Design of AI Systems (DAT410)
Assignment 4: Diagnostic Systems

Luca Modica
Hugo Manuel Alves Henriques e Silva

February 11, 2024

1 Reading and reflection

1.1 First paper: Machine learning techniques to diagnose breast cancer from image-processed nuclear features of fine needle aspirates

The main goal of the article is to explain how, through ML techniques and information from digital scans of fine-needle aspirate (FNA) slides, is possible to improve the performance of a breast cancer diagnoses system [3].

In the initial part, a detailed explanation of the diagnoses process and the characterization of the nuclear features is given. The FNA procedure was subjected to a sample of 569 patients (212 cancer and 357 benign); then, the related cell nucleus information were obtained with the following steps:

1. The area of the aspirate slides were first prepared with a proper camera equipment and to have a minimal nuclear overlap.
2. Then, through a GUI to analyze the digital images, an active contour model (called "snake") is used to locate the boundary of the cell nucleus; in particular, the "snake" is a deformable spline that seeks to minimize a defined energy function, with the goal of conforming to the boundary of a cell nucleus.

The snakes allowed to obtain several features related to the cell nucleus: for each of them, the mean, the worst value (mean the biggest 3 values) and the standard error of the feature are taken into account, for a total of 30 features for each case study. The features will be now grouped based on how they compared the digital feature assessments with the visual characteristic: that is, by generating and measuring 4 different possible shapes (also called *phantoms*).

- The first 2 features are **radius** and **area**, used to measure the *size* of the generated shapes.

- Those features are followed by: **smoothness**, **concavity** (in a cell nucleus), **compactness**, **concave points**, **symmetry** and **fractal dimension**. In the context of Phantoms measurement, they are used to compute their *shape*.

As It will described later in the report, the above notions of *size* and *shape* used for the generated shapes will be also assumed by our model implementations, especially for the Rule-based classifier development. Finally, the other features extracted by the snakes are the **perimeter** and the **texture** (variance of the grey scale intensities in the component pixels). Moreover, as It will be assumed by our Machine Learning experimentation, the features are also modelled such that higher values are typically associated with malignancy.

The Machine Learning procedure is then described. The classifier consists in a linear programming iterative method, which place separating planes in the feature space until mostly points in a region belong to the same category. In order to generalize well in unseen cases, the model is built with the goal to minimize the number of separating planes and the number of features used; this because simpler classifier usually performs better than complex ones and, as it will be described, they are more interpretable.

In the last part of the paper, to conclude, the results of the classifier are discussed. With the described approach and a tenfold cross-validation to estimate diagnostic predictive accuracy, the accuracy reached by the classifier is 97%; despite the results different possible improvements are described from using more appropriate image processing methods to a better training for using the dedicated computerized system for the analysis.

1.2 Second paper: The Mythos of model interpretability

This article seeks to refine the discourse on interpretability [1]. Firstly, to introduce the topic, nowadays predictive models are not capable of reasoning at all. Models are essentially trained to take some input and predict the corresponding output. Although, these ML-based systems don't know exactly the reason why a given input should receive a certain label, but only the fact that some inputs are correlated with that label.

This is where interpretability plays a huge role. Interpretability refers to the extent to which a human can understand the cause behind a decision made by a model. Models' workings would, therefore, need to be presented in an understandable manner, allowing users to understand why certain predictions were made instead of others.

When talking about areas such as medicine, the criminal justice system and financial markets, the inability of humans to understand these models can be greatly problematic as without understanding how predictions are made, we can't ensure that decisions are made fairly, accountably, and can be audited or challenged if necessary.

Standard ML objectives, like minimizing a function and maximizing accuracy may not fully align with complex real-world applications, such as medical research or dynamic environments like online retail. On the other hand, interpretability could possibly address these mismatches by revealing deeper insights and adapting to evolving data, ultimately making models more applicable and understandable in practical scenarios.

The paper focuses mainly on supervised learning in real-world applications as well as an interest in a common claim that "linear models are interpretable while deep neural networks are not". Understandable models will be considered *"transparent"* and incomprehensible models will be called *"black boxes"*.

Other studies have been made "post hoc interpretations" in machine learning, which provide explanations for model predictions without explaining the model's internal mechanisms. This approach somewhat contrasts with direct interpretability as it desires for clear explanations and reasoning behind decisions.

Research in interpretability is motivated by the discrepancy between the formal objectives of machine learning models, as referred before (accuracy), and the complex, multifaceted requirements of real-world applications. The demand for interpretability in models obviously adds more objectives to them which are challenging to quantify mathematically and make models take them into account.

