

Sara Aljamal

Artificial Intelligence assignment part 2

Contents

[**1** **Part II** 3](#_Toc138196838)

[1.1 Introduction 3](#_Toc138196839)

[1.2 Deployment 3](#_Toc138196840)

[1.2.1 Materials: 3](#_Toc138196841)

[1.2.2 Methods: 7](#_Toc138196842)

[1.2.3 Results and Discussion: 10](#_Toc138196843)

[1.3 References 11](#_Toc138196844)

# **Part II**

## Introduction

Flight delays have negative and huge impact on both airlines and passengers, because a delay results in financial loss, passengers’ and crew schedules, and bad reputation for the airline. To reduce the negative impacts of flight delays, we can predict them accurately which will improve the traveller experience, minimize the inconvenience, in addition to an efficient decision making for airlines.

This project seeks for accurate prediction for flight delays by using some models such as a ML model and an ANN model, by utilizing historical data for flights, the model will analyse, learn, and predict delays according to some features and relationships. This project will provide airlines with valuable information and enable them to make proactive decision, avoid delays, and take actions to mitigate disruptions. This project will also affect passengers positively by ensuring them accurate times for their flight with no delays and providing them with great travel experience without delays or disruptions.

Our contribution in this project will be shown by using two AI models to predict flight delays which will show us which one is better in the case of prediction delays a ML model or an ANN model according to the results. The prediction will enhance passengers experience with airlines which will benefit airlines that will take advantage of such project. This project will increase the research and work done in predicting flight delays section, also sharing the results with other researchers and data scientists would aid them in similar project and lead them to better results, I believe that sharing results and experiences in this topic will enhance the results of every new project, leading us to very accurate prediction and results. The project will also prove that integrating AI in all industries would improve the overall organizational process.

## Deployment

### Materials:

The dataset contains information about flights, such as the day, destination, and duration. The source of the dataset that this project will work on is Kaggle, you can find the link below in the references part. The Authors and publishers are Albert Bifet, Elena Ikonomovska, they brought the data from the Data Expo competition. The number of projects related to the same dataset uploaded to Kaggle is 30.

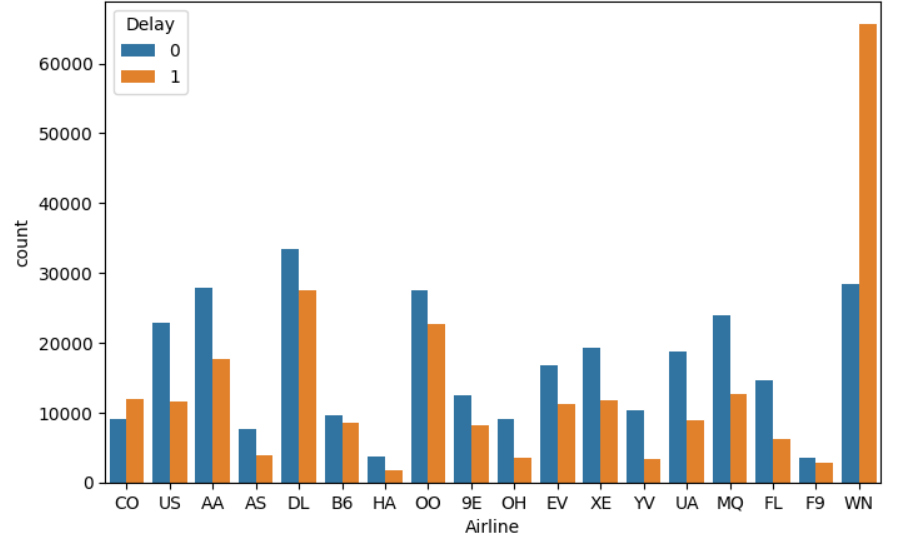
* The dataset consists of 539383 rows, where every row represents a flight, and 9 columns including the target variable:

1. ID: Which represents the serial number. (Numeric)
2. Airline: The name of the airline the flight belongs to. (Categorical)
3. Flight: The type of the aircraft of the flight. (Numeric)
4. Airport From: The airport which the flight departure from. (Categorical)
5. Airport To: The destination airport which the flight will land in. (Categorical)
6. Day Of Week: The flights day of the week. (Numeric)
7. Time: The departure time in minutes, starting from midnight (range from 10- 1439), 1440 is the number of minutes in a day. (Numeric)
8. Length: The flight duration in minutes. (Numeric)
9. Delay: Which is our target variable, represents whether the flight was delayed on not. (Numeric)

**Data preprocessing and EDA:**

The dataset doesn’t contain a lot of features, that’s why it doesn’t need much preprocessing.

