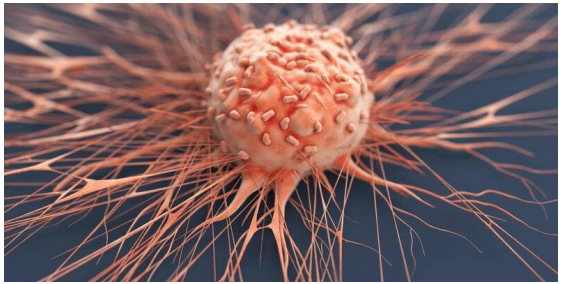
Artificial Intelligence Cancer Data Analysis

João Alves - up202007614@fe.up.pt
José Ribeiro - up202007231@fe.up.pt
Rúben Monteiro - up202006478@fe.up.pt

Problem Description - Cancer Data

Objective: Develop a successful machine learning model that can predict whether or not a cell is benign or malignant.

Dataset: 30 features of 570 different cells, along with the id and the diagnosis of each case. The diagnosis is represented as either B (benign) or M (malignant).



Algorithms and Tools

Python Libraries:

- Numpy
- Pandas
- Matplotlib
- Seaborn
- Imbalanced-learn
- Scikit-learn

Algorithms used (so far):

- Decision Tree
- KNN
- SVM
- Neural Network

Progress

Data preprocessing:

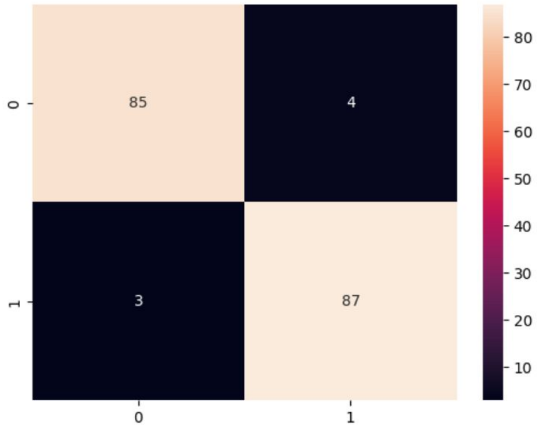
- Get a better understanding of the dataset
- Eliminate unnecessary column
- Remove outliers from the dataset
- Balancing results (currently by oversampling)

Development of model:

- Use of GridSearchCV to find the best parameters for each of the current 4 algorithms.
- Analysis of the results in a confusion matrix

Current Results:

- So far the algorithms with the best results are the Decision Tree and the Neural Network.
- Data preprocessing can still be improved in order to have better models



References

Scikit-learn documentation: https://scikit-learn.org/stable/index.html

Theoretical Class Slides