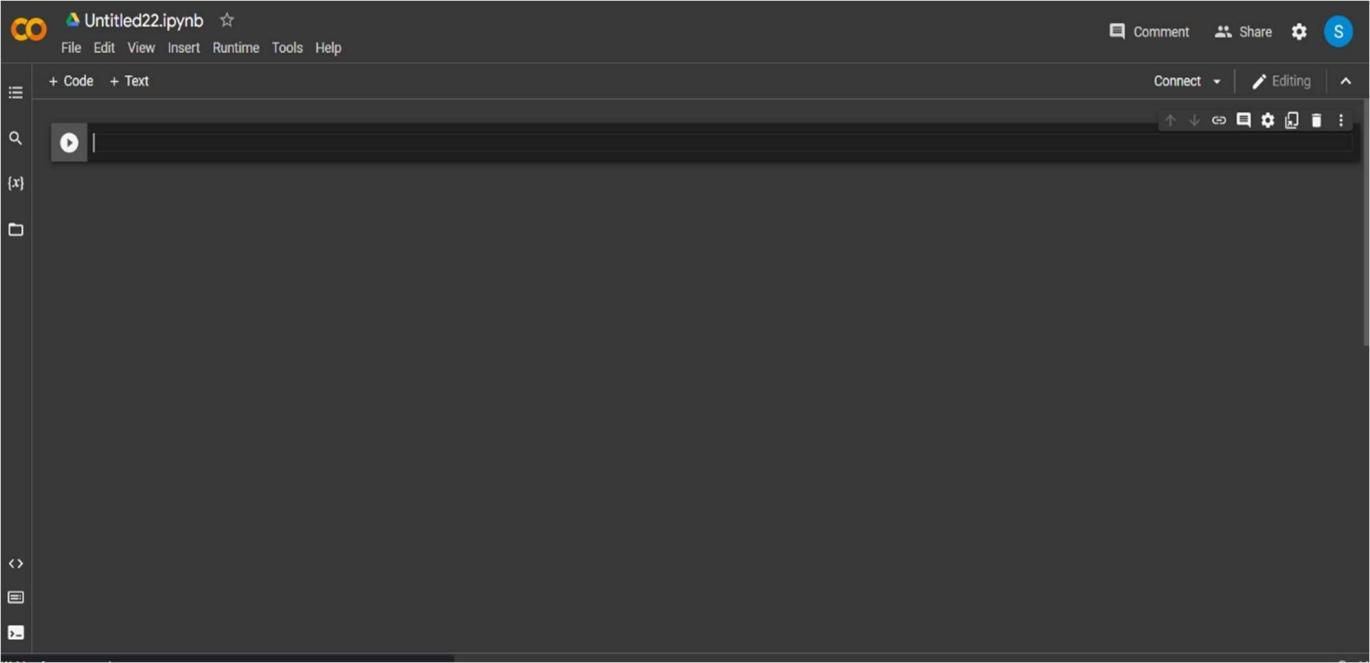
The overall program code does the prediction of flight delays with error calculation using machine learning classifiers for an instance data of a flight delay using the machine learning the algorithms such as Logistic Regression, Random Forest Classifier, Bayesian regression, Decision tree and gradient boosting classifier. Google colab is used to implement the entire project. Colab is a free Jupyter notebook environment that runs entirely in the cloud. Colab supports many popular machine learning libraries which can be easily loaded in your notebook. Most importantly, it does not require a setup and the notebooks that you create can be simultaneously edited by your team members

* just the way you edit documents in Google Docs. Google Colab provides the features that we cam write and execute our code in python. we can create/upload/share our notebooks to others for verification and validation.one of the main feature is we are able to import external datasets.



## Execution

* + Create a File with name Mini project with the extension of ipynb

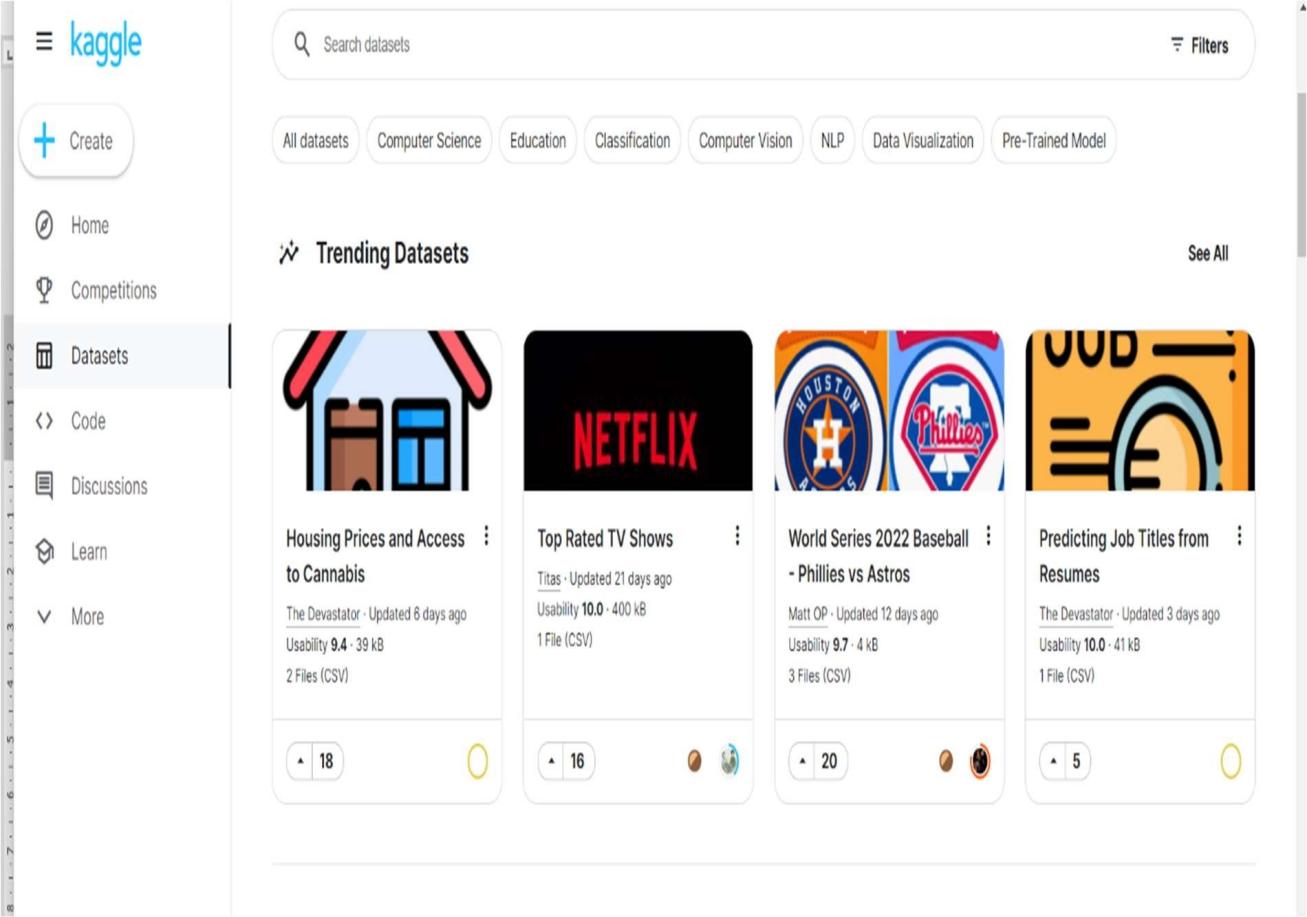
After opening the Collaboratory it will display in the below format to create a new notebook.

* Here we write our python code and run. This is the new ipynb notebook.we select language as python in the home page .

## Collecting the Datasets from Kaggle

1.First, we must go to the google chrome and have to search for the flight delay dataset. 2.It will display some website called Kaggle where we can find the desired data.

