Universidad Autónoma de Madrid

Escuela Politécnica Superior

23|24

Bachelor thesis

Comparative Analysis of Classifiers for Breast Cancer Detection with Visualizations

Iván Sotillo del Horno

Campus Internacional excelencia UAM CSIC

UNIVERSIDAD AUTÓNOMA DE MADRID
ESCUELA POLITÉCNICA SUPERIOR

Bachelor as Ingeniería Informática (modalidad bilingüe)

BACHELOR THESIS

Comparative Analysis of Classifiers for Breast Cancer Detection with Visualizations

Author: Iván Sotillo del Horno
Advisor: Alejandro Bellogín Kouki

mayo 2024

Francisco Tomás y Valiente, n^o 1
Madrid, 28049
Spain

Iván Sotillo del Horno
Comparative Analysis of Classifiers for Breast Cancer Detection with Visualizations

Iván Sotillo del Horno

PRINTED IN SPAIN

A mi madre y a mi abuela, cuya lucha contra el cáncer de mama me ha inspirado a realizar este trabajo.

Resumen

Este Trabajo de Fin de Grado presenta un análisis comparativo de clasificadores para la detección del cáncer de mama y el uso de Inteligencia Artificial Explicable (XAI) para interpretar los resultados. En la fase inicial se realizará la construcción y optimización de los modelos de clasificación, estos clasificadores analizarán los resultados de las biopsias de aguja fina y clasificarán las muestras como benignas o malignas.

Posteriormente, se realiza una comparación de rendimiento comparando métricas como la puntuación F1 o la *recall*. El objetivo es identificar el mejor clasificador de acuerdo a nuestras métricas. Una vez encontrado el mejor modelo, nos adentramos más en él para entender cómo funciona. Para esto, utilizaremos SHAP (SHapley Additive exPlanations), un método de XAI que nos permite ver la importancia de cada característica y cómo contribuyen a la decisión final del modelo. Esto nos permitirá no solo clasificar las muestras, sino también entender por qué el modelo ha tomado esa decisión, lo que puede ser un avance en la comprensión de los modelos de Inteligencia Artificial para fines médicos.

Por último, para mostrar todos los resultados de los clasificadores y el análisis XAI, se desarrolla una aplicación web. Esta aplicación contendrá las métricas resultantes de los clasificadores, así como ejemplos de cómo cada clasificador tomó algunas de sus decisiones y cuáles fueron las características más importantes para la decisión del clasificador.

Palabras clave

Detección de Cáncer de Mama, Clasificadores, Análisis Comparativo, Interpretabilidad, SHAP, Inteligencia Artificial Explicable, Visualización

Abstract

This Bachelor thesis presents a comparative analysis of base and ensemble classifiers for breast cancer detection. The classifiers are used to analyze the results from fine needle aspiration biopsies and classify the samples as benign or malignant. The use of Artificial Intelligence (AI) is crucial in this process. A specific type of AI, known as eXplainable AI (XAI), is employed to interpret the results. The initial phase involves constructing and optimizing the classifier models, these classifiers will analyze the results from fine needle biopsy aspirations and classify the samples as benign or malignant.

Following this, a performance comparison is conducted comparing metrics such as the F1 score or the recall. The aim is to identify the best classifier regarding our metrics. Once the best classifier model is found, we dive deeper into it to understand how it works. For this, we will use SHAP (SHapley Additive exPlanations), a method of XAI that allows us to see the importance of each feature, and how they contribute to the final decision of the model. This will allow us to not only classify the samples but also to understand why the model has made that decision which can be a step forward in understanding AI models for medical purposes.

Lastly, to display all the results from the classifiers and the XAI analysis, a web application is developed. This application will contain the resulting metrics from the classifiers as well as examples of how each classifier made some of its decisions and which features were the most important.

Keywords

Breast Cancer Detection, Classifiers, Comparative Analysis, Interpretability, SHAP, eXplainable AI, Visualization

Table of Contents

Lists

List of algorithms

List of codes

List of equations

List of figures

List of tables

1

Introduction

1.1. Motivation

Breast cancer is the most common cancer type among women [1]; in 2020, there were more than 2.26 million women diagnosed with breast cancer [1], being the second leading cause of death among women in the United States [2]. Early detection is a crucial step for improving survival rates. With the current analysis techniques of Fine Needle Aspiration Biopsy (FNAB), we have a sensitivity (the ability of a test to identify positive cases correctly) of 0.927 [3]. Therefore, there is a need for a more accurate interpretation of those tests.

Machine learning is a branch of artificial intelligence that focuses on developing algorithms that can learn from data and extract patterns from it to be then able to generalize it to unseen data. In this case, we care about classifiers, whose potential is in the ability to learn from a dataset and then on unseen data being able to classify it as one class or another; in this case, we will be able to classify as benign or malign the results of a fine needle aspiration.

The potential of classifiers in breast cancer detection is immense. However, the effectiveness of the different classifiers can vary; this is why it is crucial to understand how each classifier works, how to tweak it, and how to make them as precise and effective as possible. Then, once we have the best classifier, we can use Explainable Artificial Intelligence (XAI) techniques to understand how the classifier works. Then, once we have the best classifier, we can use XAI techniques to understand how the classifier works so that other types of users, such as, healthcare professionals, or in general, people who are not educated in computer science or mathematics can understand the results of the classifier and trust it. This is crucial for the adoption of machine learning techniques in the healthcare sector, where the results of the classifiers can have a direct impact on the patient's life.

Finding the best possible classifier for this problem would impact cancer detection tasks, facilitating healthcare professionals in their diagnostic responsibilities and, ultimately, improving patient outcomes.

1.2. Objectives

1.3. Structure of the document

2

State of the art

In this chapter, we are going to see the state of the art of the different techniques used in this project. We will see what the current situation of building and optimizing classifiers is, what the most common techniques currently used in the field of XAI are, and, finally, we will present the state of the art regarding web development.

2.1. Classifers: Basic Concepts

Classification is the process of taking input and assigning it to a class, it is a fundamental task in machine learning. There are different types of classification, but for this work, we are going to focus on binary supervised classification. Binary means that the classification will only be carried out between two classes, in our case, benign and malignant; supervised means that the training process will be performed with labeled data, that is, we will train that classifier with FNAB samples that have been labeled as benign or malignant by a pathologist. While classifiers may use different types of information as input, in this work we will focus on exploiting the features extracted from the images, and leave the other types of information for future work.

A classifier is an algorithm that performs the aforementioned task, it receives an input and assigns it to a class. To train a classifier we must first have a dataset, this dataset will be divided into two parts, the training set, and the test set. The training set will be used to train the classifier, which means that the classifier will learn from this data trying to find patterns that allow it to classify the samples correctly. The test set will be used to evaluate the classifier, this set will be used to see how well the classifier performs with data that it has not seen before, this is done to ensure that the classifier is not overfitting (memorizing the training data) and that it can generalize to new data.

Classifiers can be split into two main groups, base classifiers and ensemble classifiers [4]. Their main difference is that base classifiers are single classifiers, while ensemble classifiers are composed of multiple base classifiers. In this section, we will give an overview of the most common base and ensemble classifiers.

2.1.1. Base Classifiers

Base classifiers are single classifiers; they are the simplest form of classifiers, and they are the building blocks of ensemble classifiers [5]. There are many base classifiers, but for this work, we are going to focus on the most common ones [6].

Logistic Regression

Logistic Regression (LR) is a probabilistic model that uses a logistic function to predict the probability of a binary outcome. The central idea is to find the best-fitting model to describe the relationship between the binary dependent variable and one or more independent variables.

Multilayer Perceptron

A Multilayer Perceptron (MLP) is a neural network composed of perceptrons, a perceptron is a simple model of a biological neuron, it takes several binary inputs and produces a single binary output using a weighted sum of the inputs and an activation function.

Stochastic Gradient Descent

Stochastic Gradient Descent (SGD) is not a classifier itself, but an optimization algorithm used instead of the standard gradient descent algorithm, it is used to minimize the loss function of a classifier. It is in this list because in *Scikit-learn* [7] it can be used as a classifier by indicating the linear model to be used [8].

Support Vector Machine

Support Vector Machine (SVM) is a statistical model that separates data into classes by finding a hyperplane that separates the classes, since there can be multiple hyperplanes that do so, the model tries to find the one that maximizes the margin between the classes.

