Flight delays prediction: a practical approach

**Farid Ganem**

**Sidharth Gupta**

**Jeet Patel**

**Milena Rocha**

**Lakshya Saxena**

# INTRODUCTION

Since the first successful powered flight in history made by the Wright brothers in 1903, airplanes have become a popular mode of transport especially covering distant geographic locations. In 2022, US carriers transported around 750 million domestic passengers in approximately 6.7 million flights. According to the Bureau of Transportation Statistics, from the total number of flights in 2022, 76.56% departed on time whereas 20.75% were delayed and 2.69% of them were cancelled [1]. A flight is considered delayed if it departed or arrived at the gate fifteen minutes or more than the scheduled time reported in the system.

Flight delays usually have negative impacts on the passengers, as they need to adjust to new schedules, the airlines, as they have to provide compensation for passengers in certain cases, and the airports as well as delays might cause bottlenecks in their operations. Usually, the most affected party in this equation is the passengers. They mostly rely on the schedules provided by airlines and have little or no information about the likelihood of a flight being delayed when they book it. Only regular passengers might have previous knowledge of the behaviour regarding delays of a certain flight with a certain carrier.

In this report, we will explore a dataset of flights originating and arriving in cities in the United States in 2004. The purpose of this analysis is to implement and choose the most adequate machine learning algorithm to predict whether or not a flight will be delayed according to the characteristics associated with it. The findings obtained in this report will be used as input to solve a business problem for a metasearch engine such as Kayak.

# DESCRIPTION OF THE BUSINESS PROBLEM

Kayak is a metasearch engine that provides travel agency services, such as booking flights, hotels, and rental cars. With the aim of providing outstanding service to their customers, Kayak wants to give information about the likelihood of a flight being delayed. With this information, customers can choose different alternatives (i.e. flying with a different carrier, flying from or to a different airport, etc.) to avoid delays.

Different machine learning methods are going to be applied to a dataset containing historical data of flights in the USA, to select the one with the highest accuracy. Once the most accurate algorithm is selected, the company will be able to develop a model for the new functionality. With this new feature, Kayak wants to provide added value to its customers and position itself as the top online travel services provider.

# PRELIMINARY ANALYSIS

The preliminary analysis will include a description of the fields and columns in the dataset, as well as summary statistics of the quantitative variables included in it. This analysis will give us a first glance at the characteristics of our data, which will be later useful when applying diverse machine learning algorithms.

*A. Description of the dataset*

The Flight Delays dataset contains 2,201 records of flights departing and arriving at United States locations in 2004. Each row corresponds to a flight and each column corresponds to a variable associated with the flight, such as carrier, departure time, destination, and origin airport. The predicted value is Flight Status which indicates if the flight was delayed or on time. The following table summarizes the characteristics of the variables included in the dataset.

Table 1 Description of the variables in the dataset

| **Column name** | **Description** |
| --- | --- |
| CRS\_DEP\_TIME | Scheduled departure time |
| CARRIER | Carrier (airline) code |
| DEP\_TIME | Actual departure time |
| DEST | Destination IATA airport code |
| DISTANCE | Distance between origin and destination flights |
| FL\_DATE | Flight date |
| FL\_NUM | Flight number |
| ORIGIN | Origin IATA airport code |
| Weather | Weather conditions |
| DAY\_WEEK | Day of the week 1-Monday – 7-Sunday |
| DAY\_OF\_MONTH | Day of the month 1-30 |
| TAIL\_NUM | Plane tail number |
| Flight Status | Weather the flight was delayed or not |