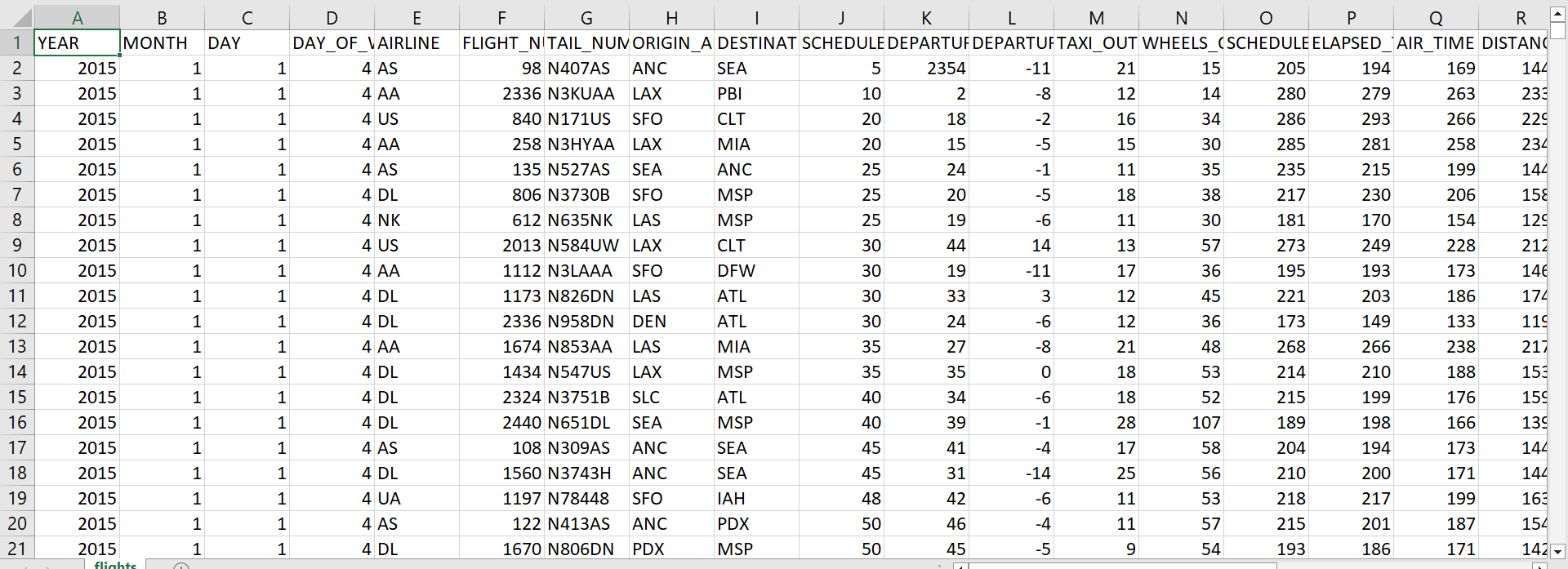
1. Click on the link open the website.
2. There will be many data sets on the delays displayed in the Kaggle. 5.Choose the on which best suits to our project and click on download.



1. The datasets are downloaded in the form of .csv. ex: flight delay.csv.
2. There are many datasets specific to different domains like netfilx,world series baseball 2022,house prices etc.
3. We have to move to top of the page; we will find search bar.
4. Our project is based on flight delays.so we downloaded the data set called Feb-20 -usflight delay. 10.We consider only data set that contains one data file for our project.

## 6.2 Data set

**Flights\_dataset.csv**

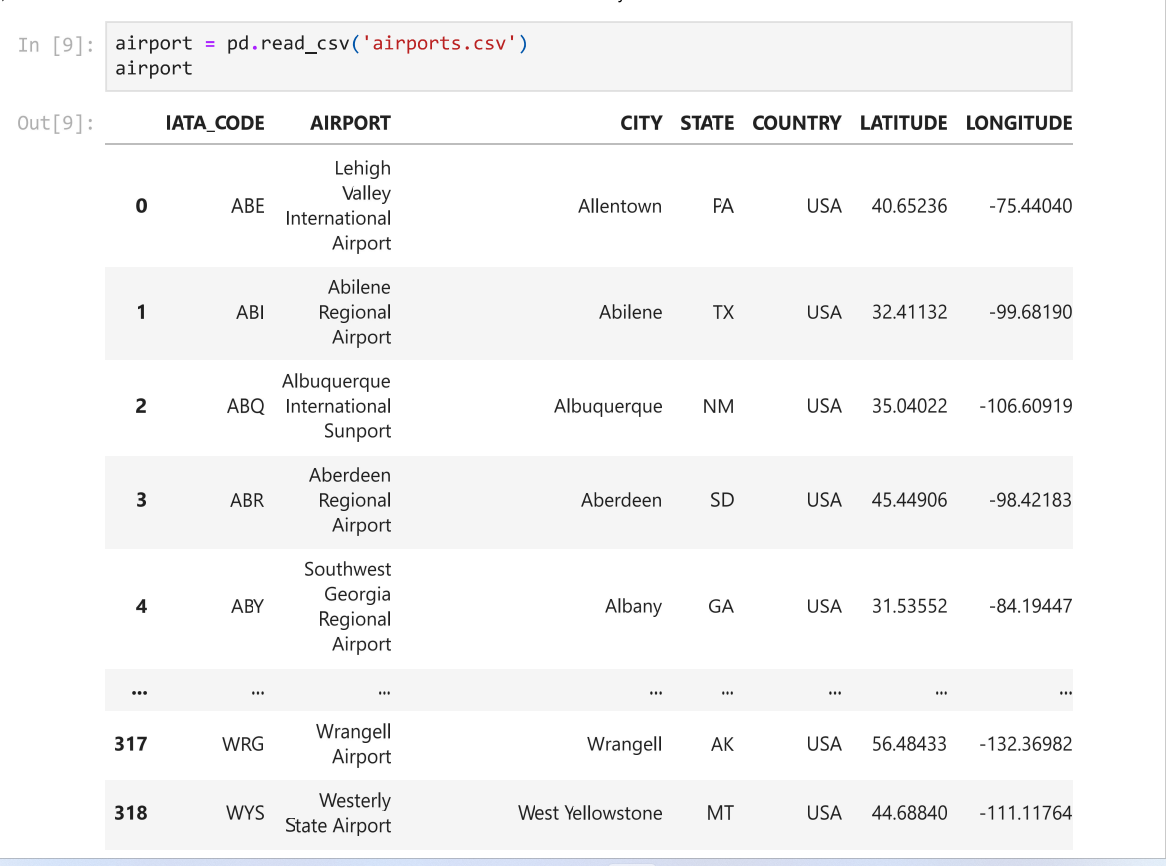


* + These are first 20 samples of our data set.
  + There are total 100000 samples in our data set.
  + Our data set contains total 31 columns.
  + The data set contains the columns like month, day, ORIGIN\_AIRPORT

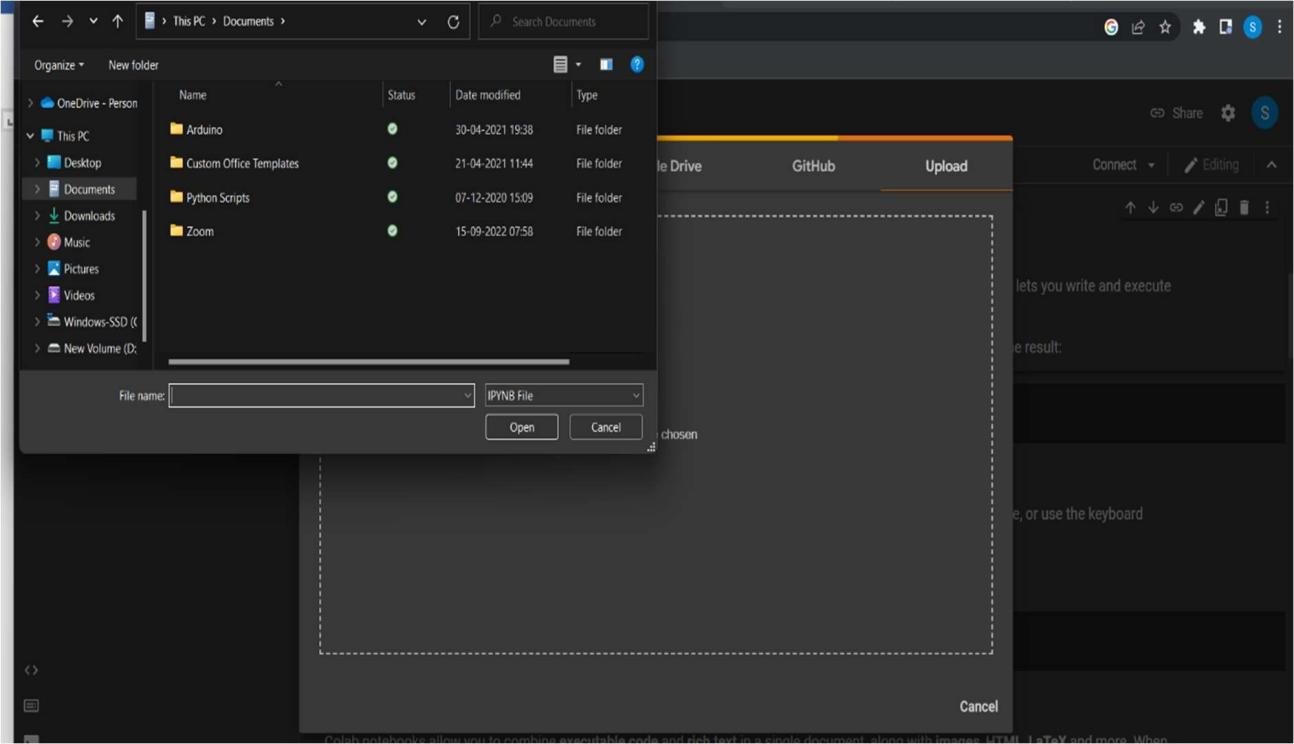
DAY\_OF\_WEEK, DESTINATION\_AIRPORT and distance.

**Aiports dataset.csv**

* Mixed data sets are used in the project.
* One is flights data set and the other is airports dataset.
* In the airports dataset ,totally there are 322 rows and 7 columns.