K-Nearest Neighbors

K-Nearest Neighbors (KNN) is a model that classifies the sample based on the majority of its neighbors, the number of neighbors, k, is a parameter that must be set by the user.

Linear and Quadratic Discriminant Analysis

Two different models, Linear Discriminant Analysis (LDA) and Quadratic Discriminant Analysis (QDA). They are used to find a linear or quadratic decision boundary that separates the classes.

Decision Tree

Decision Tree (DT) is a model that separates the data into classes by asking a series of questions, each question is based on the value of a feature, and the answer to the question determines the next question. It is a tree-like model, where the leaves are the classes.

Gaussian Naive Bayes

Gaussian Naive Bayes (GNB) is a probabilistic model that uses Naive Bayes' theorem with the assumption that the features are independent, given the class. It is called Gaussian because it assumes that the features are normally distributed.

2.1.2. Ensemble Classifiers

Ensemble classifiers are a type of classifier that combines the predictions of multiple base classifiers to improve the accuracy of the predictions. Some of the ensemble classifiers allow the base classifiers to be chosen, which means that we can use the best base classifiers to build the ensemble classifier. In this section, we are going to give an overview of the ensemble classifiers that we are going to use in this project.

Gradient Boosting

Gradient Boosting (GB) is an ensemble machine learning technique that combines the predictions from several models to improve the overall predictive accuracy. It works by first building a weak model (a model that is slightly better than random guessing) and then building a second model that corrects the errors made by the first model and so on [9], we can see a graphical representation of the GB in the Figure 2.1.

AdaBoost

AdaBoost (AB) is similar to GB in the sense that it combines the predictions of multiple models but it does so by giving more weight to the misclassified data points in each iteration.

Random Forest

Random Forest (RF) is based on using DT as base classifiers. It builds multiple DT and combines their predictions, it is called random because it uses a random subset of the features to build the DT [11], this randomness is called bagging, so Random Forest is a bagging classifier that uses DT as base classifiers [12], [**?**]. We can see a graphical representation of the RF in the Figure 2.2.

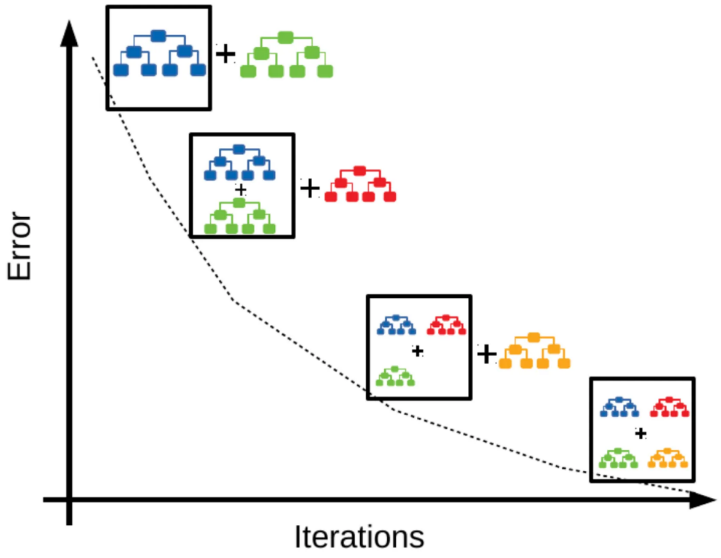


Figure 2.1: Graphical representation of Gradient Boosting [10].

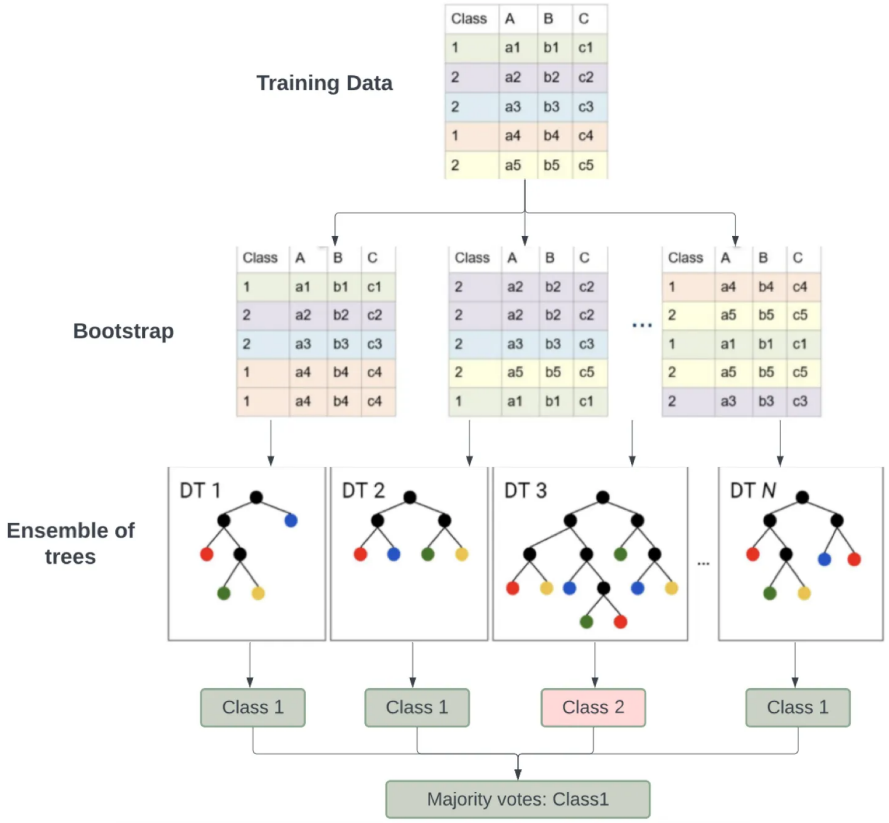


Figure 2.2: Graphical representation of Random Forest [13].

Voting Classifier

A Vote Classifier (VC) as its name suggests, is an ensemble classifier that combines the predictions of multiple classifiers by voting. It can be hard or soft voting, hard voting is the usual voting where the class that gets the most votes is the final prediction, and soft voting is when the average of the probabilities of the classes is used to make the final prediction.

Bagging Classifier

Bagging, also known as Bootstrap aggregation is a technique to build an ensemble classifier by combining multiple base classifiers and combining their predictions by averaging or voting, it was proposed by Breiman [14]. In bagging the base classifiers are built using a random subset of the features [15], we can see a graphical representation of the Bagging Classifier (BC) in the Figure 2.3.

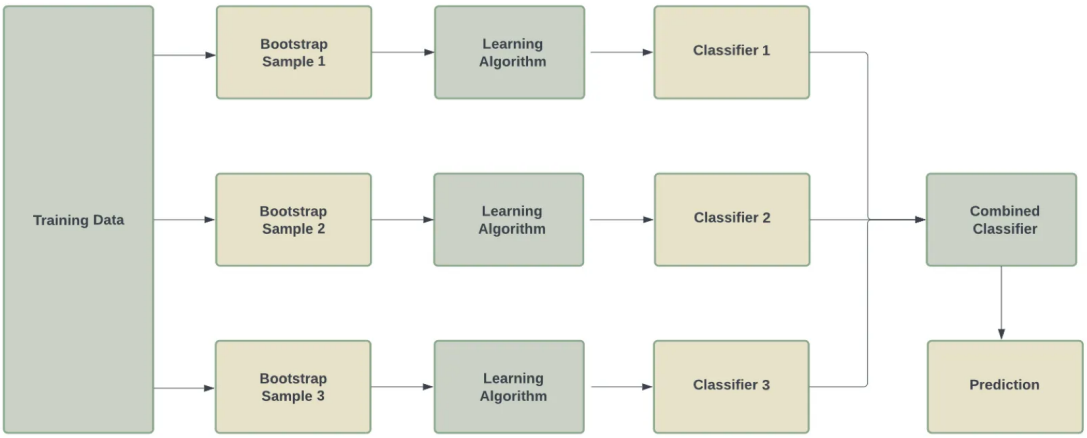


Figure 2.3: Graphical representation of Bagging [16].

Stacking Classifier

A Stacking Classifier (SC) is an ensemble that combines the base classifiers into a meta-classifier, that is, the sample to be classified goes through the base classifiers and then the predictions of the base classifiers are used as features to build the meta-classifier [17]. We can see a graphical representation of the SC in the Figure 2.4.

