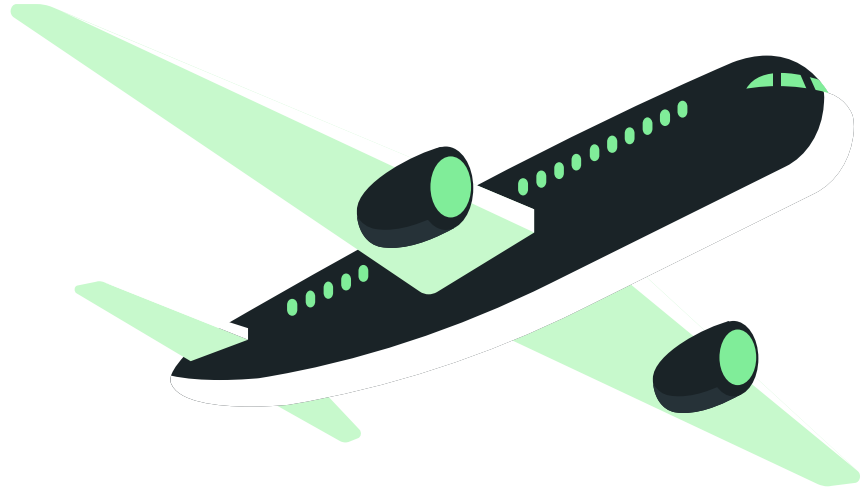# Airlines Delay

Supervised Learning

Artificial Intelligence
2022/2023   3LEIC10   Group 27

Daniela Tomás   up202004946
Diogo Nunes   up202007895
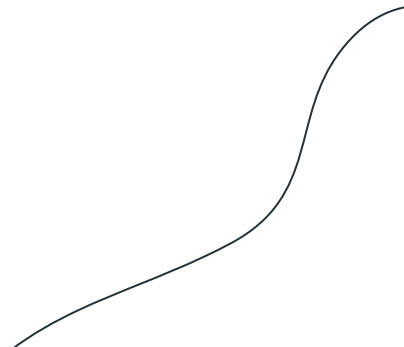João Veloso   up202005801

Work Specification

The task at hand is a binary classification problem, to predict whether a given flight will be delayed, given the information of the scheduled departure.

Dataset attributes:

- **Flight -** Flight ID
- **Time -** Time of departure
- **Length -** Length of Flight
- **Airline -** Airline ID
- **AirportFrom -** Which airport the flight flew from
- **AirportTo -** Which airport the flight flew to
- **DayOfWeek -** Day of the week of the flight
- **Class -** Delayed (1) or not (0)

Tools and Algorithms

- **Programming language:** Python

- **Python libraries:** Pandas, Seaborn, Scikit-Learn, …

- **Development environment:** Jupyter Notebook / VSCode

- **Machine learning algorithms:** Decision Trees, K-Nearest Neighbours, Support Vector Machine, …

Work Already Done

The dataset has been analysed and pre-processed. The following conclusions were drawn:

- There are 539382 rows and 8 cols in the dataset
- There are 216618 duplicates
- There are no null values
- Length has some outliers

Attribute	Type	Properties	Discrete vs Continuous
Flight	Nominal	Distinctness	Discrete
Time	Interval	Distinctness, order, and meaningful differences	Continuous
Length	Ratio	Distinctness, order, meaningful differences, and ratios are meaningful	Continuous
Airline	Nominal	Distinctness	Discrete
AirportFrom	Nominal	Distinctness	Discrete
AirportTo	Nominal	Distinctness	Discrete
DayOfWeek	Ordinal	Distinctness and order	Discrete
Class	Nominal	Distinctness	Discrete

Data Pre-processing

Our dataset presented to be ready to use as:

- There are no **null values**

- Although there are **duplicate values**, we decided not to remove them because they could represent regularly scheduled flights that are usually on time/delayed.

- There are few (<5%) **outliers** in 'Length', but they don't necessarily represent errors, just very long flights, and so we choose not to remove them.
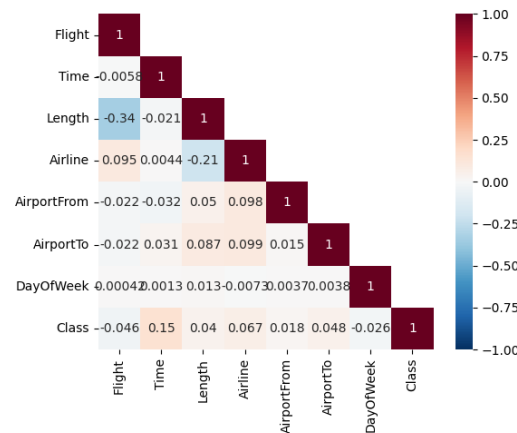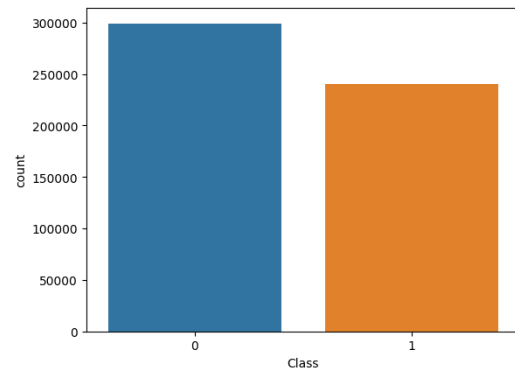
Developed Algorithms