Important model characteristics:

- **Trust**: user confidence in models' accuracy and fairness, ensuring they align with real-world objectives and ethical considerations.
- **Causality**: being able to identifying potential causal relationships in data and by interpreting the model itself, to generate hypotheses for further scientific testing.
- **Transferability**: the ability of models to maintain their predictive power under conditions they were not explicitly trained for, a task which humans are mostly capable of doing; models should preferably be robust and adaptable to changes in the environment where they are applied.
- **Informativeness**: being able to provide information and insights to the users, beyond just accurate predictions. Such ways could be pointing similar cases to support decisions serving as a bridge between simple error reduction and delivering real-world value.
- **Fair and ethical decision making**: As mentioned before, there are demands for decisions that are ethical and do not discriminate, aligning with fairness standards.

1.2.1 The Transparency Notion of Interpretability

- **Simulatibility**: Transparency at the level of the entire model. Simplicity is the main characteristic of this type of transparency as it emphasizes on total comprehension of the model by a person. Interpretability is not intrinsical to any model specifically and it can vary on their complexity.
- **Decomposability**: Transparency at the level of individual components such as parameters. This relates more in interpretability on each component (input, parameter, calculation). This means that different parts of a model, like decision tree nodes or linear model parameters, should have clear, plain-text explanations, linking features directly with their influence on the outcome. However, not all inputs are interpretable, and parameters might be misleading if the full context of the data and the model structure are not taken into consideration.
- **Algorithmic transparency**: Transparency at the level of the training algorithm. There are contrasting views on linear models and deep learning ones: while in the first the error surface and convergence to a unique solution are well-understood, which offers predictability in behavior, even when new data is used, deep learning models are not fully understood. These do not provide certainty in performance on new data. There is, therefore, a gap in predictability and understanding between traditional and modern machine learning approaches.

1.2.2 Post Hoc Interpretability

Post hoc interpretations often do not describe exactly how a model works, but they could provide useful information for practitioners and end users of ML. Below, some common approaches to post hoc interpretations will be described.

- **Text Explanations**: This concept revolves around training two models, one to make decisions and another, such as a recurrent neural network, to generate verbal explanations. While these explanations may not describe fully accurately the decision-making process, they are based on mimicking observed human explanations or interpreting latent features, which can be seen in recommender systems for instance.
- **Visualization**: Renders visualizations in the hope of determining qualitatively what a model has learned. Popular methods such as t-SNE (t-distributed stochastic neighbor embedding) or gradient descent modifications on input images share this goal. However, there is still a lack of rigorous standard of correctness as it focuses more on the appeal and interpretative potential of the visualizations rather than their accuracy in explaining model decisions.
- **Local Explanations**: The goal is to clarify the way neural networks make decisions based on specific inputs, using methods like saliency maps

for pinpointing influential features. Although, these explanations can be unstable and can change with small input modifications.

- **Explanation By Example**: Mechanism for explaining the decisions of a model by reporting other examples that are most similar to it. By mirroring human analogy-based reasoning, this improves understanding by drawing parallels with familiar examples.

1.2.3 Conclusion

From the paper we can retain that linear models are not strictly more interpretable than neural networks, as it depends on the notion of interpretability employed. Trade-offs have to be made when choosing between these 2 models, as neural networks tend to operate on raw or lightly processed features while linear models have to operate on heavily processed features to get comparable results.

One of the main points to retain is that interpretability must be qualified. Assertion regarding interpretability should fix a specific definition. Moreover, it is important to be careful when to give up predictive power in order to obtain transparency. Finally, relying on post hoc interpretability has some dangers, as it can be tailored to meet subjective preferences and may mislead by presenting plausible yet inaccurate explanations.

As far as future work in interpretability is concerned, developing more sophisticated loss functions and metrics that connect the real-world and ML objectives and diving deeper into other paradigms such as reinforcement learning could model interactions and learn causal relationships, even though it may involve some real-world experimentation risks. Innovative approaches beyond traditional metrics and model paradigms are going to be needed to address complex issues such as fairness in ML.

2 Implementation

In this section we will implement and evaluate the three classifiers in the proposed task. The task is to classify different instances of tumors between benign or malignant, based on their cells' characteristics. In the given dataset, which is dataset of patients tested for breast cancer using a fine needle aspirate (FNA), we have the multiple features. Each patient was tested multiple times, and the mean value, the standard deviation and the worst value were recorded for each patient. The features are: **radius, texture, perimeter, area, smoothness, compactness, concavity and concave points**.

The goal of this section is to compare this three classifier (rule-based, random forest and decision tree) in the given context, both in terms of performance and interpretability. While in the first part the models' setup will be described, in the next one their result and level of interpretability will be discussed.