2.2. Evaluation of Classifiers

Now that we know what a classifier is, how it works, and the different types of classifiers, we need to know which classifier is the best for our problem, for this, we need to use metrics to evaluate the

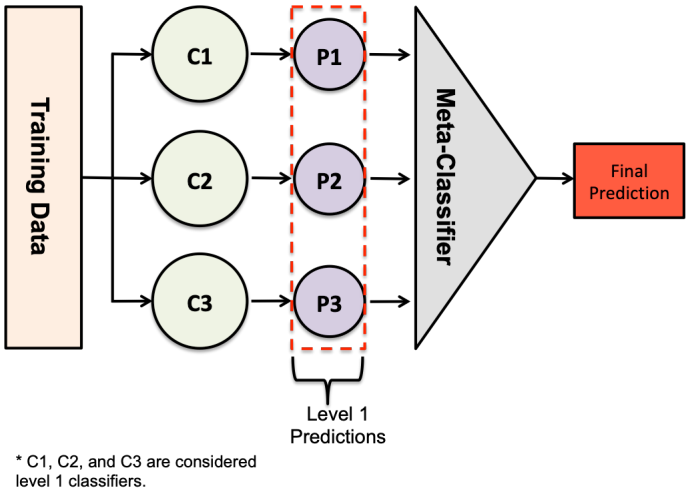


Figure 2.4: Graphical representation of Stacking [18].

performance of the classifiers. Before we describe the metrics we need to understand the values that we are going to use to calculate these metrics, the values are:

- True Positive (TP): The number of instances that are positive and the classifier correctly classified as positive.
- True Negative (TN): The number of instances that are negative and the classifier correctly classified as negative.
- False Positive (FP): The number of instances that are negative but the classifier incorrectly classified as positive.
- False Negative (FN): The number of instances that are positive but the classifier incorrectly classified as negative.

Now that we have our building blocks for the metrics, we can describe the main four metrics [19]:

- **Accuracy**: Proportion of correct predictions over the total number of instances evaluated.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{2.1}$$

- **Recall**: Proportion of TP results among the number of all positive instances.

$$Recall = \frac{TP}{TP + FN} \tag{2.2}$$

- **Precision**: The proportion of TP results among the number of cases that were predicted to be positive.

$$Precision = \frac{TP}{TP + FP} \tag{2.3}$$

- **Specificity**: The proportion of TN results among the number of cases that were actually negative.

$$Specificity = \frac{TN}{TN + FP} \tag{2.4}$$

A confusion matrix is a table that is used to see the performance of the classifier in a more detailed way, as it shows the actual number of TP, TN, FP, and FN. An example is presented in Table 2.1. We can use this table to calculate all the metrics we have seen before. The table is split into four quadrants, the top left quadrant is the TP, the top right quadrant is the FP, the bottom left quadrant is the FN, and the bottom right quadrant is the TN, what we want is to have as many TP and TN (the diagonal) as possible since these are the ones that have been correctly classified.

	Predicted Positive	Predicted Negative
Actual Positive	True Positive	False Negative
Actual Negative	False Positive	True Negative

Table 2.1: Confusion Matrix

The problem with the accuracy metric is that it is not a good metric when the classes are imbalanced, for example, if we have 95 % of the instances of class 1 and 5 % of the instances of class 2, a classifier that always predicts the class 1 will have a 95 % accuracy, but it is not a good classifier. For this reason, we need to use other metrics. For example, we can fix the accuracy metric by using the *Balanced Accuracy*, which is the average of the recall of each class; this metric is better than accuracy when the classes are imbalanced. The formula for the balanced accuracy is shown in Equation 2.5.

$$BalancedAccuracy = \frac{Recall + Specificity}{2} \tag{2.5}$$

With this new accuracy metric, we can see a better performance of the classifier when the classes are imbalanced. We can clearly see the difference between the accuracy and the balanced accuracy with a toy example: if we have a dataset with 100 instances, 97 of them are positive and 3 of them are negative, and we have a classifier that predicted 96 samples as positive and 4 as negative, this is expressed in the Table 2.2. Let us now calculate the accuracy and the balanced accuracy.

	Predicted Positive	Predicted Negative
Actual Positive	94	3
Actual Negative	2	1

Table 2.2: Confusion Matrix Example

- **Accuracy** (Equation 2.1): $\frac{94+1}{94+1+2+1} = \frac{95}{100} = 95{,}00\,\%$
- **Recall** (Equation 2.2): $\frac{94}{94+3} = 96{,}91\,\%$
- **Specificity** (Equation 2.4): $\frac{1}{1+3} = 25{,}00\,\%$
- **Balanced Accuracy** (Equation 2.5): $\frac{0{,}9691+0{,}25}{2} = 60{,}95\,\%$

As we can see, we obtained a 95 % accuracy, but the balanced accuracy is 60.95 %, this is a big difference and represents better the performance of the classifier which misclassified a high percentage of the negative instances.

Apart from these basic metrics we have other metrics that are based on these, for example, we have the *F-Beta Score*, which is the weighted harmonic mean of the precision and recall, depending on the value of β we can give more importance to the precision or the recall. For example, if we want to give more importance to the recall we can use the *F2 Score* (Equation 2.8), which means that the β is 2, and if we want to give more importance to the precision we can use the *F0.5 Score* (Equation 2.9),

which means that the β is 0.5. The general formula for the F-Beta Score is shown in Equation 2.6. This metric can be really useful when we want to give more importance to one of the metrics, for example, if we are predicting a disease, we want to give more importance to the recall since we want to detect as many cases as possible, even if we have some FP, but if we are predicting if a person is going to buy a product, we want to give more importance to the precision, since we want to avoid FP.

$$F_\beta = (1 + \beta^2) \cdot \frac{precision \cdot recall}{(\beta^2 \cdot precision) + recall} \tag{2.6}$$

$$F1 = 2 \cdot \frac{precision \cdot recall}{precision + recall} \tag{2.7}$$

$$F2 = 5 \cdot \frac{precision \cdot recall}{4 \cdot precision + recall} \tag{2.8}$$

$$F0{,}5 = 1{,}25 \cdot \frac{precision \cdot recall}{0{,}25 \cdot precision + recall} \tag{2.9}$$

We also have the *Matthews Correlation Coefficient*, this metric tries to summarize the confusion matrix into a single value, it does not take into account the class imbalance, so when the classes are imbalanced Fβ Score is better [20], but when the classes are balanced the Matthews Correlation Coefficient can be useful, as it measures the correlation between the true and predicted values [21]. The formula for the Matthews Correlation Coefficient is shown in Equation 2.10.

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{2.10}$$

The last metric we are going to mention is the *ROC AUC* it is the area under the Receiver Operating Characteristic (ROC) curve; this curve is a plot of the true positive rate (Recall) against the False Positive rate ($FPR = \frac{FP}{FP+TN}$), the ROC curve is a good way to see the performance of the classifier, the closer the curve is to the top left corner, the better the performance of the classifier. The area under the curve is the ROC AUC, and it is a value between 0 and 1, the closer to 1 the better the performance of the classifier, if the area is 0.5 or the curve is close to the diagonal, the classifier is not better than random guessing. An example of an ROC curve is shown in Figure 2.5, we can see that the curve is somewhat close to the top left corner so it is a good classifier but there is still room for improvement. This metric is really useful when the classes are balanced, if the classes are imbalanced the ROC AUC can be misleading, and thus we will want to use the aforementioned metric Fβ Score.

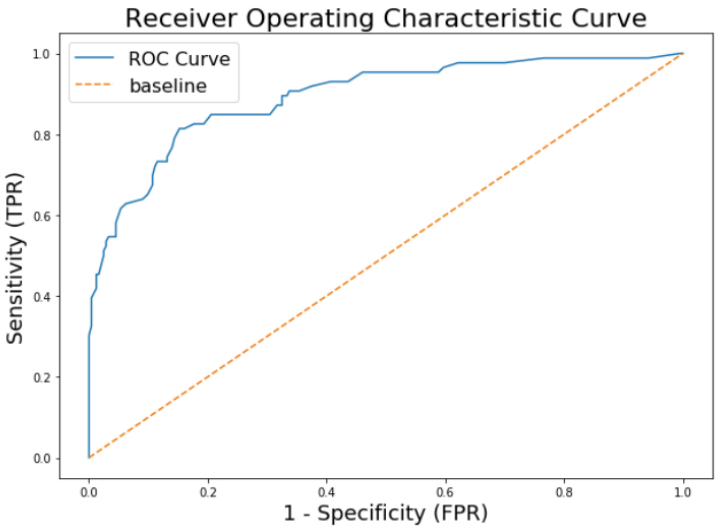


Figure 2.5: Example of a ROC Curve, the y-axis is the True Positive Rate or Recall, and the x-axis is the FP Rate [22].