## Upload The Data Set



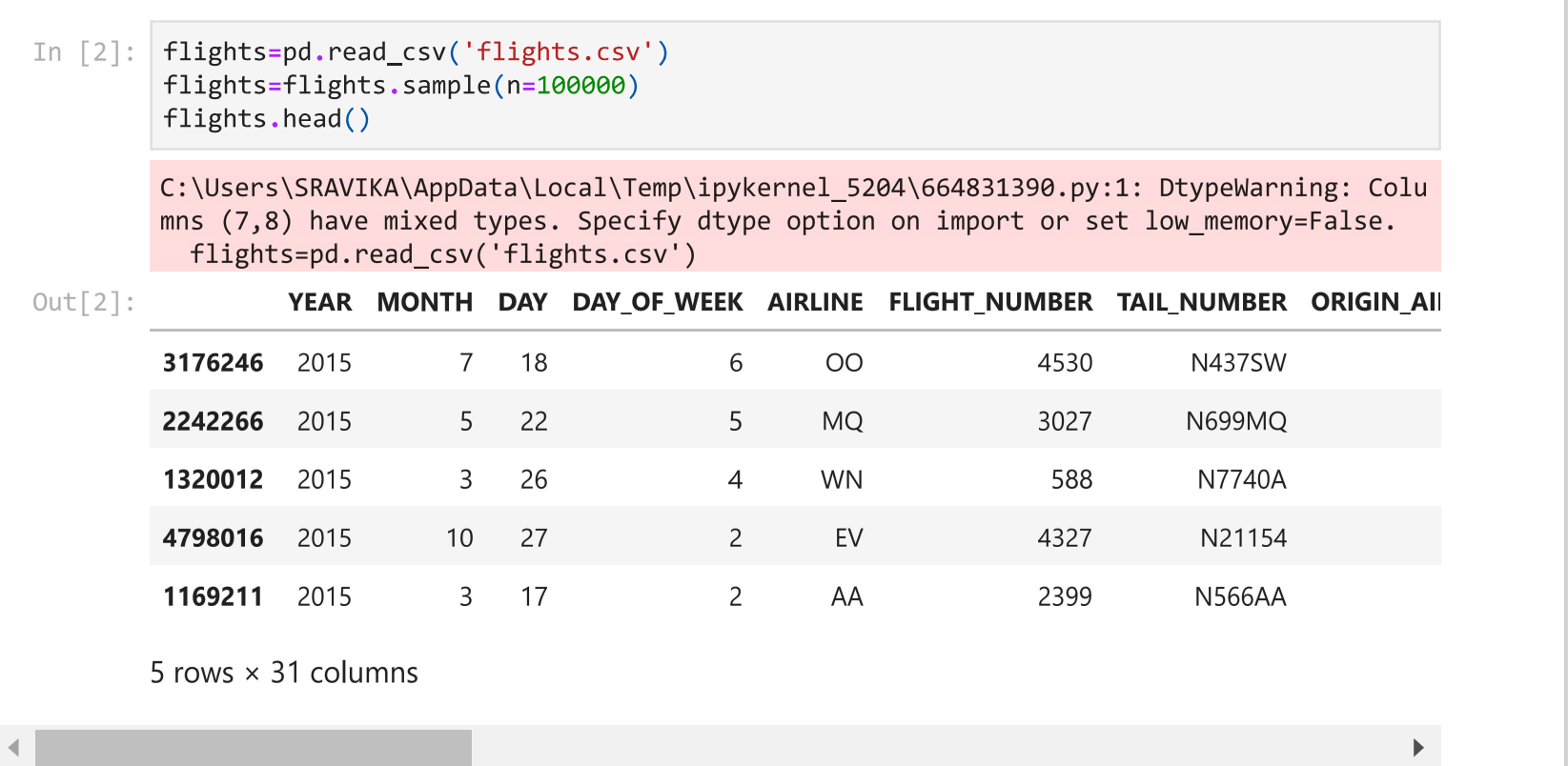
**Install necessary packages and import libraries.**

* + - Import the libraries like pandas
    - Pandas is used for the analysing and cleaning of data set
    - Install thee packages buy using the !pip command
    - Once the installation is done successfully it will display like successfully installed.



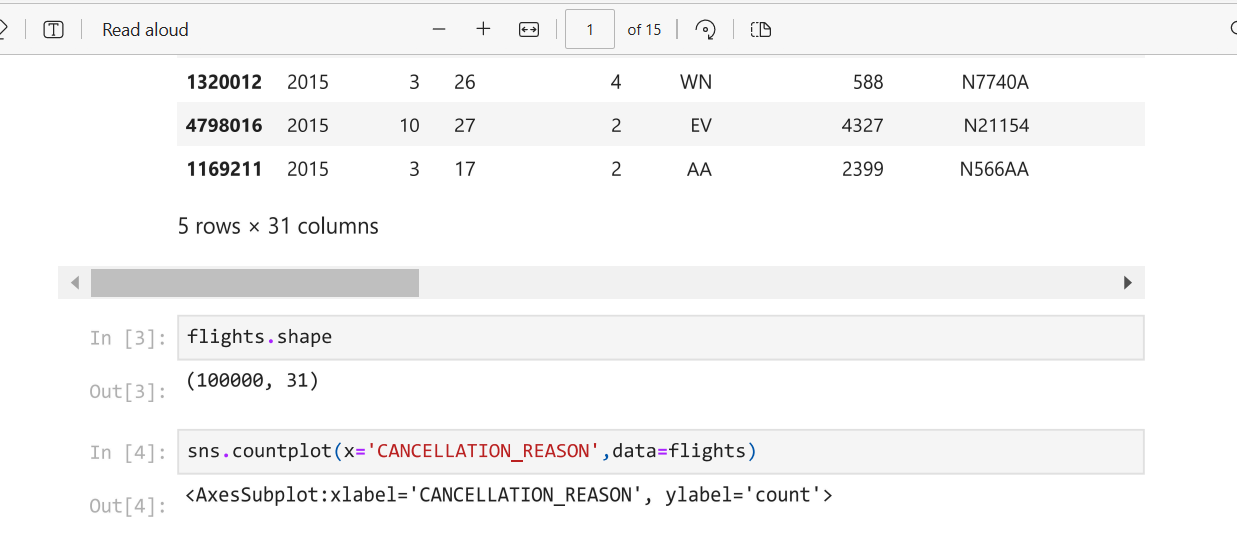
## Reading The Dataset

* + - By using ‘pd.read\_csv we are uploading our data set
    - this function will convert the csv file into the pandas data frame.
    - Our data set contains totally 5 X 31 columns samples.

