Supervised Learning
Benign and malignant cancer

Class 9, group 96
António Maria Gameiro Campos Sampaio de Matos – 202006866
Diogo Ferreira Neves – 202006343
Fábio Almeida Teixeira - 202006345

Project Description

The task at hand is a supervised learning problem, where we want to learn how to classify cancer cells into benign or malignant based on 30 different features. The dataset consists of 570 cancer cells, and the goal is to build a predictive model that can accurately classify new, unseen cancer cells as either benign or malignant.

The focus of this project is to compare the performance of the different algorithms we implement to achieve our goal.

References

- Kalaiyarasi, M. (2020). Classification of Benign or Malignant Tumor Using Machine Learning https://iopscience.iop.org/article/10.1088/1757-899X/995/1/012028
- Sikder, J., Das, U. K., & Chakma, R. J. (2021). Supervised learning-based cancer detection. *International Journal of Advanced Computer Science and Applications*, *12*(5). *https://www.researchgate.net/profile/Juel-Sikder/publication/352080337_Supervised_Learning-based_Cancer_Detection/links/60befeed458515218f9f31a2/Supervised-Learning-based-Cancer-Detection.pdf*
- Assegie, T. (2020). An optimized K-Nearest Neighbor based breast cancer detection *https://journal.umy.ac.id/index.php/jrc/article/view/8593*

Tools and Algorithms

Planned algorithms

Decision Tree	Builds a tree-like model of decisions (nodes) and their consequences (branches). By selecting the best attribute to base the split of the data, it keeps recursively building the tree until a certain criteria is met.
K-Nearest Neighbors (KNN)	Predicts values of a query point by finding the K nearest data points in a training dataset and using those to formulate a prediction.
Support Vector Machines (SVM)	Finds the “best hyperplane”, separating data points into different classes. This is done in a way that maximizes the margin between the closest points of different classes, increasing effectiveness of the algorithm in its predictions.

Libraries

NumPy	Provides support for numerical computations.
Pandas	Provides a wide range of tools for data manipulation.
Scikit-learn	Includes implementation for the chosen algorithms.
Seaborn	For statistical visualizations.
Matplotlib	Creating plots for data visualization.

Implementation work carried until checkpoint

We have started working on analyzing our dataset to check if any data pre-processing is required.

This is being done with the help of the pandas library, which allows us to easily get some insight on the state of the data such as the existence of missing values, outliers, or inconsistent data types.