U.PORTO

FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

Cancer Data

Checkpoint Assignment 2 - Theme E

L.EIC - AI - T2 - Group 9

André Tomás da Cunha Soares
Diogo Alexandre da Costa Melo Moreira da Fonte
Jorge Carlos Baptista Duarte

Specification, Tools and Algorithms

The objective for this project is to develop a machine learning model with supervised learning algorithms that can help determine if cancer cells in our data are benign or malignant.

The dataset used has 570 cells with 30 features. All the features are represented with continuous values and focus on different aspects of a cancer cell's shape and size.

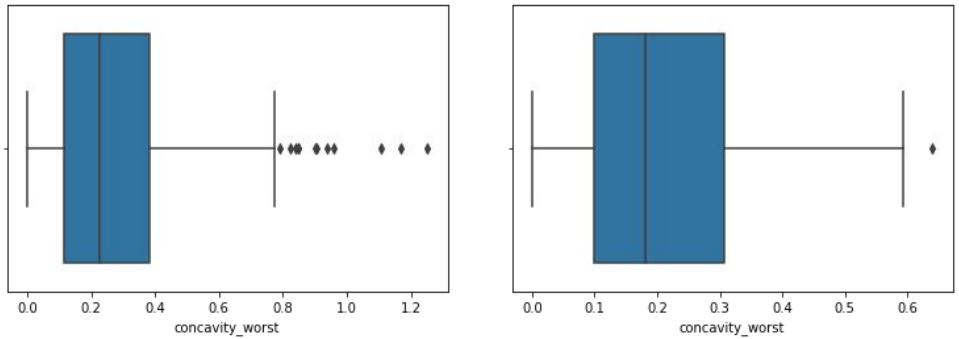
Programming Language:
We used python in the Jupyter Notebook environment and Scikit-Learn for the analysis of the dataset and to process the data.

Algorithms:
- Decision Tree
- Support Vector Machine
- Neural Network

Data Preprocessing

A first analysis of the dataset and treatment of some fields has already been done. There was a field (“Unnamed: 32”) that was full of null values, so we deleted it. We decided to create a correlation matrix to see how certain variables were connected to each other.

After that our approach was removing the outliers found outside of the range of the mean plus or minus the st_deviation. This is an example of one variable to show that the outliers have been removed. Not all of them, because sometimes if the supposed outlier is close enough, it may not be an outliers but actually an important piece of data.



Balancing the Dataset - SMOTE

The original dataset is unbalanced as there are several more Benign cases than Malignant ones.

In order to address this we use the SMOTE technique that works by oversampling the examples in the minority class.

We use the function fit_resample and then print the result to guarantee that the diagnosis column is perfectly balanced.

Class Benign and Class Malignant now have 308 entries each.

Feature Engineering

In an effort to increase the performance of our model, we looked into the columns with highest correlation and merged them into new ones.

Decision Tree Algorithm

Accuracy: 81.90%

Confusion Matrix:

	Predicted Benign	Predicted Malignant
Actual Benign	70	8
Actual Malignant	11	16

Precision: 0.67

Recall: 0.59

F1 Score: 0.63

Training Time: 0.01s

Support Vector Machine Algorithm

Accuracy: 73.33%

Confusion Matrix:

	Predicted Benign	Predicted Malignant
Actual Benign	76	0
Actual Malignant	28	1

Precision: 1.00

Recall: 0.03

F1 Score: 0.07

Training Time: 0.00s

Neural Network Algorithm

Accuracy: 90.48%

Confusion Matrix:

	Predicted Benign	Predicted Malignant
Actual Benign	22	7
Actual Malignant	3	73

Precision: 0.91

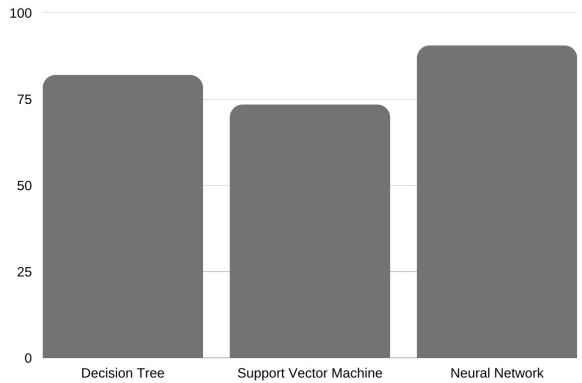
Recall: 0.96

F1 Score: 0.94

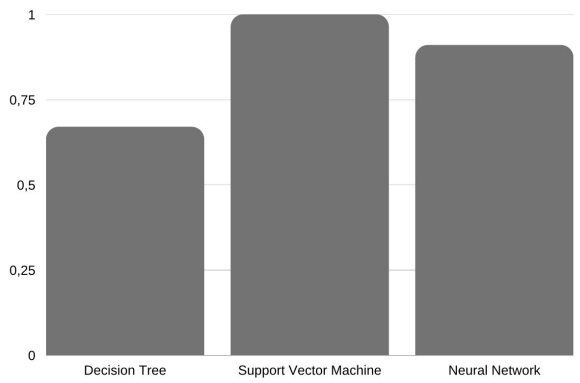
Training Time: 2.89s

Comparison between the Algorithms

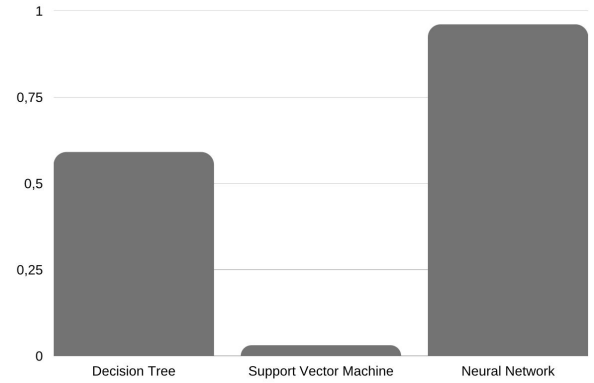
Accuracy



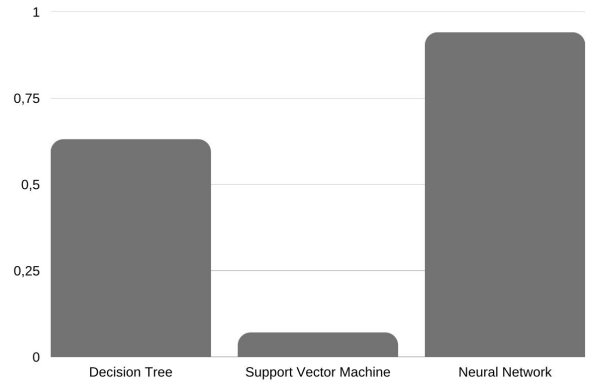
Precision



Recall



F1 Score



Comparison between the Algorithms

We can’t only compare the accuracy of the three models, it's crucial to take into account the particular context and application of cancer data detection. False positives (FP) and false negatives (FN) in medical application can both have serious repercussions.

The SVM was our worst performing model, because it had trouble distinguishing benign and malignant cells even though it has a perfect precision. It had trouble working with our dataset.

The Neural Network Classifier was our best model, because it could predict the diagnosis in most situations. The precision shown was 0.91, demonstrating a low proportion of FP and the recall of 0.96 means that almost every malignant case was detected. The F1 Score was 0.94 and this provides an overall evaluation by accounting for both FP and FN.

References

- Cancer Dataset used for the project.
- Web Page available in the project zip (index.html)