In this assignment, the objective is to explore two datasets (Iris and Breast Cancer) performing a Exploratory Data Analysis (EDA) and evaluating different classification models, aiming to compare the different algorithms and compare out how they work and which results in better statistics like accuracy, f1-score, etc.

For the Iris dataset, the EDA revealed that there are 3 rows of duplicated data and no nulls. The dataset is balanced since each of the three species contains the same number of samples. Also, using the violin plots I figured out that the Setosa is more separated than the others. All classification models (Logistic Regression, Decision Trees, SVM and Random Forest) achieved perfect accuracy on the test, including the MLP, which has been tried with different numbers of layers and neurons, although the simplest architecture (one layer with 10 neurons) struggled to converge with only 1000 iterations, that had been used for every model. We can confirm that in this case the complex models are unnecessary.

On the other hand, the Breast Cancer dataset was more challenging due to its higher dimensionality (30 features) and more complex nature. Feautres like area worst and concave points worst showed strong separation potential between malignant and benign classes. The EDA also revealed high multicollinearity among features like radius or perimeter. In terms of models, the SVM performed robustly with 98% accuracy but the best results were given from the MLP. A simple architecture with a single layer with 10 neurons achieved the best performance with a 99% accuracy, missing only one false negative. Deeper networks did not improve and showed signs of overfitting, confirming that for this datasets the simpler models the better.

During the task, i found that the Breast Cancer dataset required scaling because some features like area or smoothness had different scales. Although the Iris dataset was clean, I needed to remove the ID column from the Breast Cancer. For both datasets, I applied a label encoder for the label column for transforming it into numeric values for the classification models.