Cancer Dataset Study

First Check Point - Group 85

IA 22/23

Lia Vieira - up202005042

Marco André - up202004891

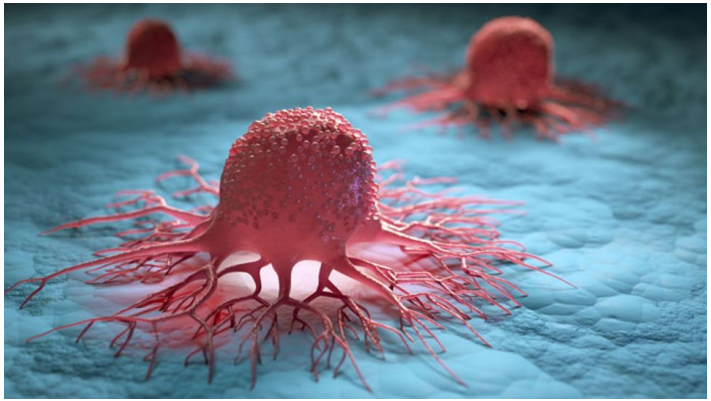
Ricardo Matos - up202007962

Specification of the work performed

Our goal is to make a model able to diagnose if a cancer is malignant or benign. To train the model and test it we have access to data that comes from Kaggle. We will then have to clean the data, looking for possible imbalances, outliers, handle missing data, standardization and normalization, data validation etc.

The project will involve the following steps:

- definition of the project
- data collection (already provided)
- preprocessing data
- feature engineering
- model selection
- training
- model evaluation



References & Bibliographic Search

For this project, we opted to use the recommended programming language: **Python 3.10**.

Because we chose Python, we are fortunate to have much of the work needed for this project already done for us as there are lots of packages available with tools to study and train models on the most varied datasets.

With that in mind, and following what we already began in the practical classes worksheets, we focused our attention on studying the **Scikit-learn** and **Seaborn** packages and what they had to offer, learning a lot their capabilities.

In conclusion, the theoretical slides worked as our research guide and we then found tools to meet our needs. Fortunately, these Python Libraries are very popular and so there are a lot of resources available (namely the official documentation for them that are very helpful and surprisingly complete with examples)

Description of the tools and algorithms used

For this project we are expected to use several well-know machine learning libraries, as :

- **Pandas**: Data manipulation and analysis for tabular data
- **NumPy**: Library for numerical computations and mathematical functions.
- **Seaborn**: Data visualization for exploring complex patterns.
- **Matplotlib**: Plotting library for custom visualizations.
- **Scikit-learn**: Machine learning library with supervised and unsupervised algorithms.

Implemented models:

- Decision Trees
- Neural Networks
- K-Nearest Neighbors (KNN)
- Support Vector Machines (SVM)
- Random Forest
- Logistic Regression

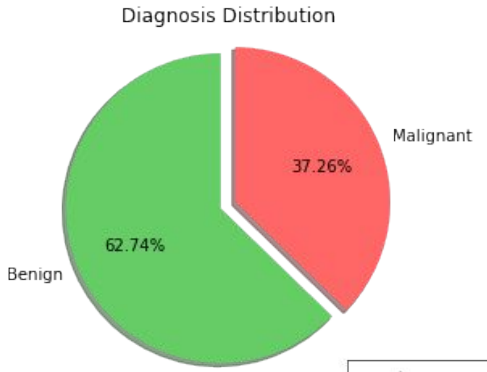
We're also interested in applying other algorithms to our dataset and study their effectiveness.

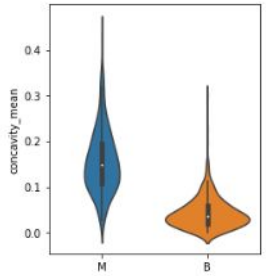
Dataset Analysis

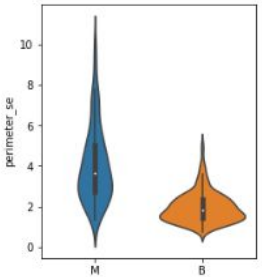
After doing data analysis, we were fortunate to already have a very clean dataset with now null, N/A, NaN, missing or duplicated values. The only thing we have to look at are data balancing and dealing with outliers.

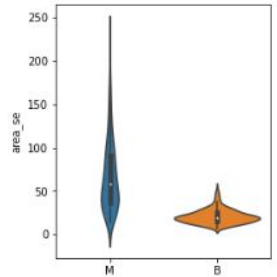
Our data has 63% **B** against 37% **M** cases which we must deal with by balancing the data to avoid having skewed results.

When concerning outliers, with violin plots we can clearly see some values that are very distant from the mean values. We'll talk more about removing them in the next slide.