2.1 The first model: Rule-Based Classifier

In this classifier, a number of different rules will be set in order to classify different instances of cells. These cells could either be malignant or benign. It is relevant to note that these rules will be based on studies described in the given papers.

To model the rules of the classifier, from the dataset we extracted values that denoted as **baseline values**: they are represented by the mean of a feature in the dataset. The specific features for which we took the mean will depend on the rule we want to model and the concept we want to express. Moreover, for the same reason, for specific rules we take into account the means of a feature of either one class (*malignant* or *benign*).

Now we will proceed to explain each of the rules (a cell is considered malignant if It satisfies at least one of them).

- **Abnormal cell size** (features considered: *radius*, *area*).

  We consider the size of a cell as abnormal when It's significantly larger than the overall mean. More in particular:

  - the absolute difference between the cell radius and area and their respective baseline values have to be greater than their respective standard deviation baselines;
  - Both the worst radius and area values have the be greater than their respective worst baseline.

- **Abnormal cell shape** (features considered: *cell compactness*, *smoothness*, *concavity*, *concave points*).

  The shape of a cell will be considered as abnormal if:

- the absolute value of the differences between the cell compactness, smoothness, concavity and concave points and their respective baseline values have to be greater than their respective standard deviation baseline values;
- The cell compactness, smoothness, concavity and concave points values have the be greater than their respective worst baseline values.

- **Abnormal cell texture** (features considered: *cell texture*, *cell smoothness*).

  The texture of a cell will be considered as abnormal if:

  - the absolute value of the differences between the cell texture and smoothness and their respective baseline values have to be greater than their respective standard deviation baseline values;
  - the cell texture and smoothness values have the be greater than their respective worst baseline. values.

- **Abnormal cell homogeneity** (features considered: *cell symmetry*, *cell fractal dimension*).

  A cell homogeneity is considered as abnormal if:

  - the absolute value of the differences between the cell symmetry and fractal dimension and their respective baseline values have to be greater than their respective standard deviation baseline values;
  - The cell symmetry and fractal dimension values have the be greater than their respective worst baseline values.

The rules described above may be used as baseline for the explanation of the later 2 models, as well as its performance.

The **rule-based Classifier** is the most interpretable out of all of the models. It shows *simulatibility* by being transparent in its entirety. An abnormal cell (and thus, a malignant one) is clearly defined with the given rules, and there is no ambiguity in the classification.

Despite being very transparent and understandable, this classifier may lack in the accuracy aspect, as its simplicity might be a drawback in complex datasets where interactions between features are not linear or easily separable by simple rules: this could lead to lower accuracy compared to more sophisticated models. Moreover, especially considering more strict rules, rule-based classifiers can also tends to overfit the data, which can be serious in the context of medical diagnosis. For more explanation, detailed classification performance are reported below (Figure 1-2).

	precision	recall	f1-score	support
0	0.93	0.96	0.94	357
1	0.93	0.88	0.90	212
accuracy			0.93	569
macro avg	0.93	0.92	0.92	569
weighted avg	0.93	0.93	0.93	569

Figure 1: Classification report of the performance of the rule-based classifier.

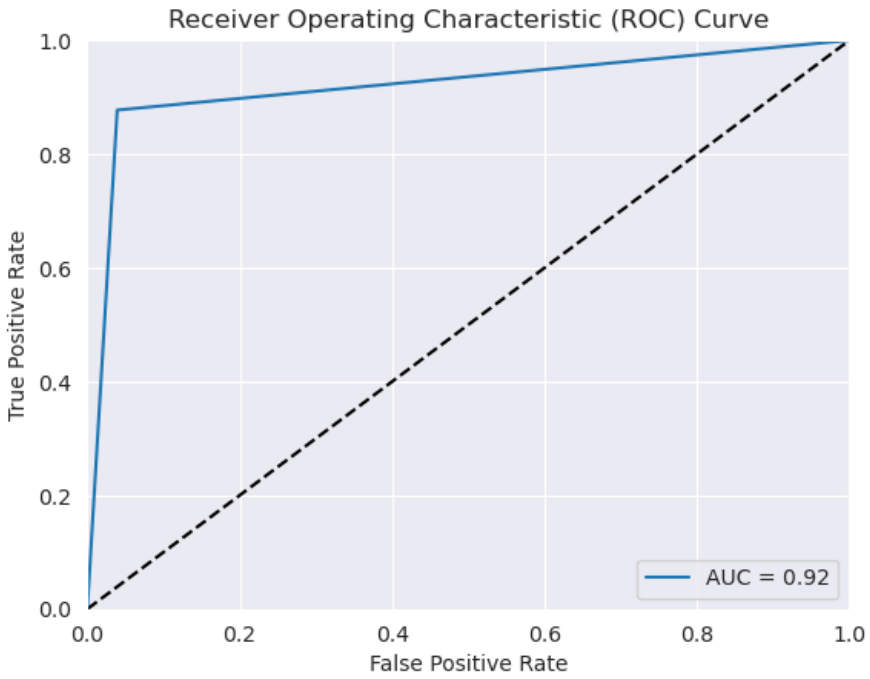

Figure 2: Receiver Operating Characteristic curve of the rule-based classifier.