2.3. Classifier Optimization

We have understood the metrics that are used to evaluate the performance of a classifier, now we can move to the optimization of the classifier. The optimization of a classifier is the process of finding the best hyperparameters for it, the hyperparameters are the parameters that are not learned by the model; for example, the number of trees in a Random Forest, the number of neighbors in a K-Nearest Neighbors, etc. These parameters can heavily influence the performance of the classifier. The two main methods to optimize a classifier are Grid Search and Random Search.

- **Grid Search**: This method is a brute force method, it tries all the possible combinations of the hyperparameters, for example, if we have two hyperparameters, one with 3 possible values and the other with 4 possible values, the Grid Search will try 12 different combinations. This method is very useful when we have a small number of hyperparameters, but it is not efficient when we have a large number of hyperparameters since the number of combinations grows exponentially.
- **Random Search**: This method instead of trying all the possible combinations, tries a random subset of the combinations. The number of combinations to try is a parameter of the method. Since it doesn't try all the combinations, it is faster than the Grid Search, but it is not guaranteed to find the best combination of hyperparameters.

Both of these methods use Cross Validation (CV), a technique that is used to evaluate the performance of the classifier without using the test set. The CV technique splits the dataset into k folds, then leaves one fold out and trains the classifier with the remaining $k-1$ folds, then it evaluates the classifier with the left out fold and repeats this process changing the fold that is left out, this process is repeated k times, and the average of the results is the final result. This technique is used to avoid overfitting the classifier to the training set. An example of a CV with 5 folds is shown in Figure 2.6.

As we saw, both methods have their advantages and disadvantages, the Grid Search is guaranteed

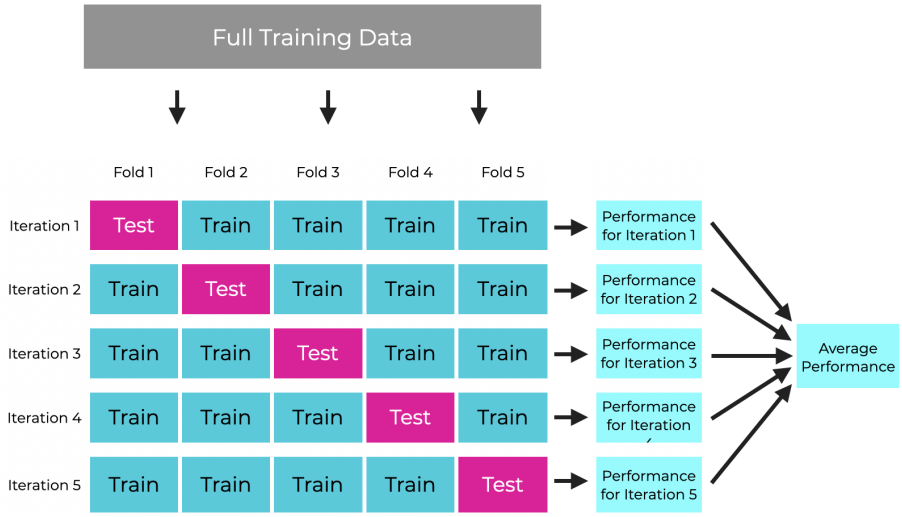


Figure 2.6: Graphical representation of a 5 fold CV [23].

to find the best combination of hyperparameters, but it is not efficient when we have a large number of hyperparameters since it can take hours and even days to finish, the Random Search is faster, but it is not guaranteed to find the best combination of hyperparameters. Researchers have found that Random Search is more efficient than Grid Search, if both methods have the same number of iterations, Random Search will find a better combination of hyperparameters [24]. Knowing this we can first use Random Search to find a boundary of the best hyperparameters and then use Grid Search to find the best combination of hyperparameters.

2.4. Explainable AI

Classifiers are a really helpful tool, but they are a black box, that is, they are fed with an input and provide an output, but we do not know why they made a certain decision. This is a problem in the medical field since we need to know why a certain decision was made; for example, if a patient has cancer, we need to know which features determine that the patient has cancer, and which ones determine that the patient does not have cancer. This is where XAI comes into play, XAI is a branch of Artificial Intelligence (AI) that is focused on making Machine Learning (ML) models transparent, explaining why a certain decision was made and allowing doctors to understand the decision [25].

SHAP

SHapley Additive exPlanations (SHAP) [26] is a model-agnostic approach to XAI, that is, it can be used with any model since it only uses the input and output. In short, how SHAP works is that it uses the Shapley values from cooperative game theory to explain the output of any machine learning model,

a concept named after Lloyd Shapley who introduced the idea that got him the Nobel Prize in Economic Sciences in 2012 [27].

The Shapley value is a concept from cooperative game theory. It is a way to fairly distribute the loot obtained by a coalition of players. The difference is that in SHAP instead of players we have features and instead of loot, we have the predicted output as a probability. The Shapley value is the average contribution of a feature to the prediction, and it is calculated by averaging the marginal contributions of a feature to the prediction over all possible orderings of the features. The marginal contribution of a feature to the prediction is the difference between the prediction with the feature and the prediction without the feature. This is done for all possible subsets of features, and then the Shapley value is the average of all these marginal contributions.

To better understand how these values are obtained let's give an example. To change topics instead of breast cancer, let's say we want to predict the price of a car, and for this, we will have three features: the year of the car, the horsepowers, and the brand. Now, we want to know how much each feature contributes to the price of the car. To do this, first we have to obtain the **powerset** of the features, which is the set of all possible subsets of the features, going from $f = 0$ to $f = 3$, where f is the number of features in the subset, we can see the power set in Figure 2.7.

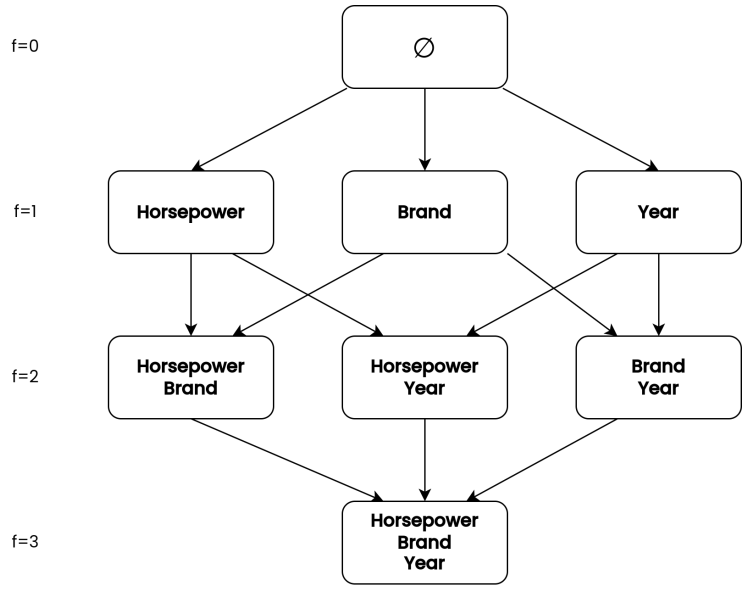


Figure 2.7: Powerset with the features of the car example.

Now that we have all the possible subsets of the features, the next step is to train a model for each subset, the number of models is given by the cardinality of the power set, 2^f where f is the number of features, in this case $2^3 = 8$. All the models are equal, same architecture, same hyperparameters, the only difference is the number of features used to train the model. Now imagine that we have trained all the models, and we have the predictions for all the models, we can see the predictions in Figure 2.8.