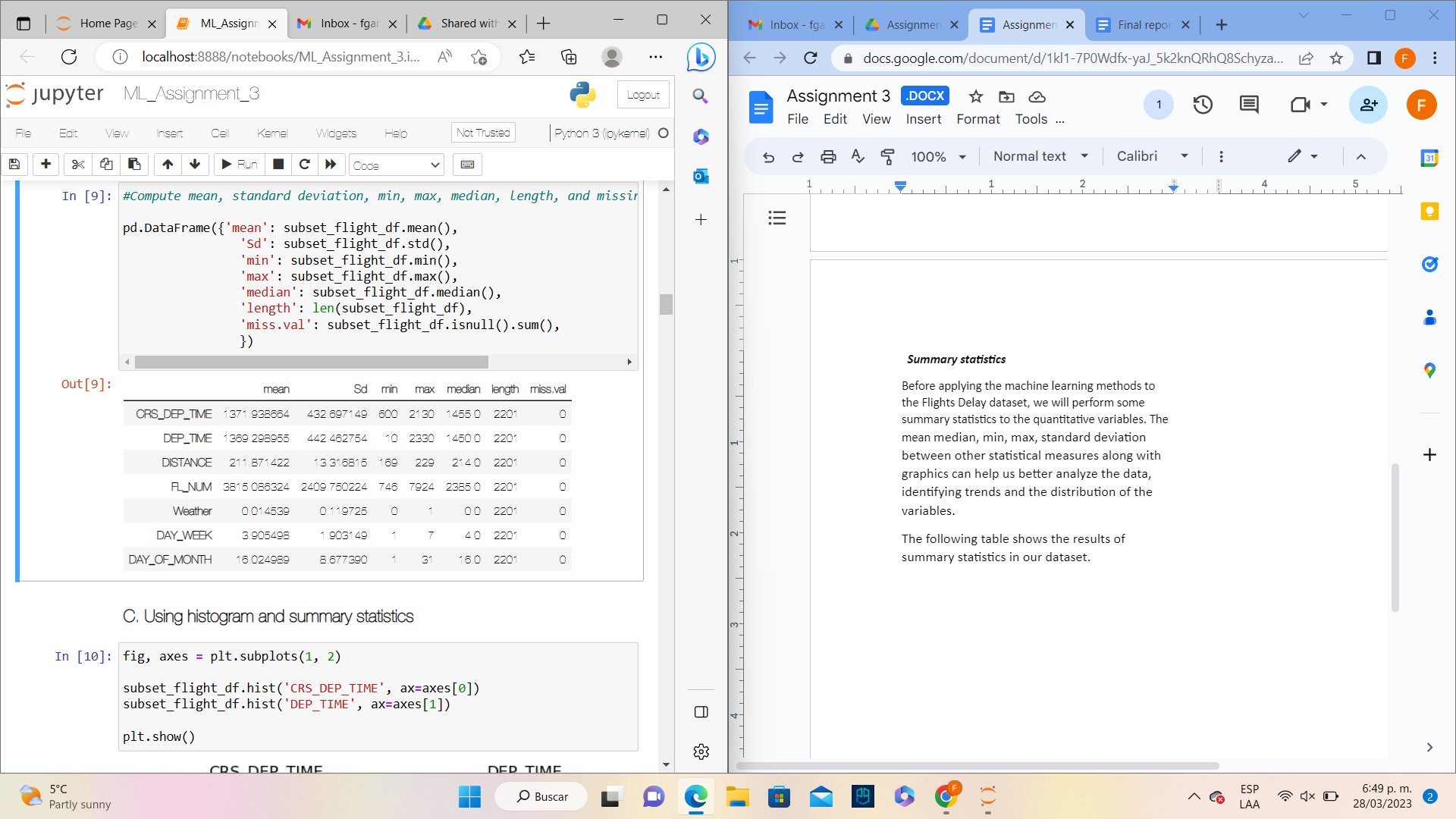
*B.* *Summary statistics*

First, the Data was imported into the libraries and was read in Python. After realizing which columns were in the table and their type, we found some unique values. After this, we created a subset data frame with only numerical values for statistical purposes.

Before applying the machine learning methods to the Flights Delay dataset, we will perform some summary statistics on the quantitative variables. The mean median, min, max, and standard deviation between other statistical measures along with graphics can help us better analyze the data, identifying trends and the distribution of the variables.

The following table shows the results of summary statistics in our dataset.

Figure 1 SUMMARY STATISTICS



For each of the values in Figure 1 we created all histograms to be able to have the results in a more visible way for data analytics purposes.