As it can be noticed, despite the performance issue of such a simple classifier, It offers a very good baseline performance for the later 2 models. Although we have to be aware in these cases, this performance confirms even more the assumptions we made for the rules, always based on the paper "Machine learning techniques to diagnose breast cancer from image-processed nuclear features of

fine needle aspirates".

2.2 The second model: Random Forest Classifier

The second classifier for the experimentation is **Random Forest Classifier**, using the sklearn framework implementation. All of the 30 features are used, and the parameters fine-tuned using a *Cross-Fold Randomized Search* (100 iterations and 5 folds). The best estimator used has the following hyper-parameters:

- *number of estimators*: 50;
- *min samples to split*: 5;
- *max features taken into account*: the log2 of the total;
- *max depth of each tree*: no limit.

Comparing the results to the rule-based and decision tree classifier, the second model appears to be the best in terms of performance and predictions robustness. The results of the predictions made with the random forest classifier are reported below.

	precision	recall	f1-score	support
0	0.97	0.99	0.98	71
1	0.98	0.95	0.96	43
accuracy			0.97	114
macro avg	0.97	0.97	0.97	114
weighted avg	0.97	0.97	0.97	114

Figure 3: Classification report of the performance of the random forest classifier.

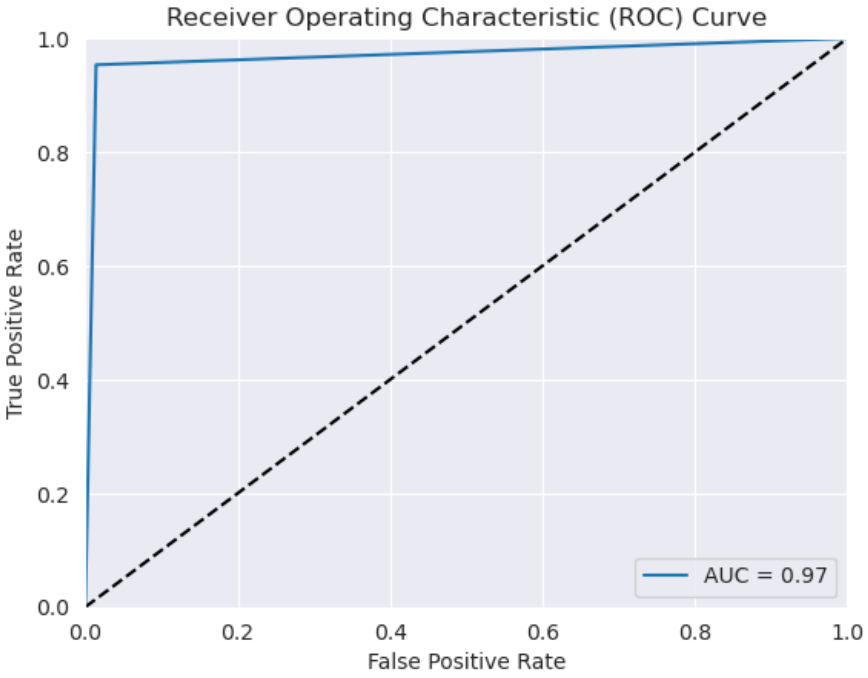

Figure 4: Receiver Operating Characteristic curve of the random forest classifier.

Alongside an overall good accuracy, it's worth mentioning the even better precision and recall values (Figure 3), as well as True Positive rate and False Positive Rate (Figure 4). This represents a good result, especially considering the task of medical diagnosis.

Despite the good model performance the random forest classifier is *not as interpretable as a rule-based and a decision tree classifier*, as there is not a clear binary decision. The final prediction is based on a majority decision of the many trees that compose the ensemble. Although a single decision trees could be one of the most interpretable models, the combination of different trees is not as clear. In these models, different trees are built on different subsets of data points and also different subsets of features, making it so that trees differ a lot from each other. To obtain full interpretability of a random forest, each tree needs to be understood which may be a hard task when scaling the model. Moreover, the decision-making process is not as interpretable since it's based on the aggregate decisions of potentially hundreds of trees.