Decision Tree

- This model was based on the implementation of the **DecisionTreeClassifier** from the **sklearn.tree** module.
- max_depth = 5

Support Vector Machine

- This model was based on the implementation of the **SVC** from the **sklearn.svm** module.
- We use a smaller subset of the data to train the model faster. If we didn't, the model would not execute in doable time.
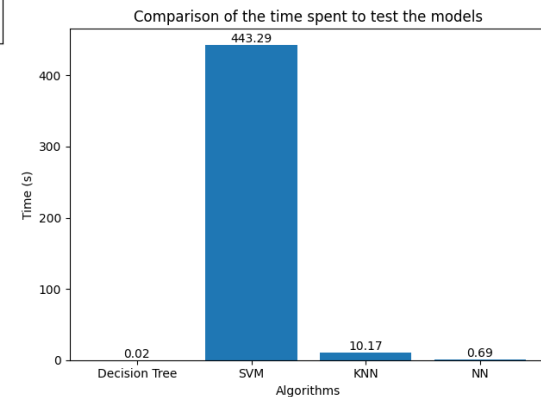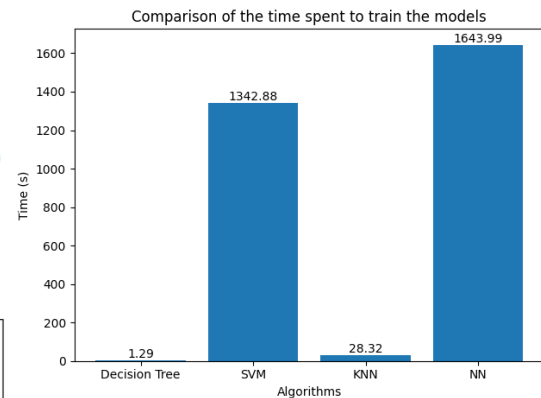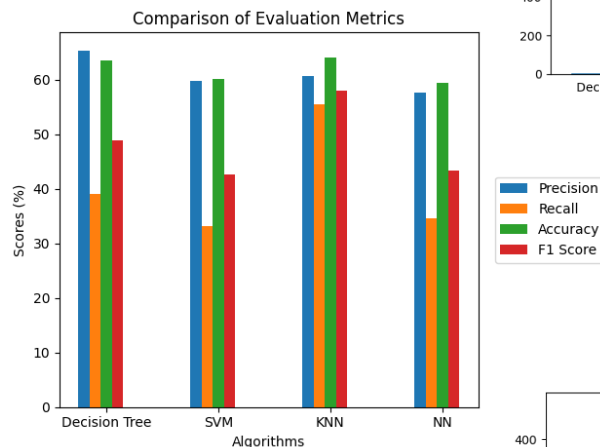- kernel = 'rbf'

K-Nearest Neightbours

- This model was based on the implementation of the **KNeighborsClassifier** from the **sklearn.neighbors** module.
- n_neighbors = 5

Neural Networks

- This model was based on the implementation of the **MLPClassifier** from the **sklearn.neural_network** module.
- hidden_layer_sizes=(100,)
- max_iter: 1000
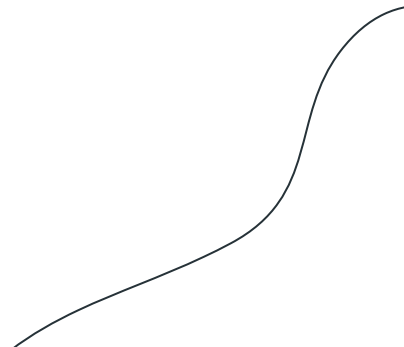
Developed Algorithms

- DT has a slightly higher **precision** score (65.4%) than the other algorithms.

- KNN has the highest **recall** (55.51%) and **accuracy** (64.06%) scores.

- DT and KNN models also have the higher **F-1** scores (48.97% and 58.0%, respectively).

- In terms of **training time**, DT was the fastest (1.29s) and NN the slowest (1643.99s ≈ 27 min).

- In terms of **testing time**, DT was once again the fastest (0.02s) and SVM was by far the slowest (443.29s ≈ 7 min).

- But we must consider that the train size of SVM (10%) was smaller than the others (30%).

Comparison of Evaluation Metrics

Comparison of the time spent to train the models

Comparison of the time spent to test the models

Conclusion

With this project, we have learned to work with various Machine Learning models directed at Supervised Learning.

Overall, **Decision Tree** and **K-Nearest Neighbors** models performed relatively better than **Support Vector Machine** and **Neural Network** in predicting flight delays. They achieved higher evaluation metrics scores and by far the fastest train/test times.

References

- IART classes slides and exercises

- https://www.kaggle.com/datasets/ulrikthygepedersen/airlines-delay?datasetId=2859795&sortBy=dateRun&tab=bookmarked