Figure 2 histograms for weather and day\_week

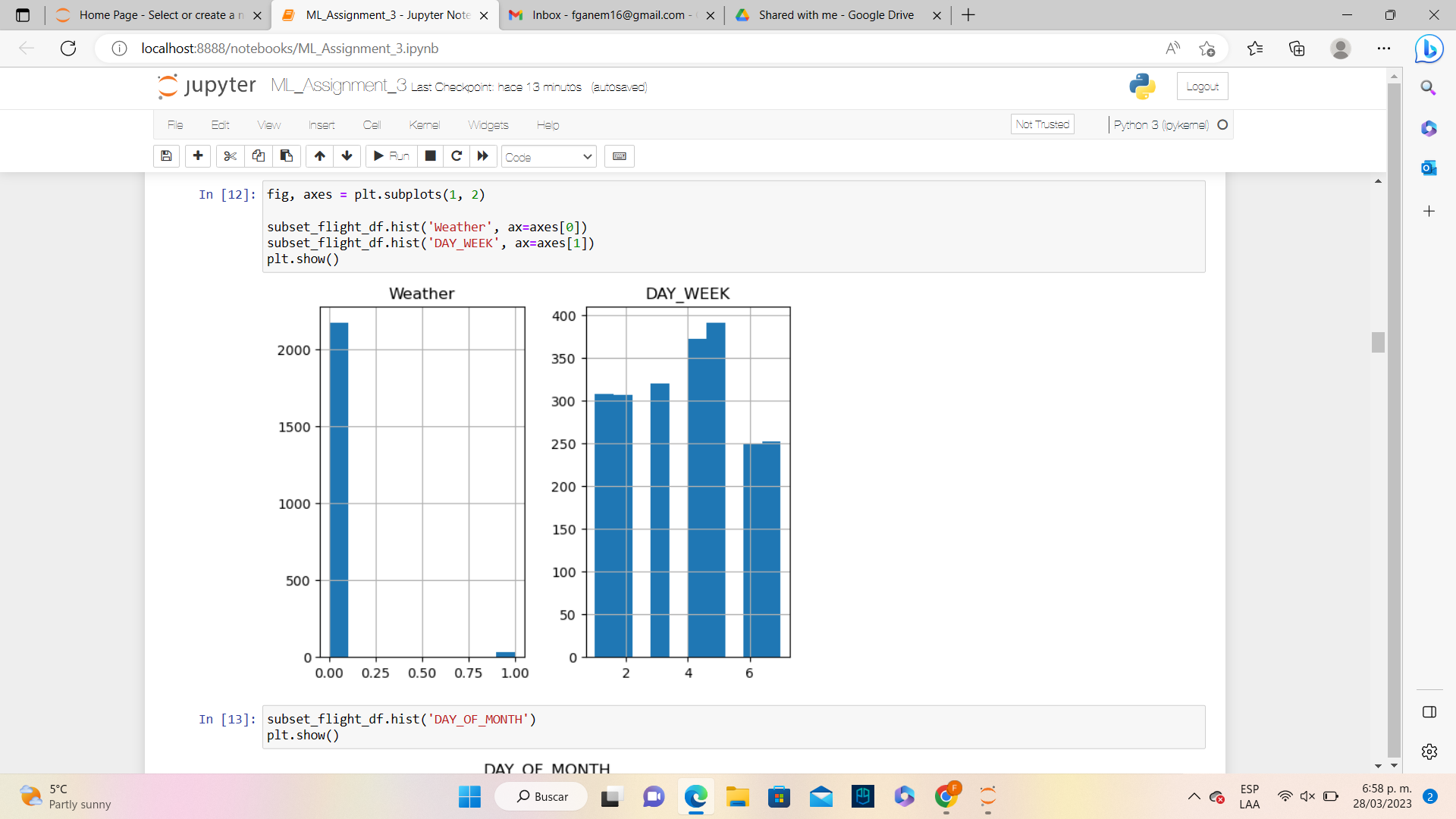


Figure 3 histograms for scheduled departure time and departure time

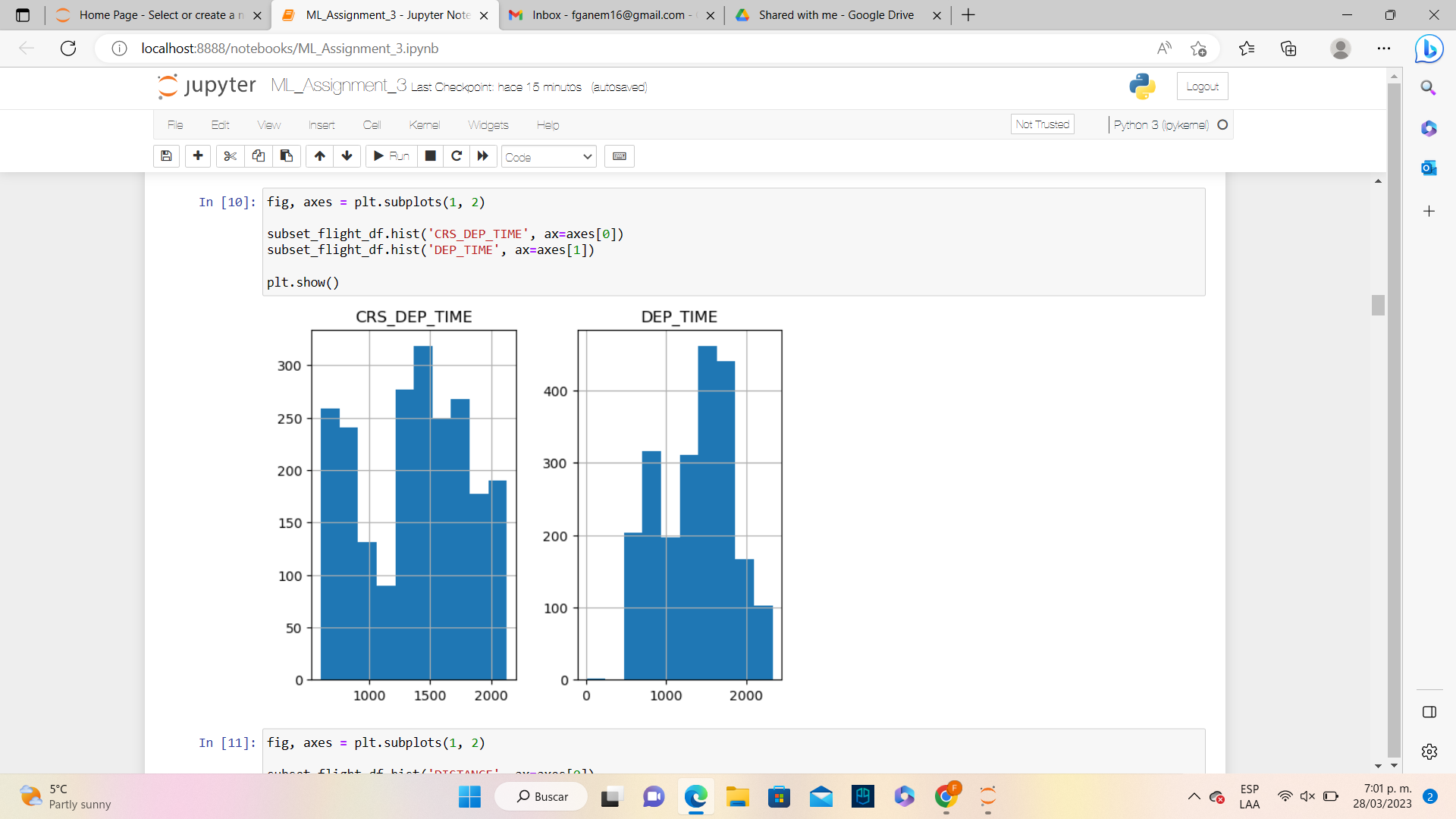
****

Figure 4 histograms for Distance and FL\_Num.