Considering the difficulties in interpreting such a model, we experiment 3 different techniques which slightly improved this aspect.

- The first approach is to compute and see the top 5 **permutation importance** of the features. By highlighting the features with the most

important scores we can improve the model explanation, since they represent the feature that most influence the decision-making of the predicted category. Down below, the top 5 most influencing features will be displayed (Figure 5).

```
Feature: concave points_0, Importance: 0.009649122807017574
Feature: concavity_0, Importance: 0.0052631578947368585
Feature: radius_2, Importance: 0.0052631578947368585
Feature: perimeter_2, Importance: 0.0052631578947368585
Feature: perimeter_0, Importance: 0.004385964912280715
```

Figure 5: Top 5 features that most influence the decision making (or most deteriorate the performance) of the random forest classifier.

What we can observe is that the best features found aligned with the ones that are most used to model the rules in the previous classifier; this can imply a first good insight of how the random forest make decisions overall. The same reasoning will be found in the permutation importance of the features in the case of the decision tree classifier.

- The second approach, as a post-hoc visualization interpretability, we create and show a **Partial Dependence Plot (PDP)**. A PDP can illustrate the the effect of a specific feature on the predicted outcome of a model: this showing how the prediction probability can vary as the feature value increase its value. Down below, we consider the related plots of the top 3 feature from the permutation importance (Figure 6).

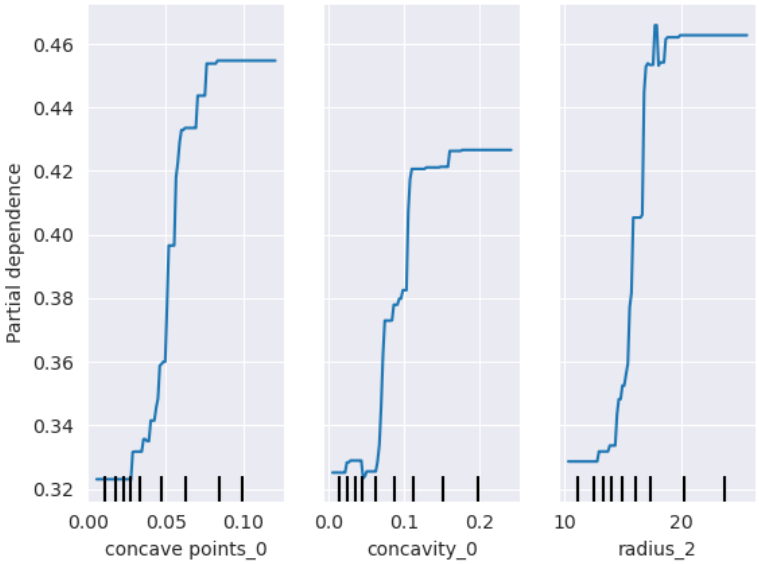

Figure 6: Partial Dependence Plot of 3 feature of the dataset: mean concave points, mean concavity and worst radius of the cells.

As assumed from the first paper of the report, we can notice that an increasing value corresponds to an increased probability that a instance is malignant. This assumption could be easily shown with this technique, since PDP also show if the relationhip between the feature and the response is linear, monotonic or even more complex.

- Since a Random Forest Classifier can represent a very complex model, It can be useful analysing more in details how it made prediction for single instances. We tried to investigate in this aspect with a third post-hoc local approach, which is using a **Local Interpretable Model-agnostic Explanations (LIME)**: It consists in the implementation and training of a local (simpler) model, to understand why the original classifier made a prediction for a specific datapoint [2]. Following this approach, down below we will show the results for both a benign and a malignant instance (Figure 7-8).

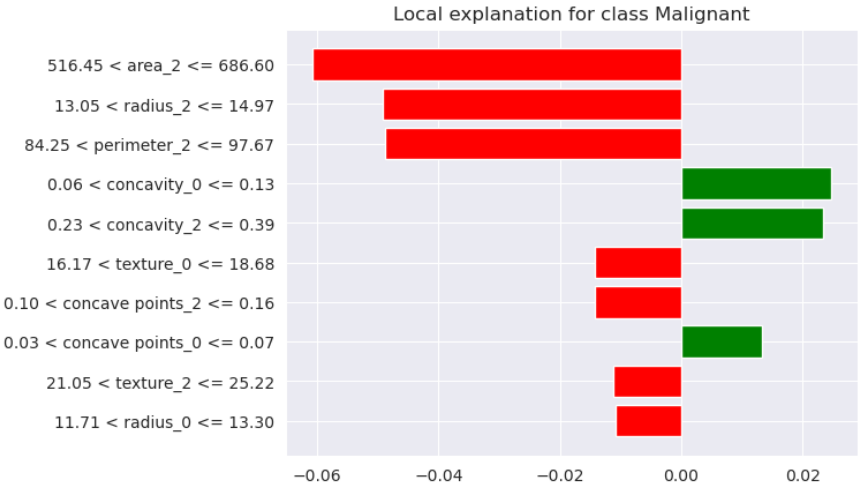

Figure 7: Explanation of how this datapoint was classified as *benign*, after applying the LIME technique.