Removal of Outliers

- It **impacts** the accuracy of the models.
- Can be a **problematic decision**
- **Removing** lots of outliers leds to overfitted models
- How can we be sure of a outlier ?
  - low dataset
  - rare cases that can depend on other feature non-available (Ex. location of cancer)
- Experiments with **unsupervised learning** to check outliers (LOF, GMM) and a self made algorithm

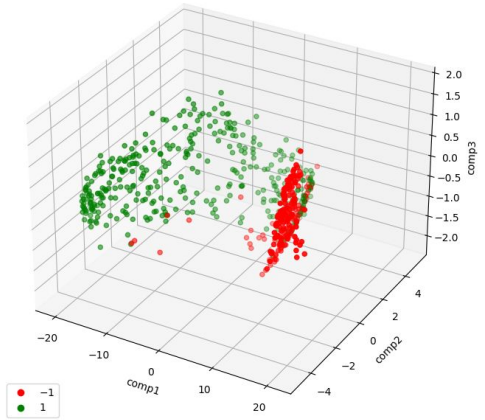


Fig.: points representing the data

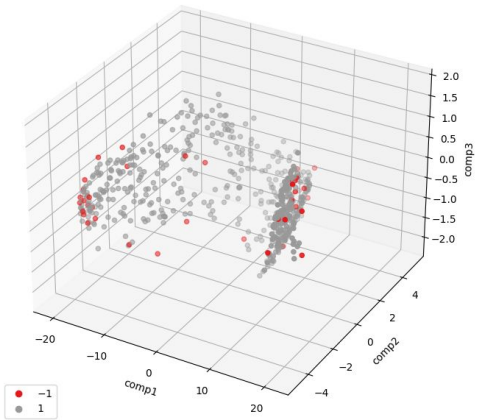
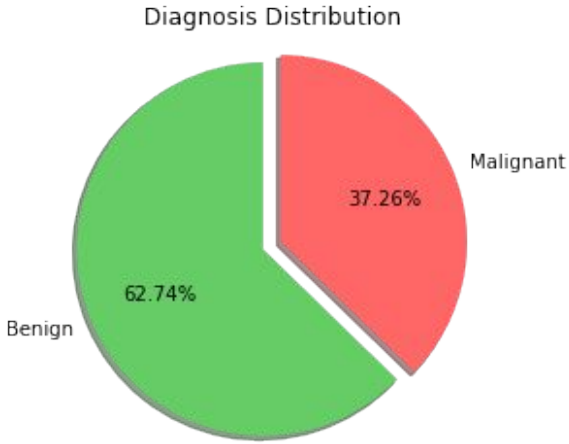


Fig.: points representing outliers (GMM)

Recursive Feature Elimination

- We used this feature selection technique to select the most **important** attributes.
- From the **30** given attributes we selected **10**, which helped us reduce the complexity of the model and the training time as well as removing redundant and irrelevant attributes.
- It also has an impact on the features that need to be collected - if we can remove some attributes, we can reduce the cost of collecting them.
- This makes the model more useful in real world applications.

Feature Balancing



- Needed to prevent biases towards the majority class
- 3 approaches followed:
  - Undersampling
    - Not a good solution (reduced dataset)
  - Oversampling
    - Duplicates a lot of data, can lead to a vicious evaluation
  - SMOTE
    - Generates synthetic samples of the minority class, can lead to a vicious evaluation

Normalization

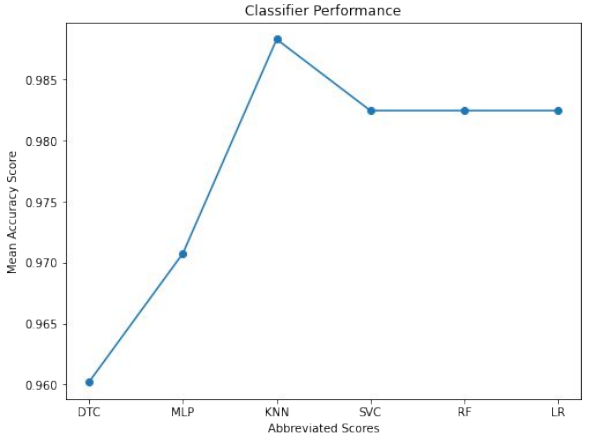
- Crucial to ensure that all input features have a similar scale and distribution
- Prevent **features** from **dominating** other
- Necessary to improve overall accuracy in neural networks, svms …
- A lot of options to do standardization in sklearn (StandardScaler, MinMaxScaler, RobustScaler)
- We use **StandardScaler**
- Normalization is very impactful in some models and not so much in others

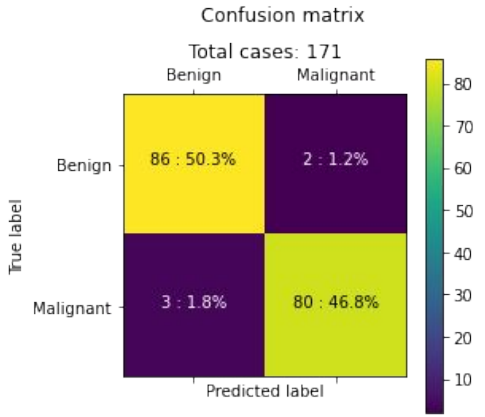
Model parametrization and evaluation

Like earlier stated, we tried **6 different models** with varied hyper-parameters and balancing solutions. In the end, the different models got very consistent **accuracy** scores of **95-98**% with very high values for **recall**, **precision** and **F1** also, while being very consistent between iterations.

Unfortunately, as the **dataset was rather small**, our conclusions were limited as the models did very well regardless of the configurations we set.

An important note is that, even though we balanced the data, the models were more prone to produce **false negative** (saying a cancer is benign when in fact it was malignant), which is much more troublesome than false positives.





Conclusions

This project enhanced our understanding of supervised models in AI and emphasized the importance of data analysis and preprocessing.

However, it is worth noting that the limited size of the dataset has somewhat constrained our research. A larger dataset would have allowed for a more comprehensive analysis and encouraged us to delve further into advanced techniques and explore additional avenues of investigation.

Despite this limitation, the project provided a solid foundation for further research and highlighted the significance of robust data treatment in machine learning tasks, enticing our curiosity for the field of AI.