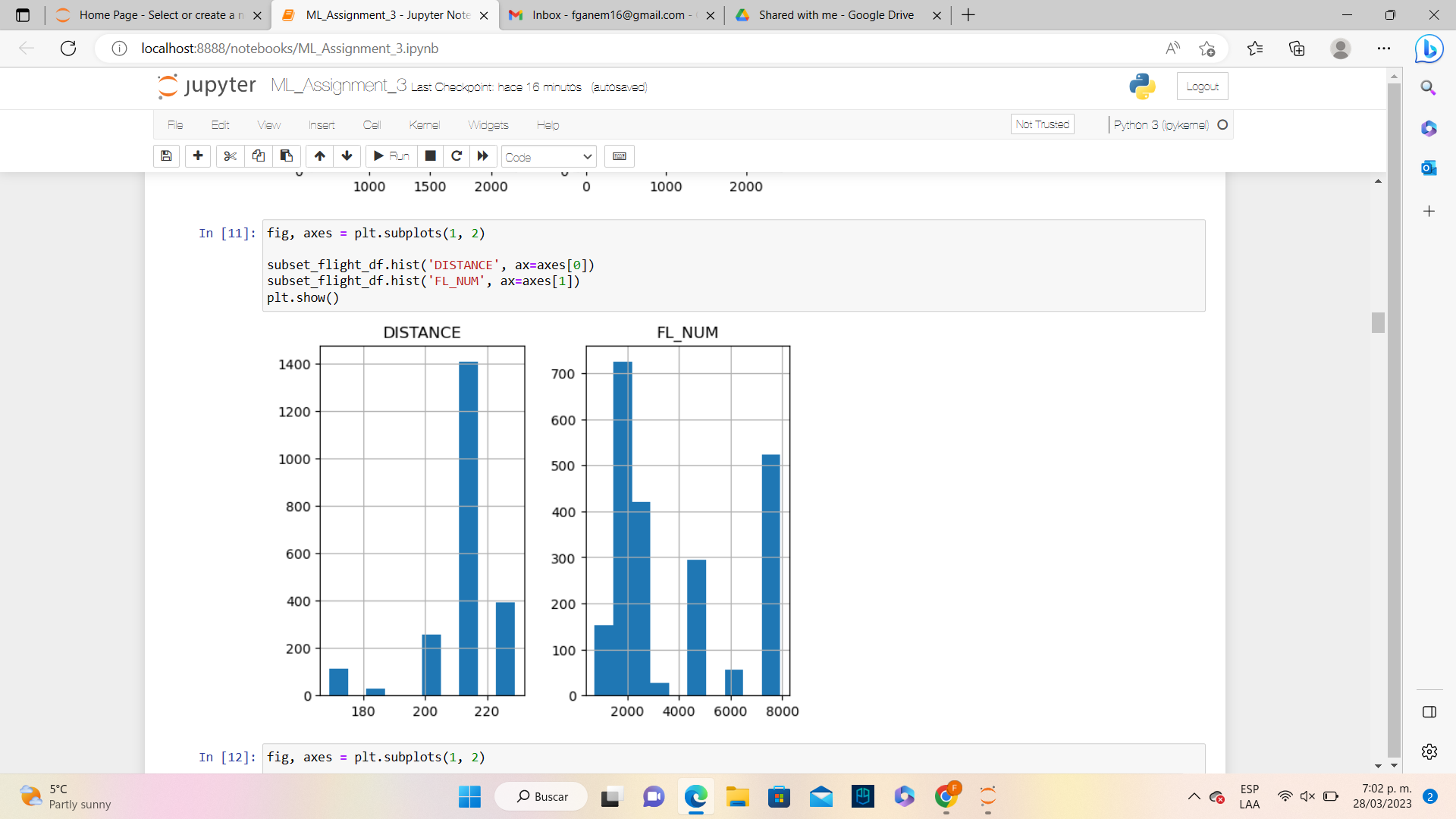


Figure 5 histogram for day\_of\_month

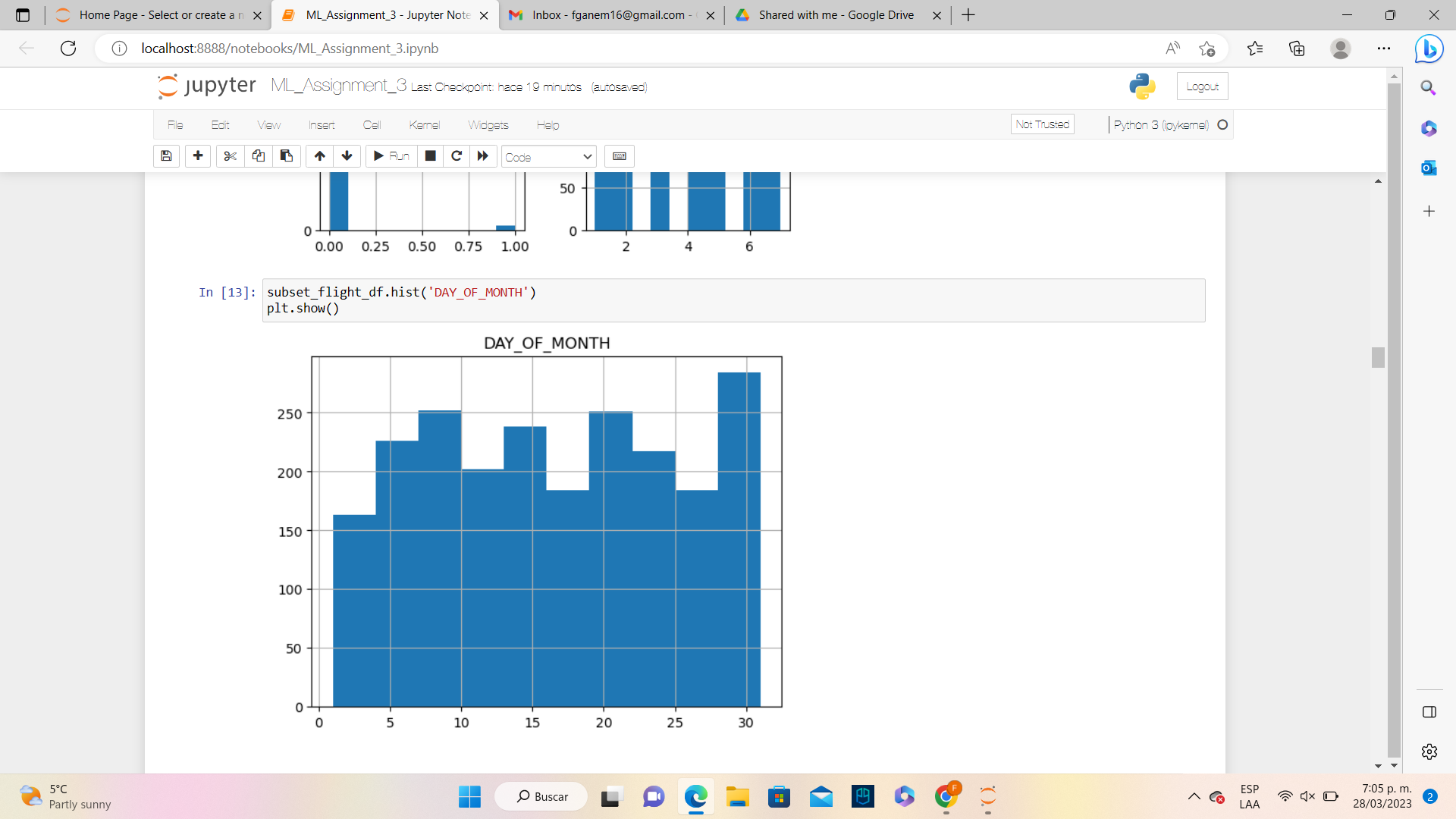