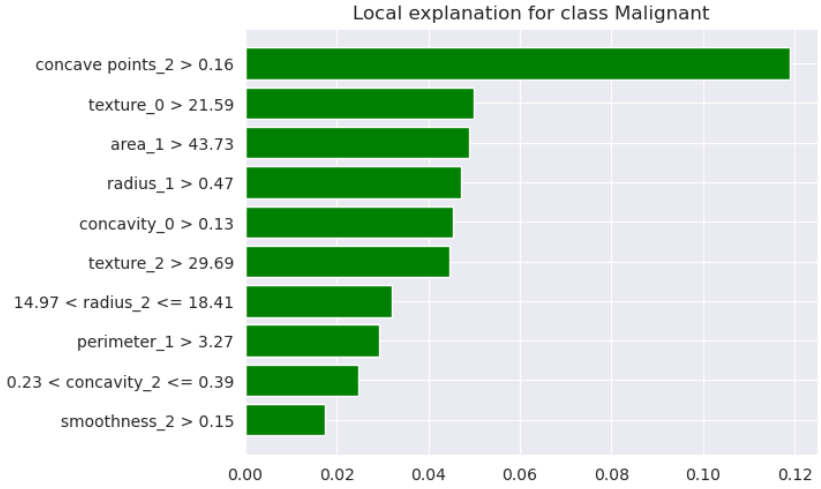

Figure 8: Explanation of how this datapoint was classified as *malignant*, after applying the LIME technique.

As It can be seen, LIME explain the results of a specific approach by showing how much a specific feature contribute in classify a datapoint as benign (in color red) or malignant (in color green), based how far the related value is far from the specified range. To proof the results shown, we compare the range values generated by LIME and the baseline values

used in the rule-based classifier, noticing they are reasonable in most cases (the overall mean of a value is greater when a LIME specify that a feature value should be greater than a threshold, for example).

Despite the interpretation above, as we specify in our take-aways from second paper, we need to be careful with local explanations: these can be unstable and can change with small input modifications.

2.3 The third model: Decision Tree Classifier

For the third classifier we attempt to chose one that better conciliate the trade-off between interpretability and classification performance. The model chosen to achieve this goal is a **Decision Tree Classifier**, for 2 main reasons:

- It's widely considered one of the *intepretable models* in the literature. This also considering the context analyzed, that is the Breast Cancer diagnosis.
- the performance compromises, which will be detailed in the later section, are acceptable.

The model is fine-tuned in the same way as the Random Forest Classifier, obtaining the following best hyper-parameters:

- *min samples to split*: 10;
- *max features taken into account*: no limit;
- *max depth of each tree*: 70;
- *criterion for split*: entropy.

The results obtained were slightly worse than the random forest classifier but better than the baseline rule-based classifier scores: It represents the trade-off we tried to achieve with a better intrinsic interpretability of the model. Down below, the performance results are visualized more in detail.

	precision	recall	f1-score	support
0	0.93	0.99	0.96	71
1	0.97	0.88	0.93	43
accuracy			0.95	114
macro avg	0.95	0.93	0.94	114
weighted avg	0.95	0.95	0.95	114

Figure 9: Classification report of the performance of the decision tree classifier.

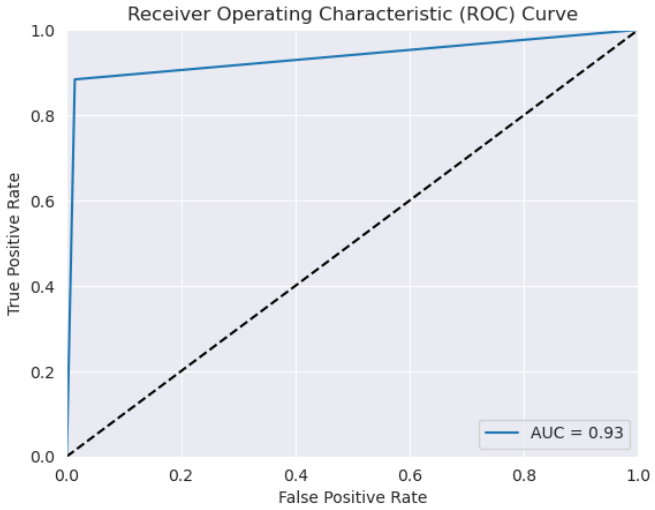

Figure 10: Receiver Operating Characteristic curve of the decision tree classifier.