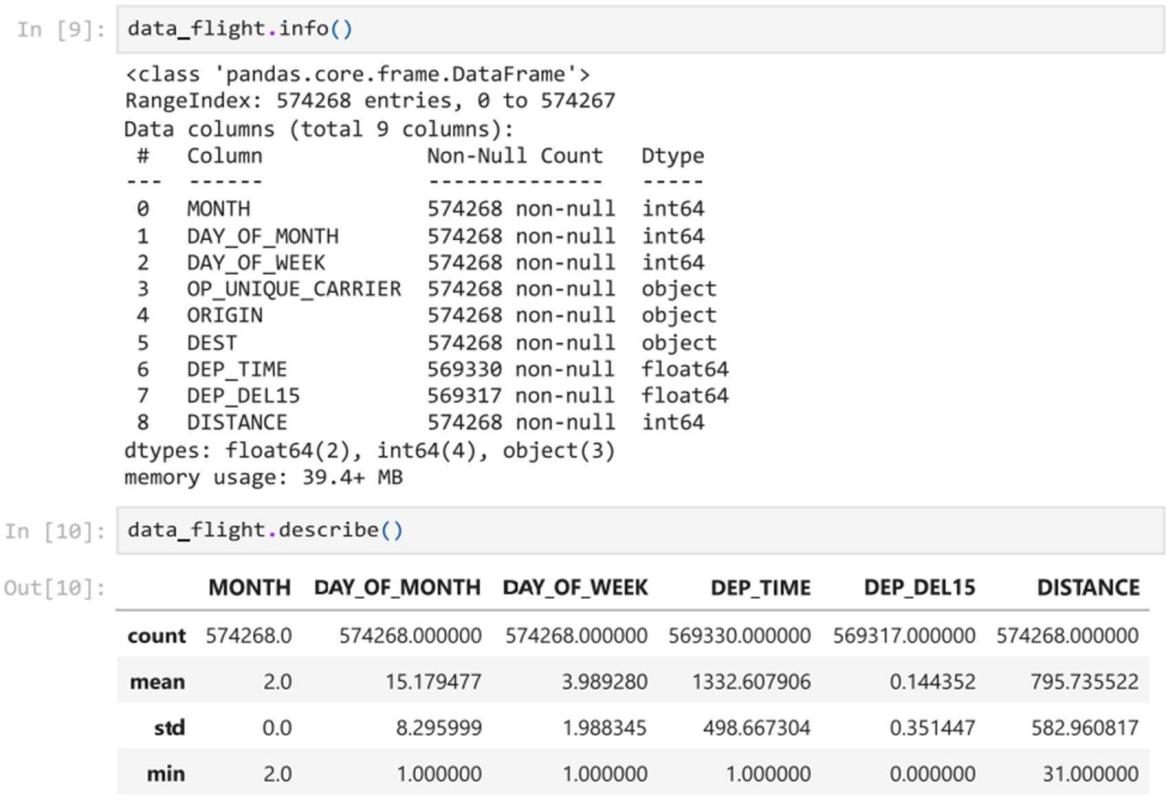


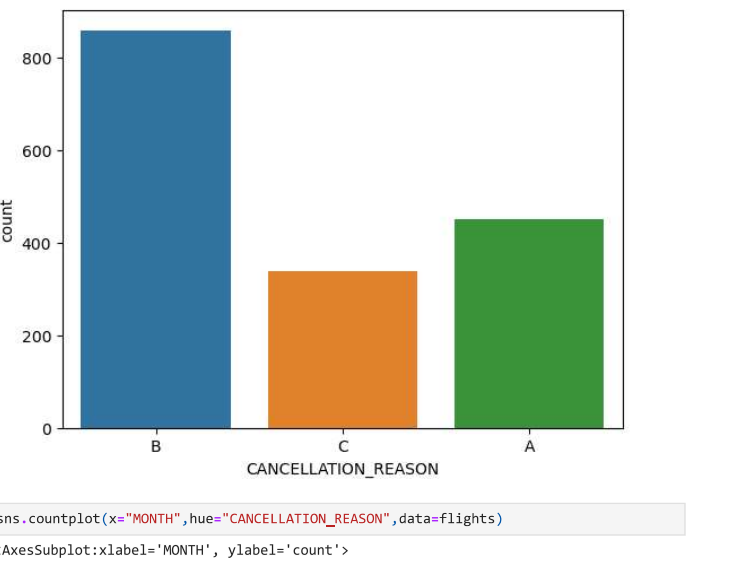
## 6.3 Data pre-processing

* Pre processing of the data set includes the modification of some unwanted samples.



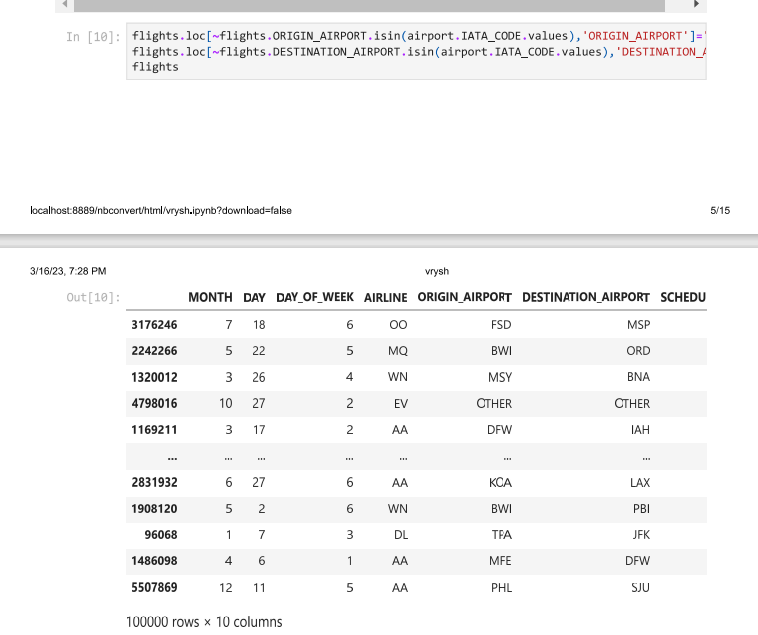
* + This step helps to the model to get good performance and accuracy.
  + This phase includes some of the functions like shape(), head ,tail() will display the first and last 5 rows of the data set.
  + **Data\_flight.info()** returns the data type of the columns and also the column names the null value count



* + There will be the one column which contains the missing values. That is ‘DEP\_DEL15’.
  + Here uses the binary classification ‘**1’** for the flights which are delayed.
  + And 0 for the flights which are arrived on time.
  + Bar plot is drawn to indentify the cancellation reason
  + 
* In the bar plot ,A-represents airline/carrier problems
* B-weather conditions
* C-national air systems
* In the above graph,the plot b is very high.
* We can observe from grapg easily that mostly weather is responsible for delays of fligh

**6.4 Data cleaning:**

Eliminating missing values from the data set.



* + - Detecting and correcting corrupt or inaccurate records from a record set,table,or database
    - After cleansing,it makes sure the dataset is consistent with other similar datasets in the system.

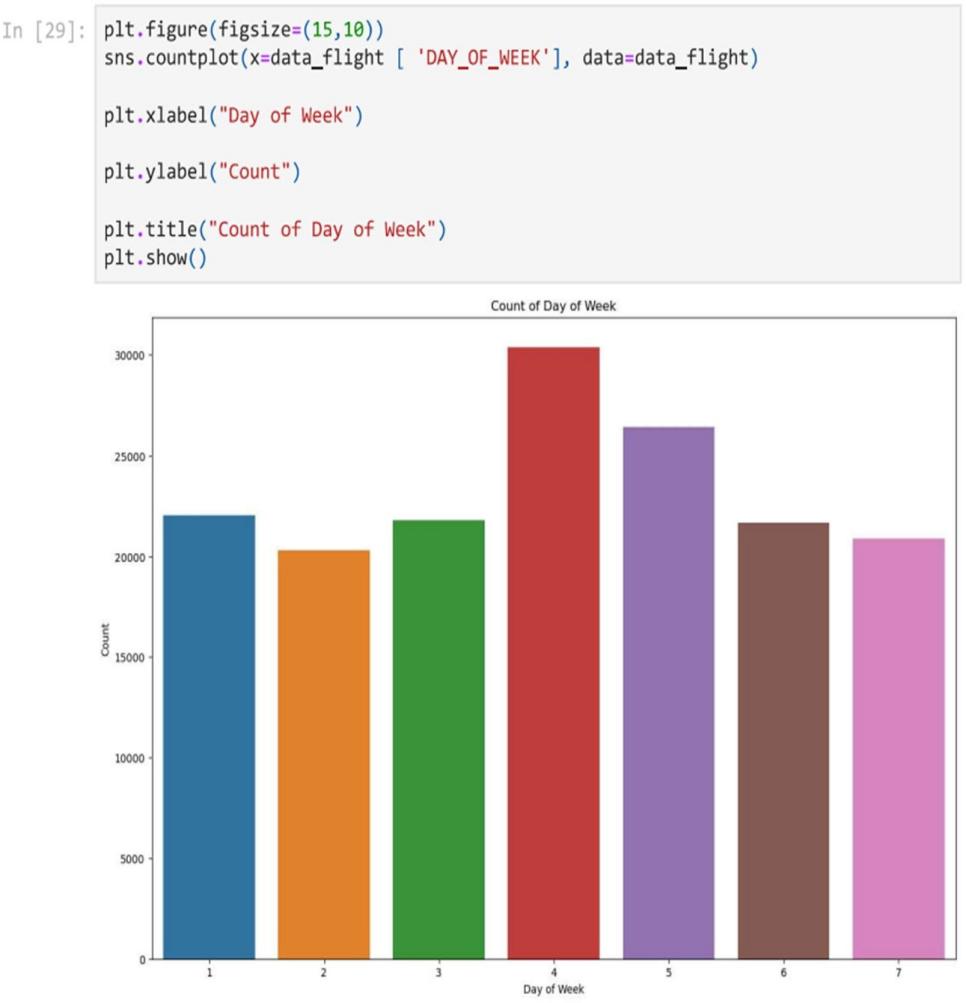
## 6.5 Data Visualization

* Data visualization is the graphical representation of the data using some the visual elements like graphs, pie charts etc.



* This the distribution of distance graph with respect to the departure delay.





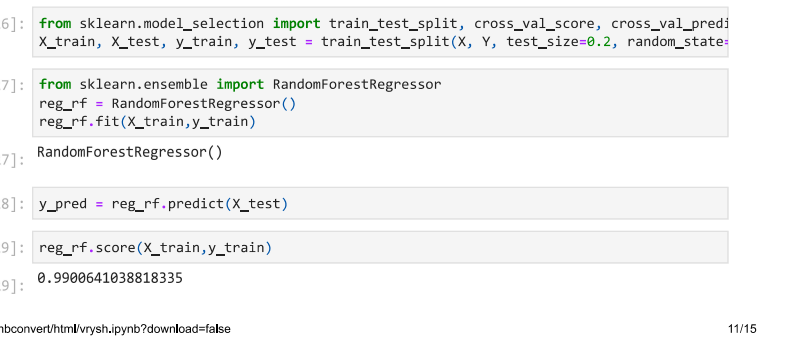
* This is the bar graph of the count of day of week.

# 6.6 Train and test process:

* + - Scikit learn imports the train-test-split helper function.
    - Train-test-split function is used to split the data frame into a training set containing 80% of the original data,and a test set containing remaining 20%.
    - X\_train corresponds to the input features.
    - Y-train labels columns which is a desired output.
    - train x.shape display the number of rows and columns in the data frame containing the feature columns for training.



**Testing training using random forest regressor:**



.