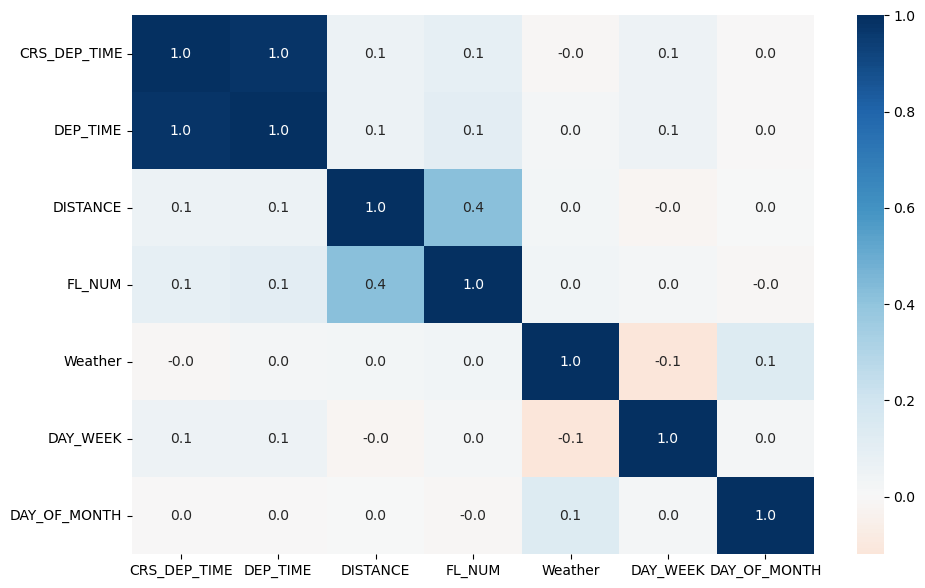
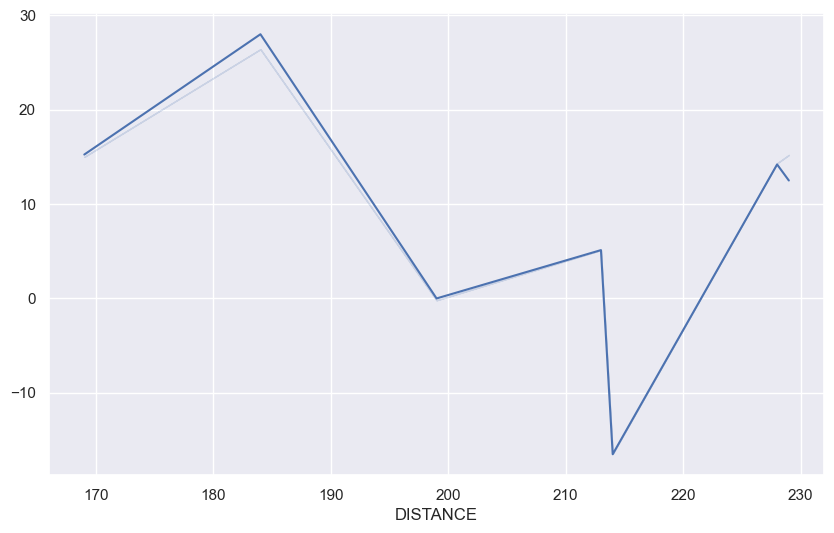
****