All the nodes that are connected differ by just one feature, knowing this we can see that, for example,

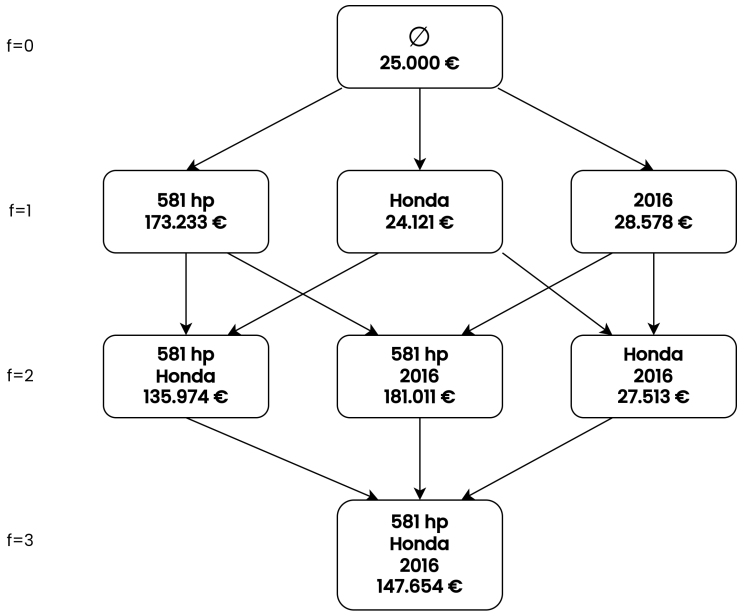


Figure 2.8: Example predictions for the powerset of the car features.

the base model, the one with no features, has a prediction of 25.000 euros, and the model with the year of the car has a prediction of 28.578 euros, this means that the year of the car contributes 3.578 euros to the price of the car, and for example, the brand, Honda, contributes -879 euros. This is called marginal contribution and the formula is shown in Equation 2.11.

$$MC_{Brand,\{Brand\}(x)} = Prediction_{\{Brand\}(x)} - Prediction_{\emptyset(x)} = 24{,}121 - 25{,}000 = -879 \quad (2.11)$$

Each edge also has a weight, the important thing to note is that the weights are the same for all the edges in a row and that the weights in a row add up to 1. The way to obtain this weight then, is by knowing the number of edges in a row and then dividing 1 by the number of edges. To obtain the number of edges we can use basic combinatorics. First, the number of nodes in a row is given by the binomial coefficient, which is given by $\binom{F}{f}$, where F is the number of features, and f is the number of features in the subset just like in the power set. Then the number of edges is given by $f \cdot \binom{F}{f}$, that is, to get the weights of the nodes that connect the base model ($f = 0$) to the models with one feature ($f = 1$) we have to divide 1 by the number of edges, which is $1 \cdot \binom{3}{1} = 3$, and then we get the weights 1/3, 1/3, and 1/3. This is done for all the rows, and we obtain the weights shown in Table 2.3.

f	$\binom{F}{f}$	$f \cdot \binom{F}{f}$	Weights
0	1	0	0
1	3	3	1/3
2	3	6	1/6
3	1	3	1/3

Table 2.3: Weights for the example

Now that we have the weights, we can do the weighted sum of the marginal contributions, and we obtain the Shapley value, which is the importance of the feature for the prediction. So if we do the weighted sum for the brand, we obtain the result shown in Equation 2.12.

$$\begin{aligned} SV_{Brand}(x) = & \frac{1}{3} \cdot MC_{Brand,\{Brand\}(x)} + \\ & \frac{1}{6} \cdot MC_{Brand,\{Brand,Year\}(x)} + \\ & \frac{1}{6} \cdot MC_{Brand,\{Brand,Horsepowers\}(x)} + \\ & \frac{1}{3} \cdot MC_{Brand,\{Brand,Year,Horsepowers\}(x)} \\ = & \frac{1}{3} \cdot -879 + \frac{1}{6} \cdot -37259 + \frac{1}{6} \cdot -1075 + \frac{1}{3} \cdot -33357 \\ = & -17801 \end{aligned} \tag{2.12}$$

With this, we can see that the brand contributes -17801 euros to the price of the car, and we can do the same for the other features, we obtain the Shapley values for all the features and see how much each feature contributes to the price of the car; we summarize the results in Table 2.4. If we add all the Shapley values we obtain the difference between the base prediction with no features, a lot of times represented as the mean of the target $E(y)$, and the prediction with all the features, which is the prediction of the model, $f(x)$.

Feature	Shapley Value
Year	6948
Horsepowers	133507
Brand	-17801
Sum	**122654**

Table 2.4: Shapley values for the example

Then we have that the general formula for the Shapley value is given by Equation 2.13. Now we understand how the Shapley values are obtained and what they are, and we can see why SHAP is model-agnostic since it does not depend on the model.

$$SV_{feature}(x) = \sum_{set:feature \in set} \frac{|set|}{\binom{F}{|set|}} \cdot (prediction_{set} - prediction_{set \setminus feature}) \tag{2.13}$$

2.5. Web Application Development

To display all the results from the classifiers and the XAI analysis, a web application will be developed. To build the classifiers and the XAI analysis the programming language Python will be used, and

thus for the web application we will continue with this language. This opens the possibility of using Flask or Django as the web framework. Django and Flask both are free and open-source web frameworks for Python, and both are very good choices, but they have different use cases.

Django allows the user to build web applications quickly, it is a high-level framework that includes a lot of features, it is really easy to use a database or validate forms for example, it takes care of a lot of things for the user, and leaves the developer to focus on the application itself. This also means that Django is not that flexible and can be a bit heavy.

Flask, on the other hand, is a micro-framework, it is really lightweight and flexible, and instead of including a lot of features, it allows the user to add the features that are needed, this means that the developer has to take care of a lot of things that Django takes care of, but it also means that the developer has more control over the application. By default, Flask does not include a database or form validation, but it is possible to add these features with libraries [28].

Taking into account that the focus of the project is not to build a web application but to build the classifiers and the XAI analysis, and that we will need to build a web application that will display the results from the classifiers and the XAI analysis, which means that we will need to use a database to store the results, we will use Django as the web framework. This will allow us to build the web application quickly and focus on the classifiers and the XAI analysis.

After choosing the web framework we will need to choose a database, the database will be used to store the different plots and metrics. Django supports a lot of databases, but the most popular ones are PostgreSQL, MySQL, and SQLite. SQLite is the default database for Django, it is a lightweight database that is perfect for development, but it is not recommended for production, so we will not use this one (except for development or local execution). MySQL and PostgreSQL are both good choices. In this project, PostgreSQL will be used since we have worked with it before and it is a robust database perfect for our described scenario.

The backend is already chosen, now we need to choose the frontend. For this, we will go with the standard choice: HTML for the structure, CSS for the style, and JavaScript for the behavior.

To deploy the web application we will use Render [1], a cloud platform that will allow us to deploy the web application for free. Even though the free accounts have some limitations, for this project such settings will be enough. For the database hosting, Neon [2] will be used, since it also has a free tier that will be enough for this project. Lastly, the application will also be able to run locally to not depend on any cloud platform.

[1] TODO
[2] TODO

3

Design and Implementation

In this chapter the design of the project will be shown, indicating the structure of the project, its requirements, and the implementation of the different parts of the project. Additionally, the implementation of each part will be explained.

3.1. Project Structure

The structure of the project is divided into four main modules corresponding to the main stages of the project: dataset, classifiers, XAI, and web application. The dataset module is responsible for loading the dataset and splitting it into training and testing set. Then in the classifiers module, we do the Exploratory Data Analysis (EDA), data preprocessing, and building and optimizing the classifiers. The XAI module is responsible for the implementation of the SHAP algorithm and the analysis of the results. Finally, the web application module is responsible for the development of the web application. The structure of the project is shown in Figure 3.1. Now we will describe each module in more detail.

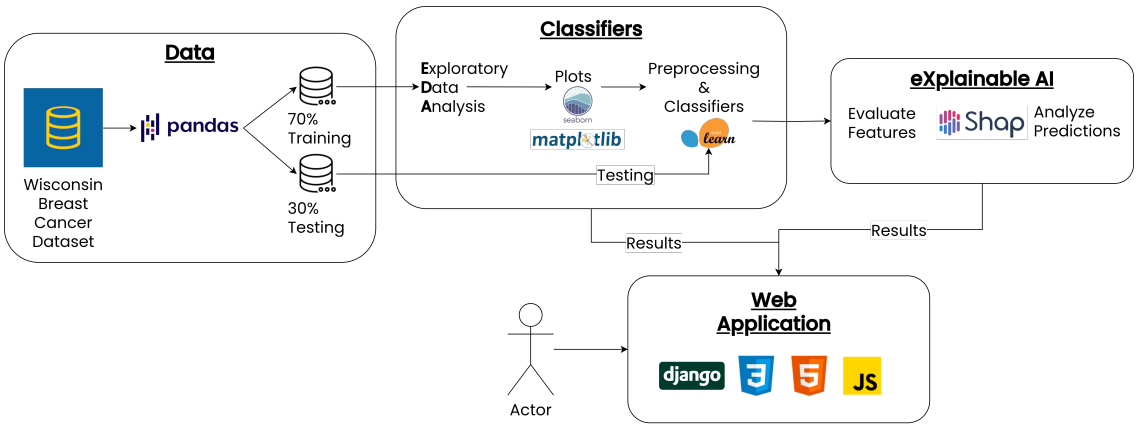


Figure 3.1: Structure of the project.