**6.7 EVALUATION METRICS**

**Error calculation results:**

The mean absolute error(MAE),mean squared error (MSE) and root mean square error (RMSE) is calculated using the below formulae.



where,

y = Actual Arrival Delay

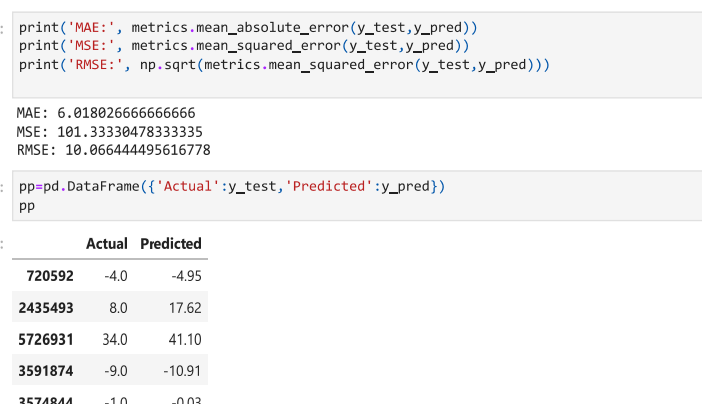
ypred = P redicted Arrival Delay

ymean = Average Arrival Delay

N = T otal Number of Delayed Flights

**Results:**

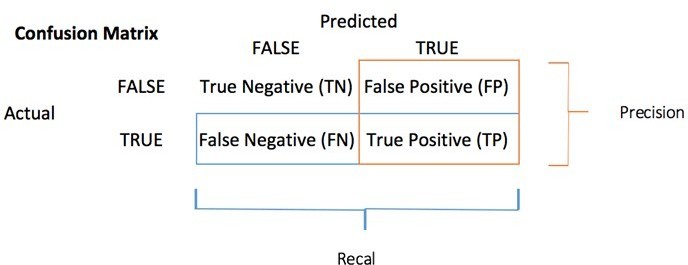
**Random forest:**





**Confusion Matrix**

A confusion matrix is drawn with each row of the matrix represents the instances in a predicted class while each column represents the instances in an actual class (or vice versa).



* TP: True Positive: Predicted values correctly predicted as actual positive
* FP: Predicted values incorrectly predicted an actual positive. i.e., Negative values predicted as positive
* FN: False Negative: Positive values predicted as negativ

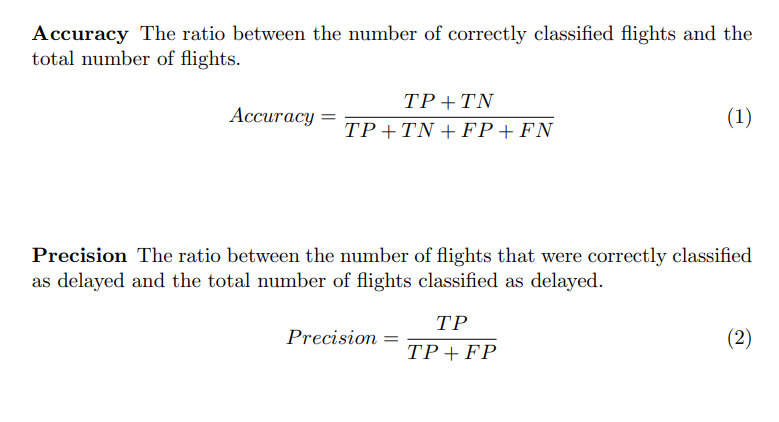
TN: True Negative: Predicted values correctly predicted as an actual negative

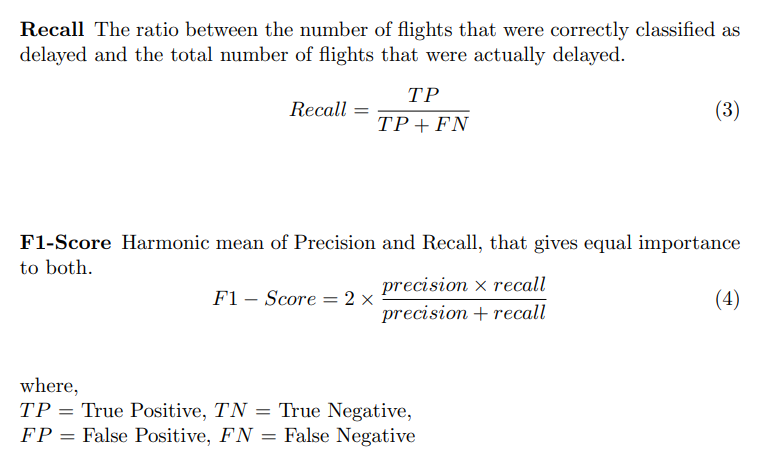


Confusion matrix

* + A **confusion matrix** is a tabular summary of the number of **correct and incorrect predictions** made by a classifier.
  + It is used to measure the performance of a classification model. It can be used to evaluate the performance of a classification model through the calculation of performance metrics like accuracy, precision, recall, and F1-score.
  + you can compute the **accuracy test** from the confusion matrix:



* In order to evaluate the performance of the aforementioned models, the following evaluation metrics are used. 



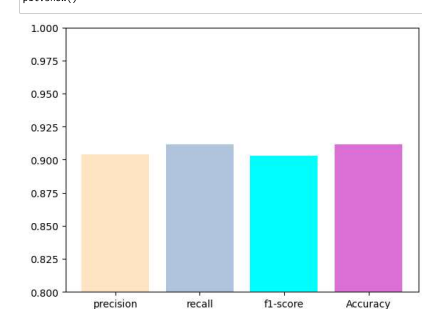
# 7 Experimental results

**OUTPUT RESULTS FOR THE ALGORITHMS**

**Random forest**

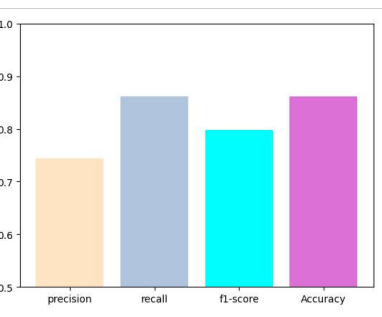


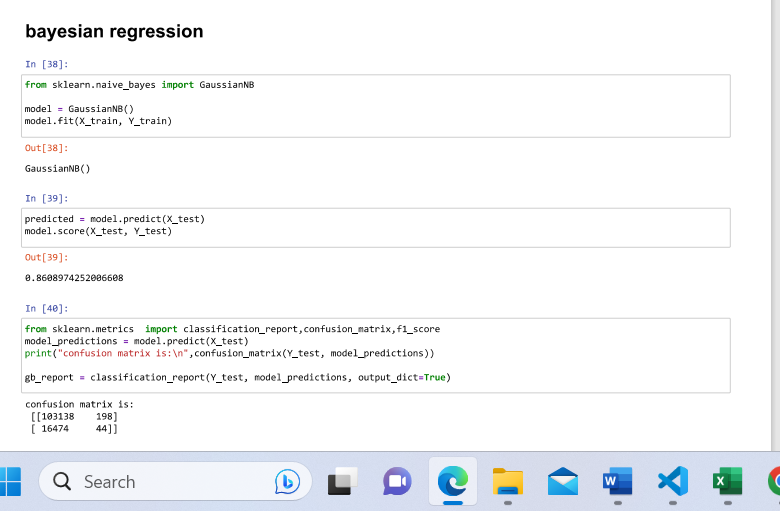
* Scikit learn includes a variety of classes for implementing machine learning models.
* Random forest is one of them,which fits the multiple decision tress to data.

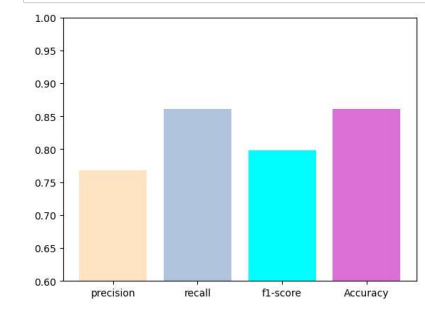




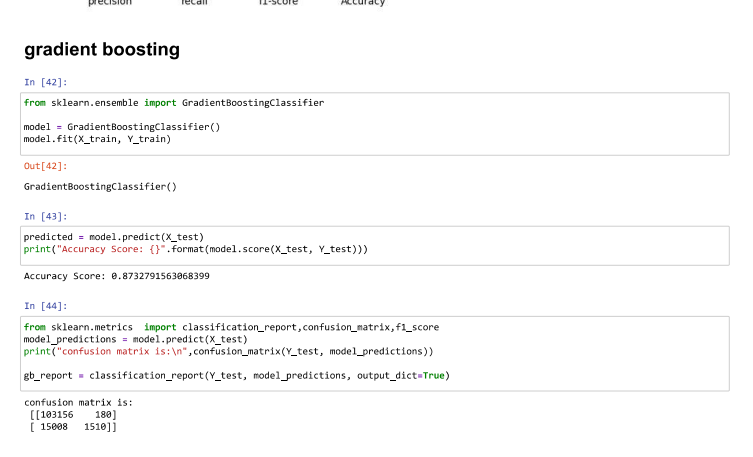
1. This type of statistical model (also known as *logit model*) is often used for classification and predictive analytics.
2. Logistic regression estimates the probability of an event occurring, such as voted or didn’t vote, based on a given dataset of independent variables. Since the outcome is a probability, the dependent variable is bounded between 0 and 1