Figure 6 correlation between variables

****

After this, we use Principal component Analysis, remove the correlated values and make a sample of 10 flights. We split the data into training (60%) and testing (40%). Split it into training and do validation.

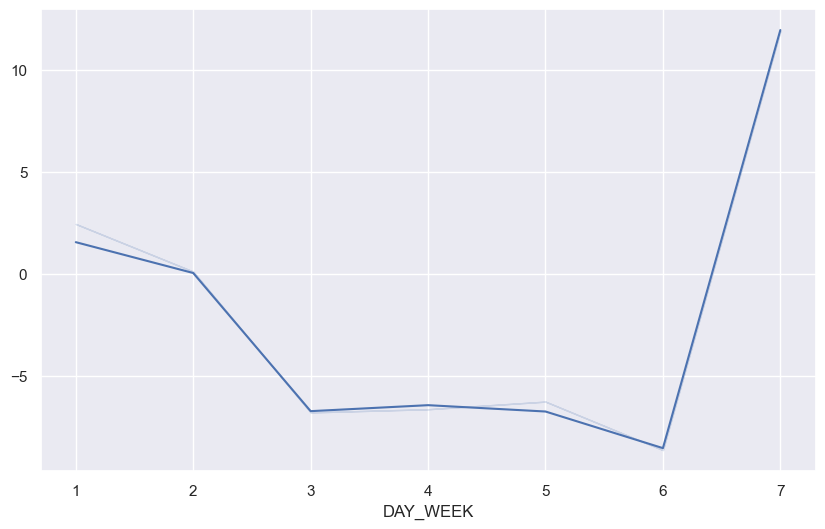
Figure 7 shows relationship between delay and distance



As we can see, figure 7 above shows the delay is higher for flights at smaller distances rather than long-distance flights.

Figure 7 does make sense since short-distance flights are cheaper flights and by that, there are more people on these. They have a full and difficult schedule and can always have a delay.

Figure 8 shows the number of delays on the WEEKDAYS.



As we can see in Figure 8, most delays happened on the day of the week, which is Sundays. On the other hand, on Saturdays, there are almost non-delays.

# MACHINE LEARNING ALGORITHMS

Once there is a clear picture of the characteristics of the dataset, machine learning algorithms can be applied to classify data and make predictions. In this report, supervised learning techniques will be applied to the flight's delay data.

In supervised learning, the goal is to infer the best mapping between an input and output dataset based on provided labelled examples [2]. The fields in the dataset are classified as input variables or predictors (X) and the outcome variable (Y). In these kinds of methods, the parameters are optimized to minimize the error between the computed output and the actual output [3].

There are two types of supervised learning: classification and regression. As its name suggests, in classification the algorithm is trained to classify the input into discrete variables such as “Yes” or “No”. Regression refers to the prediction of a continuous variable, for example, the price of a car based on its characteristics. To provide a solution to the business problem presented for Kayak, two supervised learning methods: K-nearest neighbours and Naïve Bayes Classifier will be used and compared.