1. **Dataset module**: This module is responsible for loading the dataset and splitting it into a training set and a test set. The dataset used in this project is the Breast Cancer Wisconsin dataset [29], which contains the parameters of the cell nuclei of the sample obtained by a FNAB of breast masses, an

example of a FNAB is shown in Figure 3.2. Before proceeding to look at the data, we need to perform a testing and training split, this is to avoid data leakage, which is when the model learns from the testing data causing it to overfit instead of generalize [30]. To perform the split we will use the *train_test_split* method from the *Scikit-learn* library [7]. The percentage of the data that will be used for testing will be 30 % and 70 % for training, this is because 80-20 and 70-30 are the most common splits used [31], we opted for the latter to have a larger test set to evaluate if the model generalizes well. From now on, only the training set will be used for the Exploratory Data Analysis, data preprocessing, and building and optimizing the classifiers. The test set will *only* be used to evaluate the classifiers. This ensures that we avoid data leakage.

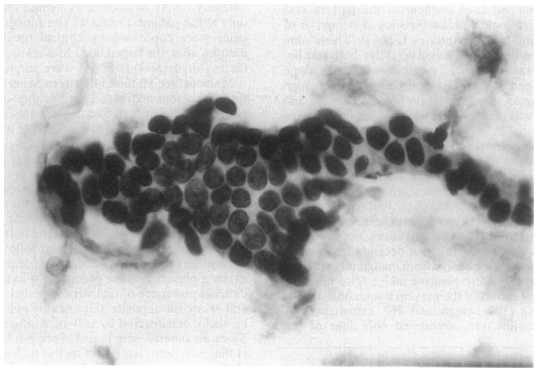


Figure 3.2: Fine Needle Aspiration Biopsy [32].

2. **Classifiers module**: This module is responsible for the EDA, data preprocessing, and building and optimizing the classifiers. The EDA is carried out using *Matplotlib* [33] and *Seaborn* [34] to obtain descriptive and visual statistics of the dataset. The data preprocessing is carried out using *Scikit-learn* to scale and normalize the data and to apply Principal Component Analysis (PCA), this is made using pipelines that facilitate the process of building and optimizing the classifiers and their reproducibility. The building and optimization of the classifiers are carried out using *Scikit-learn* using the training set to build and optimize the classifiers. Once the classifiers are built and optimized, they are evaluated using the test set.

3. **XAI module**: This module is responsible for the implementation of the SHAP algorithm and the analysis of the results. This is implemented using the *Shap* library, which is a Python library that allows us to calculate the SHAP values of the classifiers, this values will be used to understand the importance of the features for each classifier and to understand how a prediction was made. This is what makes this project more interpretable than just using classifiers since this will allow a doctor to understand why a prediction was made.

4. **Web application module**: This module is responsible for the development of the web application. This website was made using *Django* as the backend, *HTML*, *CSS*, and *JavaScript* as the frontend. The website is mainly a way of visualizing the results of the project, it allows the user to see a list of all classifiers and their metrics, which features are the most important for that classifier, and how some

predictions were made. It also allows the user to compare two classifiers and see the differences in their metrics and the importance of the features.

3.2. Requirements

In this section, we will define the requirements of the project. We will divide them into functional (what the system should do) and non-functional (how the system should do it) requirements.

3.2.1. Functional Requirements

Dataset

- **FR-1.-** The application should allow the use of the Wisconsin breast cancer dataset or one with the same features.
- **FR-2.-** The dataset should be split into a training and testing dataset using a 70-30 ratio.

Classifiers

- **FR-3.-** The application should employ the features provided in the dataset for classifier comparison and XAI.
- **FR-4.-** The application should work with multiple classifiers for breast cancer detection.
- **FR-5.-** The application should preprocess the dataset before feeding it to the classifiers. This may include steps such as normalization, and feature selection, focusing on reproducibility and repeatability.
- **FR-6.-** The application should use a test dataset for classifier comparison.
- **FR-7.-** The application should use evaluation metrics to determine the effectiveness of the classifiers taking into account the nature of the problem.

Explainable Artificial Intelligence

- **FR-8.-** The application should show which features are the most important for a classifier.
- **FR-9.-** The application should show how sure the classifier was of its prediction.
- **FR-10.-** The application should show which features determined if the sample was benign or malignant.

Web Application

- **FR-11.-** The web application should have a home page with a slideshow with information.
- **FR-12.-** The web application should display a list with all the classifiers.
- **FR-13.-** The web application should show a details page about each classifier, showing its metrics, most important features, and examples of decisions made by the classifier.
- **FR-14.-** The web application should compare any two classifiers showing their metrics and most important features.

3.2.2. Non-Functional Requirements

Classifiers

- **NFR-1.-** The programming language should be Python.
- **NFR-2.-** The classifiers should be created and trained in a Jupiter Notebook.
- **NFR-3.-** The Pandas library should be used for data manipulation.
- **NFR-4.-** The Scikit-learn library should be used for the classifiers.
- **NFR-5.-** The Matplotlib and Seaborn libraries should be used for the plots.
- **NFR-6.-** The Shap library should be used to understand the feature importance of the models and how a sample decision was made.
- **NFR-7.-** Pipelines should be used for reproducibility and clarity of the code.

Web Application

- **NFR-8.-** The web application should be designed for desktop usage.
- **NFR-9.-** The web application should have a pink color palette and a modern design.
- **NFR-10.-** The web application should be implemented in Django.
- **NFR-11.-** The web application should have an intuitive interface.

3.3. Exploratory Data Analysis

As it has been previously mentioned the dataset used in this project is the Breast Cancer Wisconsin dataset [29]. This dataset is composed of 569 samples of breast masses obtained by FNAB, then these samples were parametrized into ten different features, meaning that now we will work with numbers instead of images. Then, each of the ten features is divided into three new features: the mean, standard error, and the average of the three worst values, ending up with a total of 30 features. A graphical representation of how the different features are divided is shown in Figure 3.3.

Before using these values in the classifier we first need to understand the dataset, this is done using EDA techniques. This involves, first, obtaining a general understanding of the dataset, then visualizing the dataset, and finally understanding the relationships between the features. All of this is done only using the training set to avoid data leakage.

3.3.1. Descriptive Statistics

To have a general understanding of the dataset we can use the *pandas* library to load the dataset and then use the *.describe()* method to see an overview of the features in the dataset, their mean, standard deviation, minimum, and maximum values; and then use the *.skew()* method to see the skewness of the features. With this, we would have the needed descriptive statistics of the dataset.

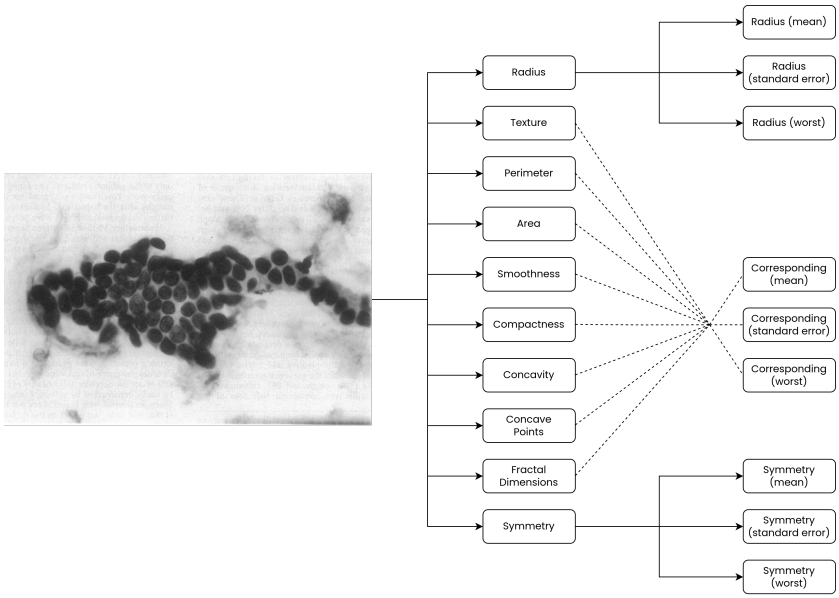


Figure 3.3: Features of the Breast Cancer Wisconsin dataset, FNAB sample picture taken from [32], the rest of the diagram is original.