As previously mentioned, using a decision tree for a classification task comes with a much greater interpretability. The main reason relies on the fact that we can visualize, through the built tree, the path the model takes to predict the category for an instance (Figure 11).

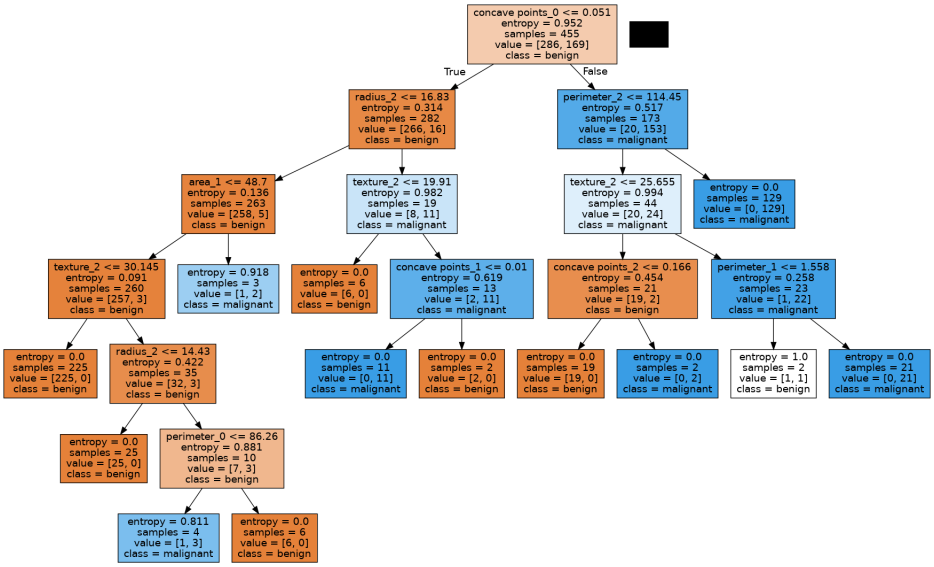

Figure 11: Visualization of decision tree classifier.

To complete the interpretability analysis of the decision tree, as for the random forest the top 5 features by permutation importance are listed below.

```
Feature: concave points_0, Importance: 0.21578947368421045
Feature: area_1, Importance: 0.1403508771929824
Feature: radius_2, Importance: 0.12456140350877191
Feature: perimeter_2, Importance: 0.02368421052631574
Feature: texture_2, Importance: 0.019298245614035026
```

Figure 12: Top 5 features that most influence the decision making (or most deteriorate the performance) of the decision tree classifier.

2.4 Considerations

Overall all the 3 models had very good results, both in terms of performance and intepretability; this also considering the relevance to the the 2 papers described in the first section of the report. The experiments and comparisons help us a lot on understanding how these Machine Learning models make decisions, in a so sensible environment such as medical diagnosis; this alongside future possible improvements and directions, from alternative models to test (Logistic Regression or Lasso), to other interpretability tools (both for direct and post-hoc intepretability) to understand predictions from other points of view (SHAP method or Individual Conditional Expectation plot).

One last consideration, more related to the dataset itself, is how the features interact to each other. The heat map below show an overall situation throughout the entire dataset (Figure 13).