To implement these methods, we first need to remove the variables that are correlated in the Flight delays dataset. As explored in the summary statistics, the variables CRS\_DEP\_TIME and DEP\_TIME are highly correlated, therefore one of them has to be dropped. The next step consists in splitting the data into training and testing and defining the predictors and outcome variables in the system. The following chart summarizes the classification of the variables.

Table 2 CLASSIFICATION OF THE VARIABLES

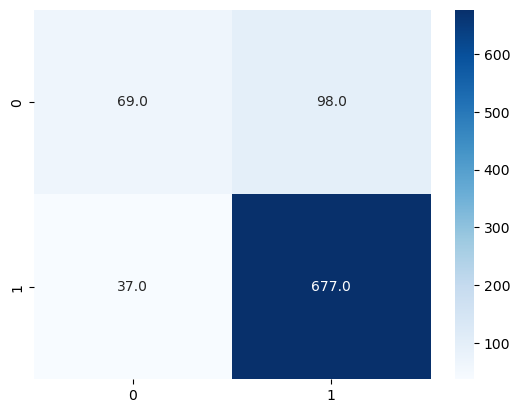
| **Variable** | **Type of variable** |
| --- | --- |
| Carrier | Predictor |
| Departure time | Predictor |
| Destination airport | Predictor |
| Distance | Predictor |
| Flight date | Predictor |
| Flight number | Predictor |
| Origin airport | Predictor |
| Weather | Predictor |
| Day of the week | Predictor |
| Day of the month | Predictor |
| Plane tail number | Predictor |
| Flight status | Predictor |

1. *The K-nearest neighbours' classifier*

The K-nearest neighbours classifier aims to identify records in the dataset that are similar to a new record that needs to be classified. This method allows classifying a data point according to the label of the other close data points in the training set.

We implemented the K-nearest neighbours’ method in the Flight delays dataset, using a fixed value of 3 for K for training. Once the model is trained, we made predictions on the test dataset and obtained the results shown in the following confusion matrix.

Figure 9 confusion matrix for k-nearest neighbours



Based on this confusion matrix, we can find the error of the model. This value is calculated by adding up the correctly classified records in each category and dividing them by the total number of cases as follows:

(i)

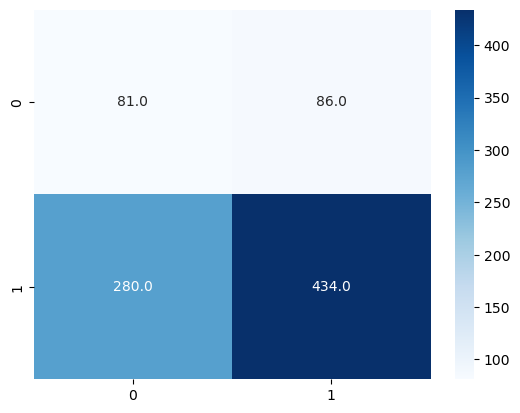
In this case, the accuracy of the K-nearest neighbours method for the flight delays dataset is 84.67%. Subsequently, the error for this method is 15.33%.

1. *The Naïve Bayes Classifier*

The Naive Bayes is a classifier method based on the Bayes theorem in which we find a probability of an event happening given that another has occurred, always assuming that predictors and features are independent [4].

To run the Naive Bayes method in Python, we first defined a 0.01 alpha value, then we predicted the probabilities of each record for each class and finally we predicted the class membership in our dataset. The results of the Naive Bayes method are shown in the confusion matrix shown below.

Figure 10 confusion matrix for naïve bayes



Based on this confusion matrix, we find an accuracy of 58.45% in the Naive Bayes method for the flight delays dataset. The error value of this method is therefore 41.55%.

1. *Comparison*

The effectiveness of the machine learning methods explored before is measured using different indicators as follows:

* Accuracy: the number of correctly classified records divided by the total number of records.
* Precision: the number of correctly predicted records divided by the total predictions.
* Recall: the number of true predicted positive records divided by the number of actual positive values
* F1-Score: corresponds to the weighted average of precision and recall.

The following table summarizes the performance indicators of each model.

Table 3 performance indicators for the models

| **Parameter** | **Naive Bayes** | **KNN** |
| --- | --- | --- |
| Accuracy | 0.584563 | 0.846765 |
| Precision | 0.224377 | 0.650943 |
| Recall | 0.48503 | 0.413174 |
| F1-Score | 0.306818 | 0.505495 |