3.3.2. Data Visualization

After seeing a statistical overview of the dataset, we can now dive into visualizing the dataset. This is done using the *matplotlib* and *seaborn* libraries. With these libraries, we will be able to see more visually the distribution of the target variable, the distribution of the features, and the relationships between the features. To achieve this, we will use histograms, correlation matrices, pair, density, and violin plots among others. This will give us a better understanding of the dataset and the relationships between the features which will allow us to make more informed decisions when preprocessing the dataset and building the classifiers.

3.4. Data Preprocessing

Once we have done the EDA and we know how the dataset is structured, how the different features are distributed, and the relationships between the features, we can start preprocessing the dataset. To achieve this we will standardize the features and then reduce the dimensionality of the features by using PCA, this will result in a dataset that is easier to work with and that will allow us to build better classifiers [35]. To facilitate this process we will use pipelines, which will allow us to easily build and optimize the classifiers and to ensure reproducibility and prevent data leakage [36], this means that we will not directly have to modify the dataset. The code snippet 3.1 shows how the pipeline is built.

Code 3.1: Building the pipeline with the StandardScaler and PCA.

```
pipeline = Pipeline([
    ('scaler', StandardScaler()),
    ('pca', PCA(n_components=0.95)),
    ('clf', Classifier())
])
```

3.4.1. Scaling and Normalization

This is done using the *Scikit-learn* library. The first step is to handle the missing values, this is done using the *SimpleImputer* class, which replaces the missing values with the mean of the feature. Then we scale the features using the *StandardScaler* class, which scales the features to have a mean of 0 and a standard deviation of 1, this is done like the Equation 3.1, where μ is the mean of all the samples of the feature, σ is the standard deviation, and x is the value we want to standardize. We can see in Figure 3.4 how the distribution of the feature stays the same but the mean is 0 and the standard deviation is 1 after standardization.

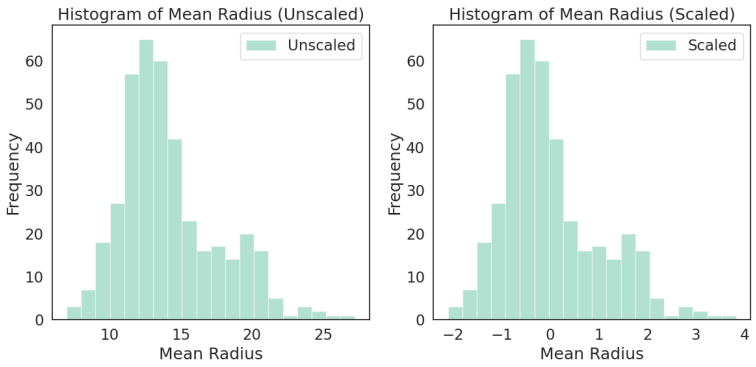


Figure 3.4: Standardization before and after of mean radius.

$$x' = \frac{x - \mu}{\sigma} \tag{3.1}$$

3.4.2. Principal Component Analysis

Lastly, we select the most important variables using the PCA. What this does is that it takes the features and transforms them into a new set of features that are linear combinations of the original features trying to maximize the variance of the new features while the correlation between the new features is 0, this is done using orthogonal transformations [37]. In our case instead of specifying the number of components we want to keep, we specify the percentage of variance we want to keep, in this case, we want to keep 95 % of the variance. This is done so that we can reduce the number of features while keeping most of the information. In our case, when we plot the cumulative explained variance we

can see that with 10 features we can keep 95 % of the variance, this is shown in the Figure 3.5.

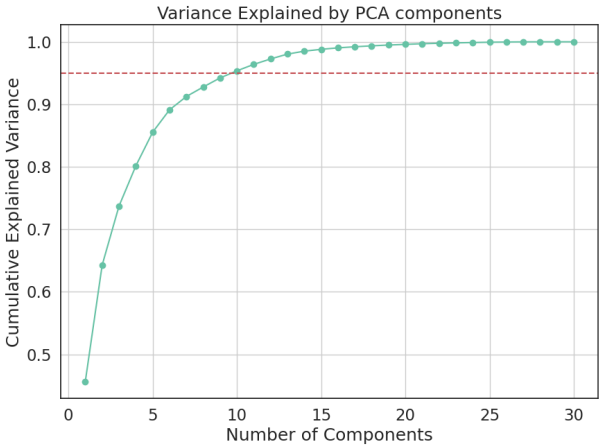


Figure 3.5: Cumulative explained variance when using PCA.

3.5. Building and Optimizing Classifiers

To build the classifiers, *Scikit-learn* was used. This library was chosen because it is the most popular library for building machine learning models in *Python* and it has a wide variety of classifiers and ensemble methods. The classifiers were built in a *Jupyter Notebook* to allow reproducibility and to facilitate the process of building and optimizing the classifiers. The classifiers can be divided into two types, base and ensemble, base classifiers are the simplest classifiers, and ensemble classifiers are a combination of base classifiers, these classifiers are explained in the Section 2.1.

The base classifiers used are:

- **Logistic Regression**
- **Multilayer Perceptron**
- **Support Vector Machine**
- **Decision Tree**
- **Stochastic Gradient Descent**
- **K-Nearest Neighbors**
- **Linear Discriminant Analysis**
- **Gaussian Naive Bayes**
- **Quadratic Discriminant Analysis**

Based on the best results of the base classifiers, the ensemble classifiers were built. Some of these ensemble classifiers allow us to choose the base classifier to use; for these cases, the best base classifiers were used. The ensemble classifiers used are:

- **Random Forest**
- **AdaBoost**

- **Gradient Boosting**
- **Bagging**
- **Voting**
- **Stacking**

3.5.1. Comparison of Classifiers

To decide if a classifier is better than another, we must use some kind of metric. In Section 2.2 we explained the metrics that are most commonly used to evaluate the performance of a classifier. Now we have to choose the appropriate metrics for this problem; the most important thing for this problem is to minimize the number of false negatives, since a false negative means that a patient with cancer was classified as healthy, and this could have fatal consequences. To minimize the number of false negatives, we will use the *Recall*, but the problem of only taking into account the recall is that to maximize it, the model could classify all samples as positive; to avoid this we will also use the *Precision* since this will penalize the model for classifying a sample as positive if it was not. To combine these two metrics we will use the *F1 Score*, which is the harmonic mean of the precision and recall. We will also use the *F2 Score*, which is the weighted harmonic mean of the precision and recall, this will give more importance to the recall since we want to minimize the number of false negatives [38]. So, the evaluation metrics sorted by the importance that we will use are:

- **F1 Score** (Equation 2.7)
- **Recall** (Equation 2.2)
- **F2 Score** (Equation 2.8)
- **Precision** (Equation 2.3)

3.5.2. Optimization of Classifiers

To maximize these metrics in the classifiers, we have to optimize the hyperparameters of the classifiers. The hyperparameters are the parameters that are not learned by the model, and they are set before the training process. These can, for example, be the number of neighbors in the KNN classifier, the number of neurons in the MLP classifier, or the maximum depth of the DT classifier. To optimize these hyperparameters, we used the *RandomizedSearchCV* followed by the *GridSearchCV*. The reason for using the *RandomizedSearchCV* first is that it can achieve similar results to the *GridSearchCV* but with a lower computational cost [24], this will allow us to explore a wider range of hyperparameters and have a range of values to use in the *GridSearchCV*. Then we used the *GridSearchCV* to find the best hyperparameters in the range of values found by the *RandomizedSearchCV*, a better explanation of these methods is given in Section 2.3.

These searches used a 5-fold CV (explained in Section 2.3) using the *F1 Score* as the metric to

optimize. All the *RandomizedSearchCV* and *GridSearchCV* were done using the same random seed, this means that the split of the dataset was the same for all the classifiers, this ends up in a more reliable result and also allows the reproducibility of the results. This process was done to avoid overfitting and to have a more reliable result, of course, as before, this process was only done in the training set, to avoid data leakage.