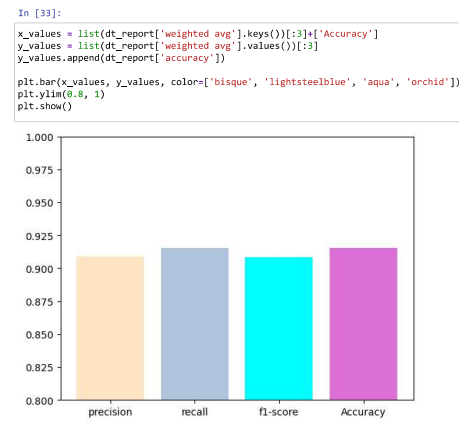


**Gradient boosting:**



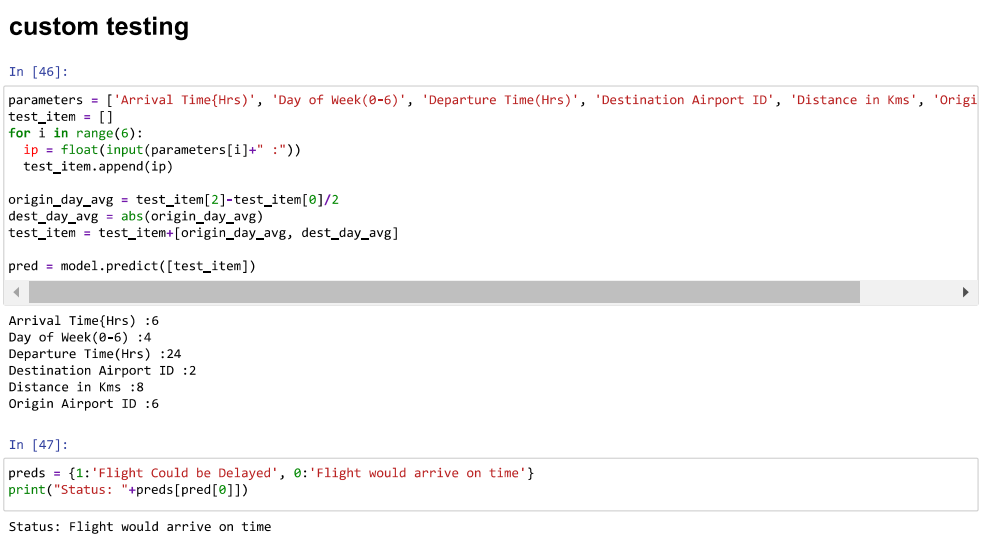


**Decision tree:**



The use of logistic regression which is a supervised learning method to predict delay in departure time of aircraft.Random forest is used for training and testing data 80:20 respectively to feed the data into random forest classifier.logistic regression gives an accuracy of 0.8612.Random forest gives an accuracy of 0.9114 ,gradient boosting gives an accuracy of 0.8732,decision tree gives an acuracy 0.9111 and Bayesian regression gives an accuracy of 0.8608.

## 7.1 Custom testing



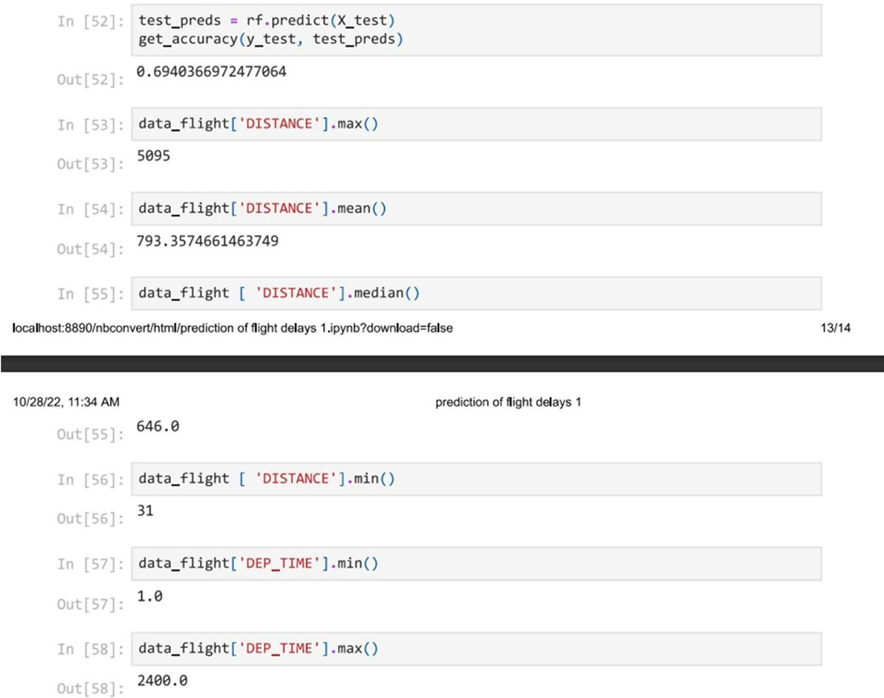
Here, the columns are passed as parameters to the variable parameters.

On testing the code ,the user should enter the values for the arrival time,day of week,departure time,destination airport,distance in kms and origin airport.

**7.2 Accuracy table**

|  |  |
| --- | --- |
| **Algorithms** | **Accuracy** |
| Decision Tree | 0.90 |
| Logistic regression | 0.86 |
| Bayesian regression | 0.86 |
| Gradient Boosting | 0.87 |
| Random Forest | 0.91 |

* + By using this we can predict the upcoming year flight dealy and we avoid cancellation of flights due to departure delay.
  + The mean, median, max distances are calculated here. The minimum departure time and max departure time is calculated to predict flight delay

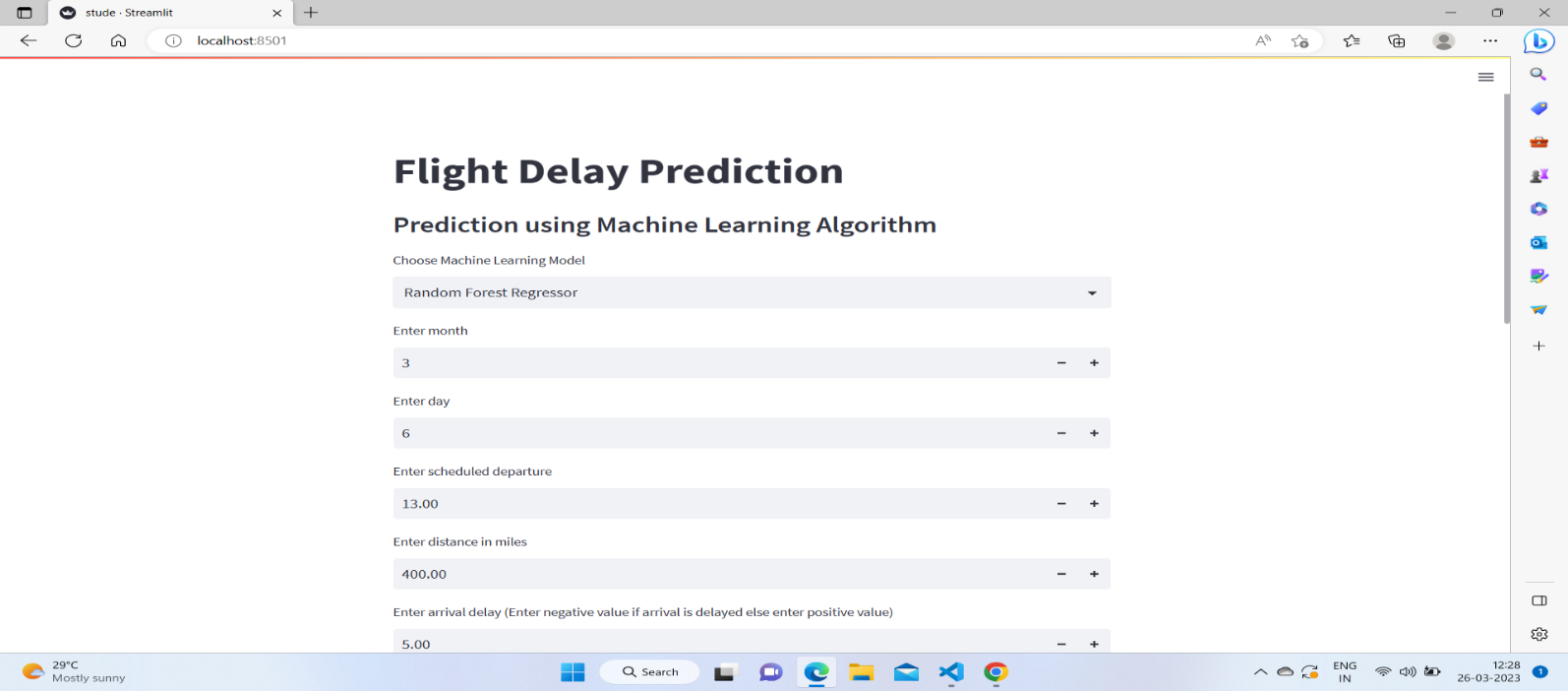


* In the above figure ,using ‘predict function’ the accuracy is calculated.Hence,the accuracy found was 0.6940366.
* The maximum distance calculated using max function is 5095
* The mean distance found was 793