# Additional Algorithms

# *C. Classification and Regression Trees*

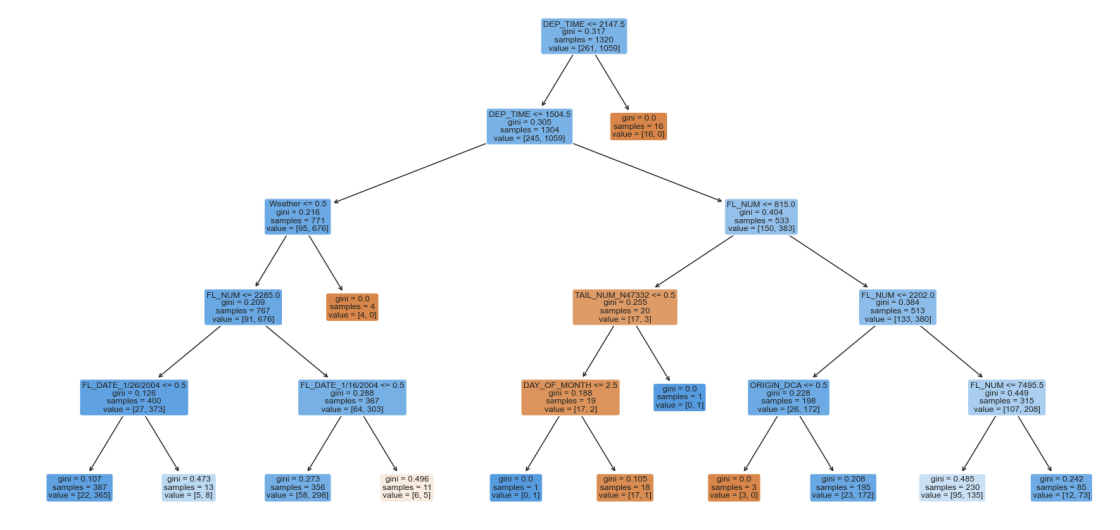
CART (Classification and Regression Trees) can be used to solve both classification and regression issues, as the name implies. The target variable is where the difference lies. [5]:

We attempt to predict a class label by categorization. In other words, classification is employed for issues where the output (target variable), such as whether or not it will rain tomorrow, can only take one of a limited number of values. [5]

Regression is then used to forecast a numerical label. This implies that the values of your output, like the price of a house, are infinitely variable. [5]

To create the decision tree classifier and fit the data, we predict the test data using the decision tree to calculate the accuracy and plot the decision tree .

Figure 11 confusion matrix for k-nearest neighbours



Based on this decision tree, we find an accuracy of 82.29% in the CART for the flight delays dataset. The error value of this method is therefore 17.71%.

*D. Logistic Regression*

A predictive analytic algorithm based on the idea of probability, logistic regression is a machine learning approach that is used for categorization problems [6].

Assigning observations to a discrete set of classes is done using the classification process known as logistic regression. The hypothesis of logistic regression tends to limit the cost function between 0 and 1 [6].

In order to map predicted values to probabilities, we created a test dataset and predicted the accuracy by comparing the actual test value with the predicted value.

Based on the prediction, we find an accuracy of 81.95% in the regression model for the flight delays dataset. The error value of this method is therefore 18.05%.

# Conclusions

As we worked on two different methods and performed them in Python to realize which one was more accurate, precise and other performance indicators for the models.

We realized that by running Naive Bayes and K-Nearest Neighbor (KNN) algorithms on the dataset of Delayed Flights, the accuracy is almost 26% higher in KNN than in Naive Bayes. In other words, the accuracy of the KNN is 84.68% while the accuracy for the Naive Bayes is 58.46%.

Also, the precision is almost three times higher in KNN (65.09 % compared to 22.44%). Even though the Recall is 7% higher on the Naive Bayes, the F1-Score is again 20% higher on the KNN procedure. The KNN method is better in this case than Naive Bayes.

Based on this analysis, we highly recommend Kayak to use KNN algorithm method for its accuracy so they can create features in their web page and add value to their customers by giving them accurate and precise options regarding flights that do not have a delay.

# References

[1] “On-Time Performance - Reporting Operating Carrier Flight Delays at a Glance,” *Bureau of Transportation Statistics. U.S. Department of Transportation*. [Online]. Available: https://www.transtats.bts.gov/homedrillchart.asp. [Accessed: 29-Mar-2023].

[2] M. Bironneau and T. Coleman, *Machine learning with go quick start guide: Hands-on techniques for building supervised and unsupervised machine learning workflows.* Birmingham: Packt Publishing, 2019.

[3] T. Jo, Machine learning foundations: Supervised, unsupervised, and advanced learning. Germany: Springer International Publishing, 2021.

[4] R. Gandhi, “Naive Bayes classifier,” Towards Data Science, 17-May-2018. [Online]. Available: https://towardsdatascience.com/naive-bayes-classifier-81d512f50a7c. [Accessed: 29-Mar-2023].

[5]Dobilas, S. (n.d.). *CART: Classification and Regression Trees for Clean but Powerful Models*. Medium . Retrieved from https://towardsdatascience.com/cart-classification-and-regression-trees-for-clean-but-powerful-models-cc89e60b7a85

[6]Pant, A. (n.d.). *Introduction to Logistic Regression*. Medium . Retrieved from https://towardsdatascience.com/introduction-to-logistic-regression-66248243c148#:~:text=Logistic%20 regression%20is%20a%20 classification,Fraud%2C%20 Tumor%20 Malignant%20 or%20 Benign.