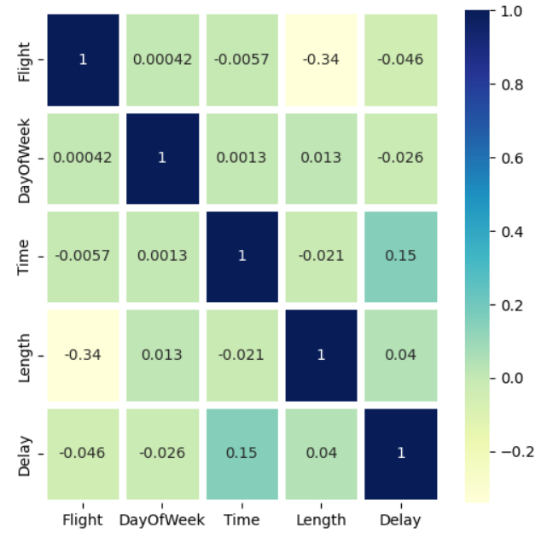
1. The first thing I did is dropping the ID column, because it’s unnecessary and it won’t benefit us in anything.
2. Secondly, I checked the outliers in the dataset, but the thing is we can’t consider having outliers in such dataset, because outliers didn’t make any sense in any column we have in the dataset, lets analyse the columns have, “ID” column was dropped, the “airline”, “airport from”, and “airport to” columns are categorical variables, moving to the “flight”, “day of the week”, and “delay” which are considered discrete numeric columns, and it doesn’t make any sense to say that day of the week is an outlier or a type of aircraft. The last two columns left are “time” and “length”, which can be considered continuous numeric variables because they represent time, but also these columns didn’t contain outliers, I tried taking the maximum value in “length” column which was 655 minutes (the longest flight in the dataset), after that I selected all flights (rows) with length 655, they all departure from the same airport which is EWR (Newark Liberty International Airport – New Jersey) to the destination airport which is HNL (Honolulu International Airport -Hawaii), 655 minutes can be considered 11 hours which is a true value for the flight length. We can say that this dataset doesn’t contain outliers, because these values are real data that describes a flight.
3. EDA or data visualization: For better understanding for the data, I’m dealing with, I visualized some features to find relations and gain insights.
4. In the airline column, we had 18 different values which means 18 airlines, I made a count plot for these airlines to show how many flights belong to each airline, as well as the airlines with the most and least flights in the dataset. In addition, each airline had 2 bar plots, the first one represents the number of normal flights and the second one represents the number of delayed flights. From the figure below, we can recognize a lot of things, for example, the airline with most delayed flights was WN airline.



1. The second plot showed the number of normal and delayed flight over the week. We can reach the conclusion that Thursday is the day with the biggest number of flights, but Wednesday is the day with more delayed flights.

A picture containing text, screenshot, diagram, font

Description automatically generated

1. An important thing to know and visualize is the relation between variables in the dataset, I used a heat map to visualize relations between numeric columns, the results weren’t very good, in fact, we can see that the relations between variables is weak, this can be an indication for bad results in the prediction step.
2. To make sure feature are on the same scale, I used min max feature scaling to scale the values for both “time” and “length” columns. But before that, I transformed both columns’ values from minutes to hours, because it’s much easier to understand, the “time” column which tells us which minute of the day the flight departure, is more readable and understandable when it’s represented by hours instead of minutes, for example, it’s easier to say the flight departure the 23rd hour of the day than saying in the 1380 minute. In addition, we must deal with data as it’s represented in real world, in our daily life, if we wanted to check the duration for our flight, we will see 2h 30 minutes and not 180 minutes.
3. An important thing that affects the predictive model’s performance sometimes is the balance of the data in the target column which is “delay” column in our case, so I checked it and it seemed that class 0 appears more in our dataset than class 1, so I lowered the number of values in class 0 to equal the values in class 1 using underdamping technique.

A blue and orange bars

Description automatically generated with low confidenceA picture containing screenshot, text, rectangle, diagram

Description automatically generated

With under sampling

Without under sampling

1. Lastly, I encoded categorical variables to numeric variables using label encoder.

### Methods:

* **Description for the models used:**

In this project, I used two models, one of them was a machine ML model which was decision tree, and the second one belongs to deep learning models which was an Artificial Neural Networks model.

Decision trees classify data according to a tree, where the root nodes represent a feature, or a condition and the leaf nodes represent decision and data classification. I used this model because it’s easy to explain how the data was classified, and it can handle non-linear relationships.

The ANN model tries to simulate the human brain neural networks. In neural networks we have 3 main layers, in input layer, hidden layer, and an output layer, and we van add more hidden layers according to our data. I used ANN because it can learn and identify complex patterns, and it’s a model I always wanted to try in predictive models.