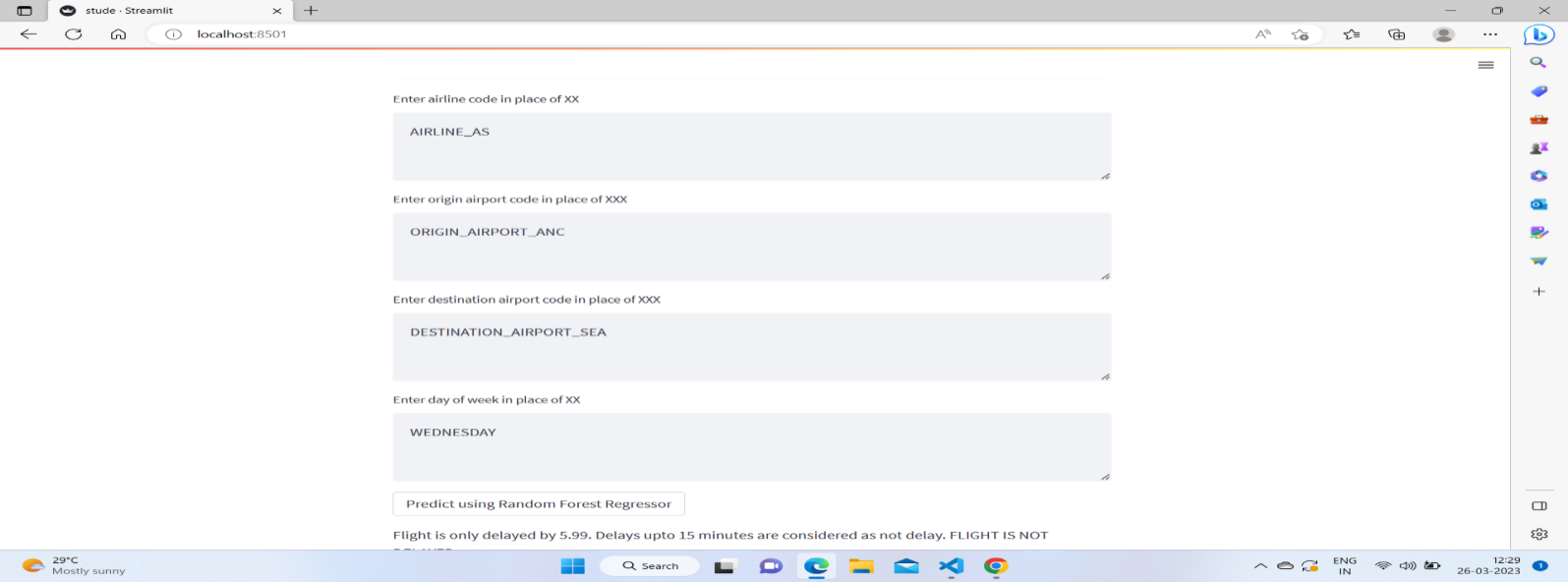
**8 Website**

**8.1 front-end:**

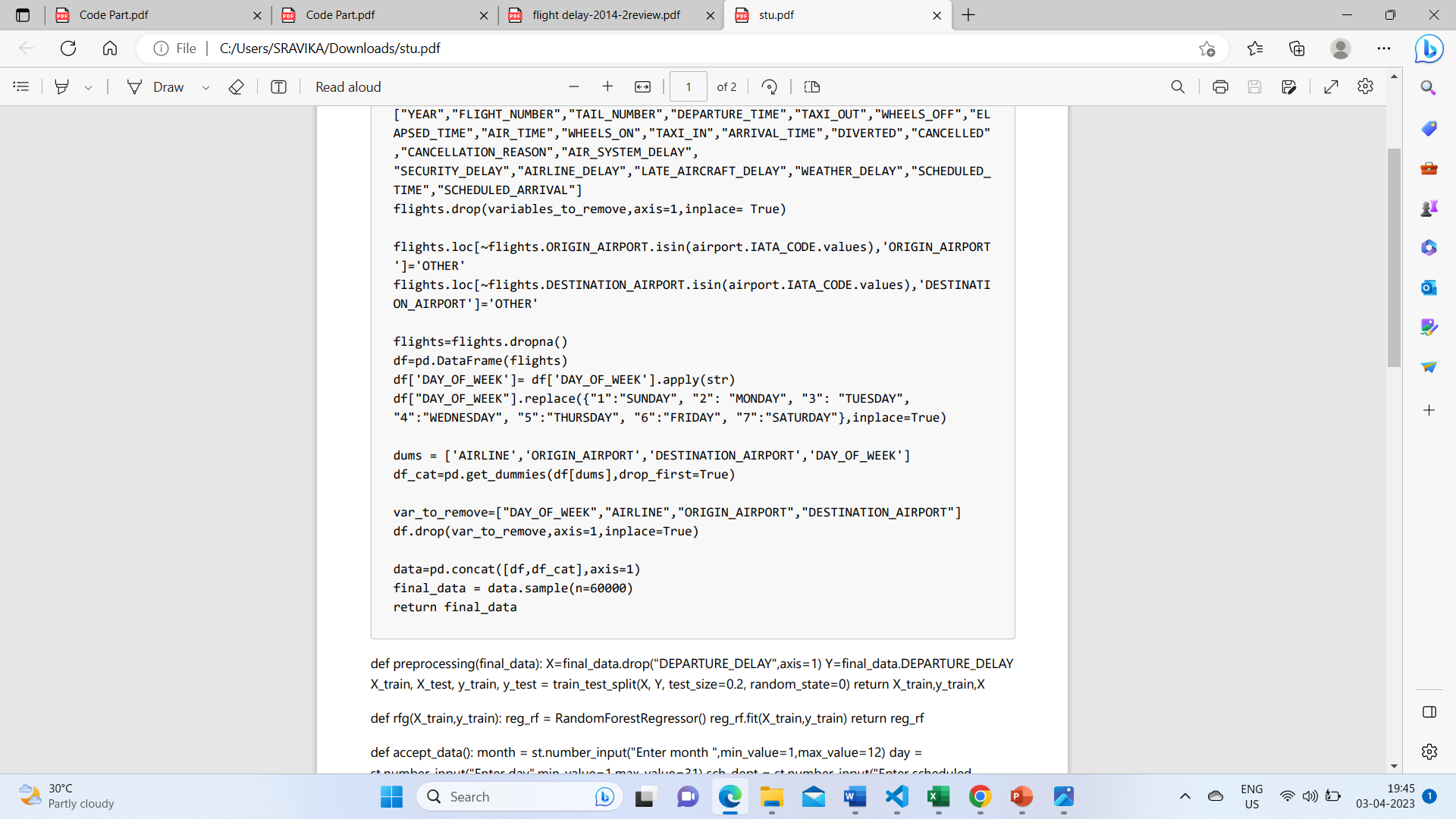
The website is developed for the user convenience using stream lit frame work.



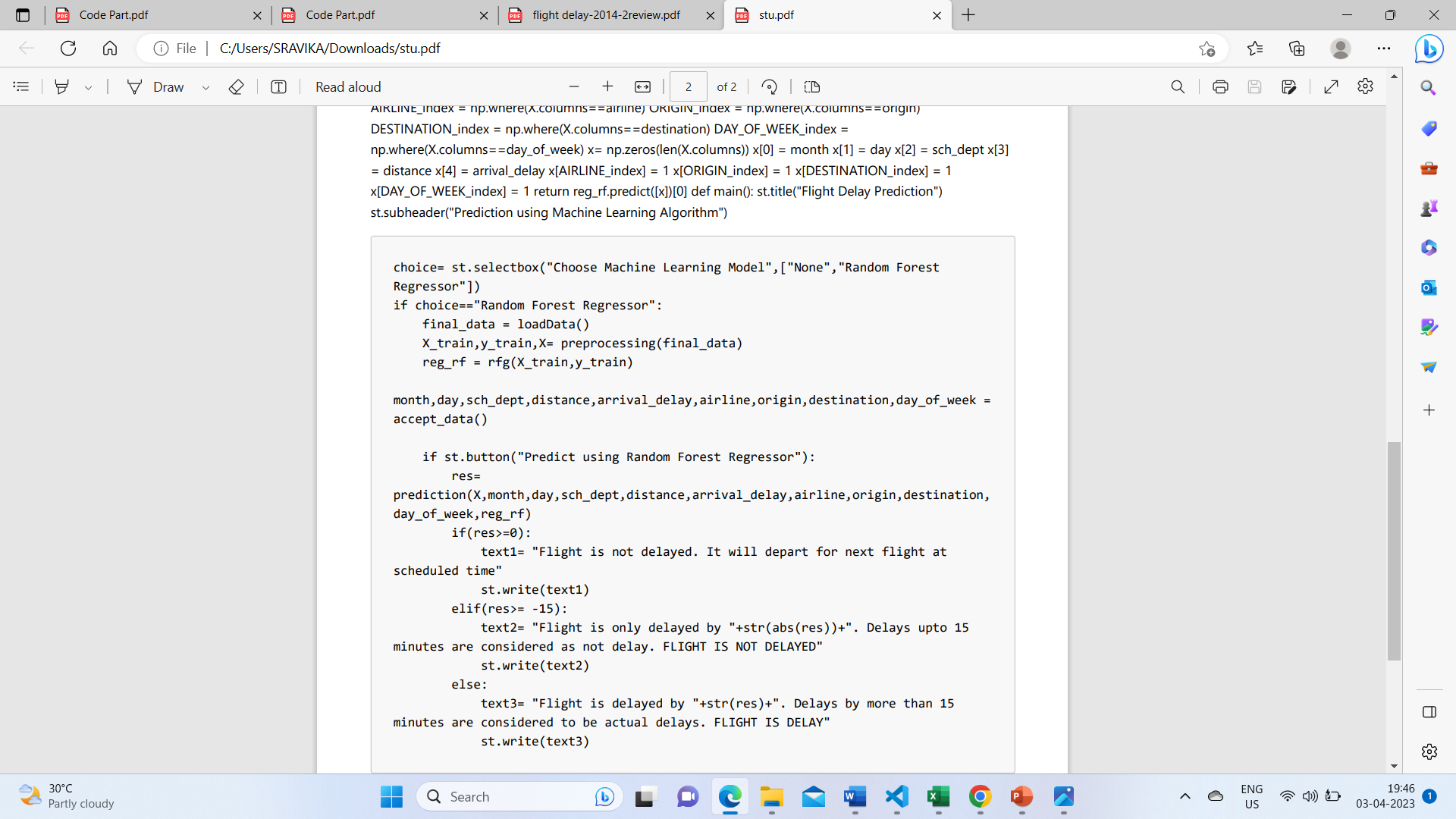
* This is the front end where we have taken stream lit object to show the title of flight delay prediction
* sub header as Prediction using machine learning algorithms
* here we also have select boxes for selecting the type of machine learning algorithm where user will have to choose random forest regress or as a machine learning algorithm.
* Here also then we can have user inputs like month date schedule departure, arrival delay,distance,airline core,origin and destination airports and day of week those are the flight details of the user after they enter the details then we click the predict button