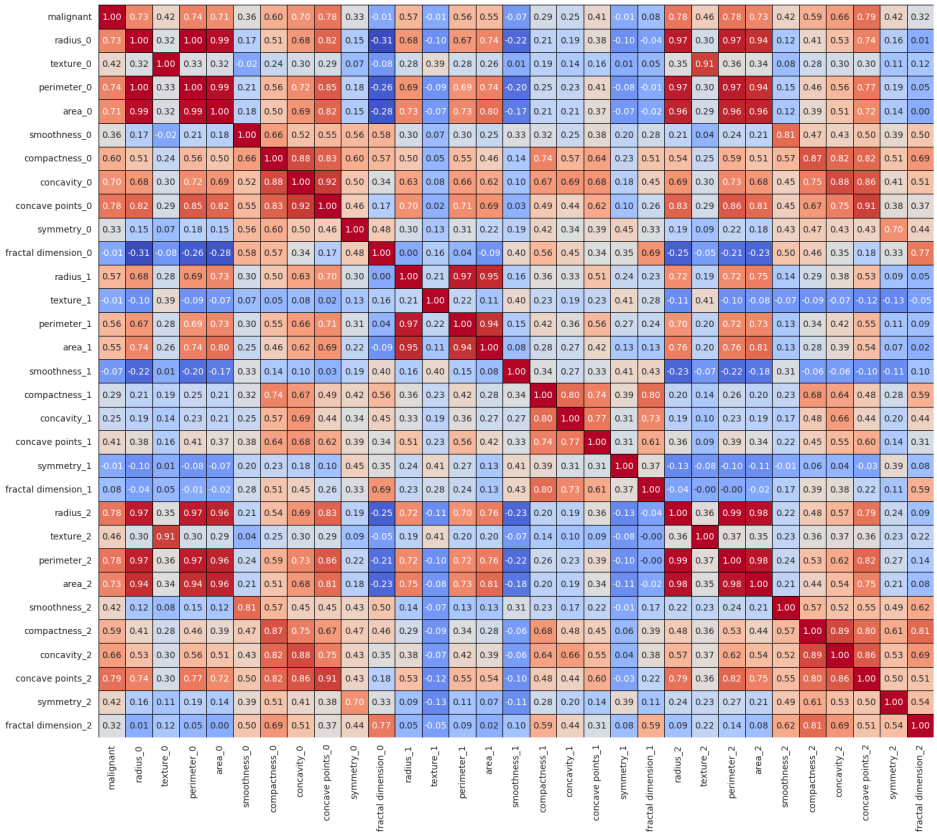

Figure 13: Heat map with all the correlation (interaction) between the features in the dataset, target variable included.

Apart from the trivial high interaction between features related to the same measurement (mean, standard deviation and worst value of the area for example), correlations worth mentioning among the measurements of *area*, *radius* and *perimeter* of the nuclear cell. In particular, they are almost perfectly correlated to each other; this is most likely due to the "snake" method identify the boundaries of cells in the image scans (as described in the first paper), alongside the geometric correlations the 3 measurements have.

While It may seem reasonable to remove highly correlated features for a simpler model, we decided to keep all of them. The main reason is that, heuristically, by keeping all the dataset information better results were obtained.

3 Discussion

By relying more and more In Machine Learning systems to make decisions, the concept of interpretability is becoming essential to understand the real reasons behind a specific predictions. This is even more crucial in more sensible in applications such as the one of this assignment, which is Healthcare and cancer predictions: medical practitioners need to trust and validate a predictive model, before using it into a real environment.

In this context, "interpretability" can mean different aspect of the model, as already cited from the paper "The Mythos of Model Interpretability".

- We may need our model be **transparent**: that is, understand how it uses the features and which of them could consider more important. Ranking or scoring the features (such as with permutation importance) can help in interpret the model inner workings behind one of its output.
- A model should be as **decomposable** as possible, such that It can be broken down into understandable components. Meaningful examples are linear models, where the coefficient of the features can be interpreted as their importance. In other words, consider the following linear model, which can be used in the context of the Breast Cancer predictions.

  $$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n + \epsilon.$$

  In this case, y is the label prediction, x_i all the features extracted from the FNA images, β_i the coefficients indicating the features importance and ϵ the error term.
- Especially in more complex model, being "interpretable" can be represented by a form of **post-hoc interpretability**: that is, using method to explain model's predictions after its training process. As we already described in the first part of the assignment, this type of interpretation can be of various type: from a local interpretation for a single instance (by using LIME technique) to explanation via examples (by providing simialr past clinical cases).

Even though It's important to maximize the interpretability in context like this, we also need to take into account advantages and disadvantages of such an approach.

- The main benefits are represented by more ethical consideration by using the model, alongside an increased trust by the users and more safety (an error in healthcare can have serious consequences).
- On the other hand, focusing too much on making the model interpretable may create drawbacks on its predictive performance. For example, over-simplifying a model to better understand can lead to not being able to campture the underlying patterns in the data; this alongside an increased

computational cost, depending on the technique used to make a model more explainable.

All in all, as mentioned before, interpretability represents an crucial concept in Machine Learning for health care and diagnosis; but, at the same time, it's also necessary to find a good trade-offs with accuracy to let the model being both explainable but also reliable.

References

[1] Zachary C. Lipton. The mythos of model interpretability, 2017.

[2] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you?": Explaining the predictions of any classifier, 2016.

[3] William H. Wolberg, W.Nick Street, and O.L. Mangasarian. Machine learning techniques to diagnose breast cancer from image-processed nuclear features of fine needle aspirates. *Cancer Letters*, 77(2):163–171, 1994. Computer applications for early detection and staging of cancer.