* **Models pipeline:**

A diagram of a flowchart

Description automatically generated with low confidenceDecision tree pipeline:

A screenshot of a computer

Description automatically generated with medium confidence ANN pipeline:

* **Model implementation:**

Decision tree implementation:

* Importing libraries.
* Importing dataset.
* Dropping “ID” column.
* Checking outliers.
* Visualization.
* Feature scaling using normalization.
* Balancing data using under sampling.
* Encoding categorical variables using label encoder.
* Splitting data to X, y.
* Creating decision tree classifier.
* Feature selection using forward sequential feature selection.
* Defining variables to save the accuracy measure values to take their mean, in addition to a list for accuracy values to draw a box plot.
* Splitting the dataset to training and testing, training the model, making predictions, calculating the accuracy measure, and appending the accuracy value in the accuracy list defined above. (This process is repeated 10 times)
* Drawing a box plot for the accuracy values in the 10 iterations.

ANN technical implementation:

* Because this model was implemented after the decision tree, then the data here was already pre-processed, including all details mentioned above. In addition to splitting the data to X and y.
* Creating a sequential classifier.
* Splitting the dataset to training and testing
* Adding the input layer with appropriate parameters such as 12 neurons, the activation function type, and the input shape.
* Adding a hidden layer with 8 neurons, and relu activation function.
* Adding the output layer with 1 neuron and sigmoid activation function as parameters.
* Compiling the model.
* Training the model on our training data.
* Making predictions using the model on our testing data.
* Evaluating the model using appropriate accuracy measures such as accuracy, recall, and precision.
* **How our models can work with other models in the organization:**

A model like this one that van predict if the flight will be delayed or not, can improve the organization decision, ensure proactive activities, and increase efficiency. This model will be the only one in the organization that tries to improve and benefit the organization, so an aviation company or organization must apply and adopt other models to predict other things such as the number of passengers in each flight and so on. Integrating the model we made with and other models in the organization will increase its efficiency, make it a technological advanced company, and if it uses and integrates a lot of AI models it may revolutionize the whole aviation field and be the first company to accomplish this.

* **Performance metrics used in the project:**

Because this project was solving a classification problem, then we have to use the classification accuracy measures which are:

1. Accuracy: This metric calculates the percentage of correct predictions the models made. For example, if the accuracy was 50%, then we can say that half of the predictions were correct.

2. Recall: This measure evaluates the percentage of how many delays the model was able to predict correctly.

3- Precision: This measure evaluates how correctly the positive prediction were, to explain more using our data, it tells how many delays the model predicted that are actual delays

4- F1 score: It’s a measure that combines both precision and recall through a mathematical equation, so in this measure we consider both precision and recall together.

* **Based on the evaluation measures, how we improved the model:**

I tried running the models multiple times, in every time I changes something or added something, let’s take the decision tree model as an example, first of all I tried it without normalization, then I improved the results by normalizing the “time” and “length” columns, after that I realized I had to balance the data to get better results and enhance the model’s performance, so I balanced the data using under sampling technique. After that I tried to change the parameters such as the depth of the tree. Lastly, I added feature selection technique to improve the results.

### Results and Discussion:

* **Results analysis:**

With all the technique I mention in the previous question to improve the models performance but the results didn’t get any better, it seemed like there is a problem with the data set, I believe the reason behind that is that the data is very large it contains 480,000 rows, and we only have 8 columns in the dataset, in addition, as mention in the preprocessing section, the correlation between features were bad, that’s why the results couldn’t get any better regardless of any techniques used to improve the results.

Decision tree results:

Accuracy: 0.642

Precision: 0.625

Recall: 0.647

F1 score: 0.636

ANN results:

Accuracy: 0.562

Precision: 0.565

Recall: 0.564

F1 score: 0.567

* **Possibility of implementing this project in Jordan:**

In Jordan, we have 3 aviation companies, Royal Jordanian, Jordan aviation, and fly Jordan, all these aviation companies can benefit from a project like that to use the same model and but using their own datasets, that’s because the model will be trained on their data o it will better results when giving it similar data to predict. I think a project like that would definitely improve the aviation companies and the aviation sector in Jordan, because some of these airlines have delays problems which made their reputation bad.

* **Further enhancements in the future:**

In my opinion, the project can be improved a lot, by changing the training dataset, we will recognize a difference and improvement in results, as well as adding more hidden layers in the ANN model to understand more complex relations in the flights. Also, it’s important to tune hyperparameters correctly and efficiently to get better results and improve the model’s performance. The limitation for this project may be the low number of available dataset.

* **My role in building and improving the project:**

In this project, I build the models, chose the dataset, and done all the reprocessing needed, I’ve done things with understanding, knowledge, and interest at the same time. I liked the topic of this project so I’m thinking of sharing my model and results on Kaggle, I will also work on improving and enhancing the project by applying things mentioned in the previous question.

## References

[**https://www.kaggle.com/datasets/jimschacko/airlines-dataset-to-predict-a-delay**](https://www.kaggle.com/datasets/jimschacko/airlines-dataset-to-predict-a-delay) **(The dataset source)**