**8.2 Backend**



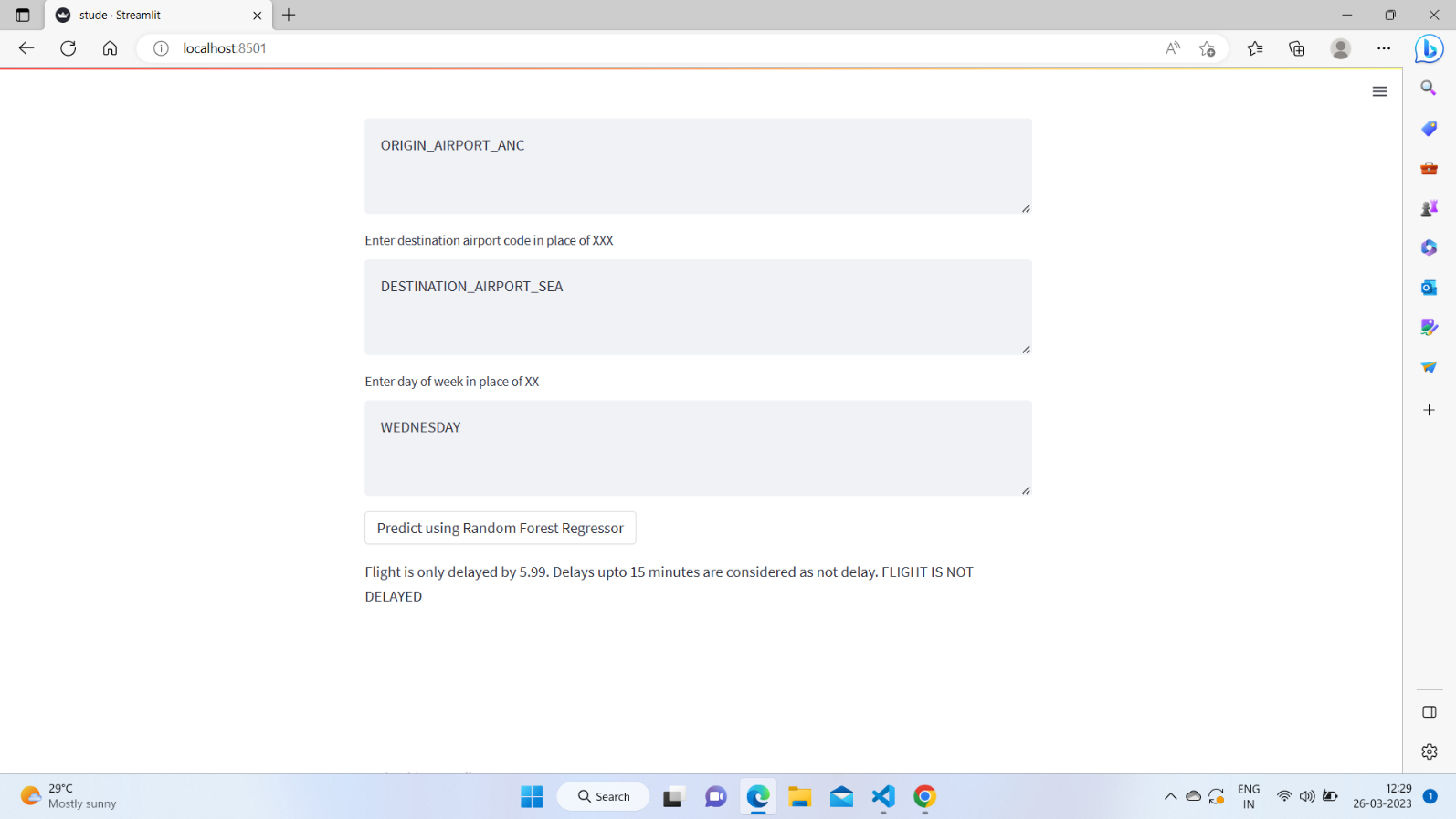
* if part gets exe,main function is called
* choice for machine learning model user made random forest ,will stored in choice variable
* if part gets executed.
* In load data we are called load data function
* we are import flight.csv,airports.csv.
* remove - we are storing irrelevant variables are removed
* drop-dropped col
* If origin airport column from fights data set is not matched with airports with then it is labeled as other
* drop col which have null values
* using pandas library we have created the d.f of the flights datasets
* we have converted int datatype of int to string for user convenience
* dummie variable for these column,that is stored in dums.



* dummie are stored df-cat.
* we can encounter column which are irrelevant,those are again stored in var,to remove and drop data df
* we have compared two datasets df&df- cat using contact fun of pandas and stored the output in data

preprocessing ,split x&y data sets

* since text size is 0.2 ,training & testing is done on 80% & 20% of data resp
* x-train passed to rfg function



* Here user gets the output where flight delays will be predicted.And then on clicking the button i get the output as the flight is only delayed by 5.99minutes since delays upto 15 minutes are considered as output flight is not delayed in this case.

# 9 Conclusion:

This project contributes the correlative application and analysis of distinct machine learning algorithms in the python software which gives an immediate mechanism for the user to use the machine learning algorithms.The project is implemented using google colab.Machine learning algorithms were applied progressively and successively to predict flight arrival & delay. We built three models out of this.Each evaluation metric considered the values of the models and compared them.The mean departure delay with respect to distance is 793.357, and the maximum departure time is 2400.0.The accuracy of random forest algorithm is 0.911.In the rest of the metrics, the value of the error of Random Forest Regressor although is not minimum but still gives a low value comparatively. In maximum metrics, we found out that Random Forest Regressor gives us the best value and thus should be the model selected.so we have developed using stream lit frame work.that website will predict the delay time in minutes.This prediction will be very helpful to user.

# 10 References

### Chakrabarty et al vol.75vol. 76, no.7, pp.4802–4837, 2020. K.S.Adewole,T.Han, W.Wu, H.Song vol. 24, no. 5, pp. 3475–3498, 2020.

https://towardsdatascience.com/a-comprehensive-guide-to-convolutional-neural- networks- the-eli5-way-3bd2b1164a53

### Juan José Rebollo and Hamsa Balakrishnan vol. 7, no. 1, pp.71–86, 2020

https://scikit-learn.org/stable/modules/tree.html#regression

### Sruti Oza, Somya Sharma vol. 10, no. 8, pp. 2143–2162, 2021.

https://machinelearningmastery.com/tutorial-first-neural-network-python-keras/

### Choi et al Retrieved 2 February 2017.

https://[www.kaggle.com/datasets/usdot/flight-delays](http://www.kaggle.com/datasets/usdot/flight-delays)

### Rebollo and Balakrishnan Retrieved 24 October 2020

[http://scikitlearn.org/stable/m odules/classes. html?source=post](http://scikitlearn.org/stable/modules/classes.html?source=post)

### Tu et al. Retrieved 2021-10-25.

[http://www.transtats.bts.gov.](http://www.transtats.bts.gov/)

### Hsiao and Hansen Retrived on 12 October 2020

https://[www.researchgate.net/publication/315382748\_A\_Review\_on\_Flight\_Delay\_Prediction](http://www.researchgate.net/publication/315382748_A_Review_on_Flight_Delay_Prediction)

### Junfeng et al. vol. 6, no. 2, pp.74–85, 2020

https://jespublication.com/upload/2022-V13I60147

### [9]Adrian et al. vol. 26, no. 4, pp. 3445–3398, 2020.

https://ietresearch.onlinelibrary.wiley.com/doi/full/10.1049/itr2.12183

### [10]McCarthy et al. Py t h o n 3 . 9 . 2 do c u m e nt a t i o n " . Ar c h i v e d f r o m th e or i gi n a l on 24

https://[www.sciencedirect.com/science/article/pii/S2046043022000119](http://www.sciencedirect.com/science/article/pii/S2046043022000119)