3.5.3. Choosing the Best Classifier

The previous comparisons were done using the training set with Cross Validation, but to choose the best classifier, we used the test set. This was the first time that used the test set since the beginning of the project, this was done to avoid data leakage and to have a more reliable result. To select the best one we chose the five classifiers that obtained the best metrics in the training set, and then we used the test set to determine which one was the best. The metrics used to compare the classifiers were the same as the ones used in the training set. We only used the five best classifiers because the other classifiers did not perform as well as these five and we did not want to waste computational resources.

3.6. SHAP Implementation

Building and optimizing the classifiers is a crucial step and can be helpful in a clinical situation, but it can only get us so far. We also need to know why the model made a certain decision, for this, there is a branch of AI called Explainable Artificial Intelligence, this branch of AI is focused on making ML models more transparent, allowing clinical practitioners to understand why a certain decision was made [25]. To achieve this we will use the SHapley Additive exPlanations library. The reason for using the SHAP library [26] is that it has been proven to achieve great explanations for ML models in the medical field [39]. A more detailed explanation of SHAP is given in Chapter 2.4.

We will use SHAP to have a global understanding of the importance of the features in the models and a local understanding of how a certain decision was made. This will allow us to understand which features are the most important for a classifier, how sure the classifier was of its prediction, and which features determined if the sample was benign or malignant. This will also be done in a *Jupiter Notebook* to ensure reproducibility.

To be able to calculate the SHAP values, we need to create an *Explainer* object, this object is created with the classifier as a parameter, but in our case, we are not using just the classifier but pipelines, this means, that we will have to find a way to obtain the predicted probability of the pipeline. To do this, we will use the *predict_proba* method of the pipeline, this method will return the predicted probability of the sample being malignant no matter which classifier is used, with this, we will be able to create the *Explainer* object and calculate the SHAP values, all this process is shown in the code snippet A.1.

After we have the SHAP values, we will be able to use them to understand the importance of the features for the classifiers by plotting a graph with the SHAP values of all the samples and also one for each single feature which will show the behavior of that feature, meaning that we will be able to see if the higher the value of the feature the higher the probability of the sample being malignant or benign. The plots, which will be done in Chapter 4 are: cases where the classifier was sure the sample was benign, cases where it was sure it was benign, the closest to the decision boundary, and all the misclassified ones.

In the code snippet A.2 we can see how to code the global feature importance, the behavior of a single feature, and the local explanation of a prediction for a certain classifier, this will be done for all the classifiers, features, and the aforementioned samples.

3.7. Web Application Development

To make the results more accessible and visually appealing, we have built a web application using *Django*. This framework was used because *Python* has been the main programming language used in the project and *Django* was the most familiar framework in that language. For the front-end, we used HyperText Markup Language (HTML), Cascading Style Sheets (CSS), and JavaScript (JS). The web application is hosted in *Render* with the database in *Neon*, which makes it publicly accessible using this link (https://breast-cancer-classification-web.onrender.com/). Since the free hosting of *Render* is not always available, we have also made a video showing the web application in action, this video can be found in this link (https://www.youtube.com/watch?v=ZASpsVNhXPA) or if not the project can be run locally by cloning the code repository (https://github.com/LittleHaku/breast-cancer-classification-web). The home page of the web application is shown in Figure B.1, an example of the comparison of two classifiers that can be prepared using the application is shown in Figure B.2.

4

Experiments and Results

This chapter will showcase the results of the aforementioned experiments. We will start by showing the Exploratory Data Analysis of the dataset, then we will compare the performance of the classifiers, and, finally, we will analyze the SHAP values to understand the importance of the features in the classifiers.

These results can be reproduced by running the code in the *Jupyter Notebook* that is available in the *GitHub* repository https://github.com/LittleHaku/BreastCancer-ClassifierAnalysis.

4.1. Exploratory Data Analysis

After loading our dataset [29] to our *Jupyter Notebook* and importing it from the *Scikit-learn* library [40], we convert it into a *Pandas* [41] DataFrame because it allow us to manipulate the data more easily.

Before proceeding to look at the data, we need to perform a testing and training split, this is to avoid data leakage, which is when the model learns from the testing data causing it to overfit instead of generalize [30]. To perform the split we will use the *train_test_split* method from the *Scikit-learn* library [7]. The percentage of the data that will be used for testing will be 30 % and 70 % for training, this is because 80-20 and 70-30 are the most common splits used [31], we opted for the latter to have a larger test set to evaluate if the model generalizes well.

Now that we have our data loaded and split in a DataFrame, we can begin the EDA. We will divide the process into two parts, the first part will be the descriptive statistics and the second part will be the data visualization.

4.1.1. Descriptive Statistics

The first step in the EDA is to look at the numerical description of the data, this will allow us to have a general idea of the data we are working with. To do this we will use the *describe* method from the *Pandas* library. From this quick statistical summary, we obtain the table available in Table C.1, which

shows the count, mean, standard deviation, minimum, 25 %, 50 %, 75 %, and maximum values of each feature in the dataset, since the dataset has 30 features, we will only show the features included in the Table 4.1; we encourage the reader to look at the full table in the Table C.1.

measure	worst area	mean area	mean fractal dimension
count	398.0	398.0	398.0
mean	876.2	655.9	0.1
std	525.5	333.5	0.1
min	185.2	143.5	0.0
25 %	524.1	432.1	0.1
50 %	688.1	557.4	0.1
75 %	1063.5	782.7	0.1
max	3216.0	2250.0	0.1

Table 4.1: Descriptive statistics of only some features of the dataset to fit the table on the page.

- The median (50 %) is usually smaller than the mean, which would mean that the data is lightly skewed, we will dig into this result in the next section.
- The difference between the 75 % and the maximum value is more than double the number in some cases (worst area 75 % = 1063 and worst area max = 3216), which would mean that there are some outliers in the data, this will also be seen later.
- The range in some values is pretty large, for example in the mean area we go from a min of 143.5 to a max of 2250.0, this means that we will have to scale the data in order to be able to use it with some classifiers.
- The standard deviation is also quite large in some cases, for example, in the worst area, we have a standard deviation of 525.5, this points out that the data is spread out emphasizing the need to scale the data.
- Lastly, some features such as mean fractal dimension have values that go from 0.050 to 0.097, and others like the worst area from 185 to 3216, this means that we will have to scale and normalize the data to be able to use it with some classifiers.

After looking at the description given by the *describe* method in Table 4.1, we can see that the data is quite spread out and that we will need to scale it in order to be able to use it in some classifiers.

It is also important to detect the skewness of the data, since it will tell us if the data is symmetric or not. This is important because some classifiers assume that the data is normally distributed, and if it is not, the classifier will not work as well. We can see the skewness of the features in Table 4.2.

The skewness' values can be interpreted as follows:

- If the skewness is between -0.5 and 0.5, the data is fairly symmetrical.
- If the skewness is above 1 or below -1, the data is highly skewed.
- If the skewness is positive, the data is right-skewed.
- If the skewness is negative, the data is left-skewed.

We will omit the target since it is not a feature but a class. From the values we have obtained in

Mean Radius	Mean Texture	Mean Perimeter	Mean Area	Mean Smoothness	Mean Compactness
0.84	0.71	0.90	1.38	0.51	1.22
Mean Concavity	Mean Concave Points	Mean Symmetry	Mean Fractal Dimension	Radius Error	Texture Error
1.44	1.14	0.84	1.41	1.68	1.78
Perimeter Error	Area Error	Smoothness Error	Compactness Error	Concavity Error	Concave Points Error
1.85	2.15	2.60	1.97	5.59	1.66
Symmetry Error	Fractal Dimension Error	Worst Radius	Worst Texture	Worst Perimeter	Worst Area
2.32	4.12	0.95	0.44	0.97	1.49
Worst Smoothness	Worst Compactness	Worst Concavity	Worst Concave Points	Worst Symmetry	Worst Fractal Dimension
0.41	1.54	1.24	0.46	1.49	1.84
Target					
0.52					

Table 4.2: Skewness of the features in the dataset.

Table 4.2, we can see that all of them are positive, which means that all the features are right-skewed. Some of them are highly skewed with values of up to 5.45 in the case of area error. This unevenness in the distribution is not only a product of the nature of the data, but also because of outliers which we will see in the data visualization section.

4.1.2. Data Visualization

After looking at the descriptive statistics of the data, we can now proceed to visualize the data. The first step is to look at the distribution of the classes, this is important to see if the data is balanced or not, since if the data is not balanced the classifier will have a hard time learning the minority class [42].

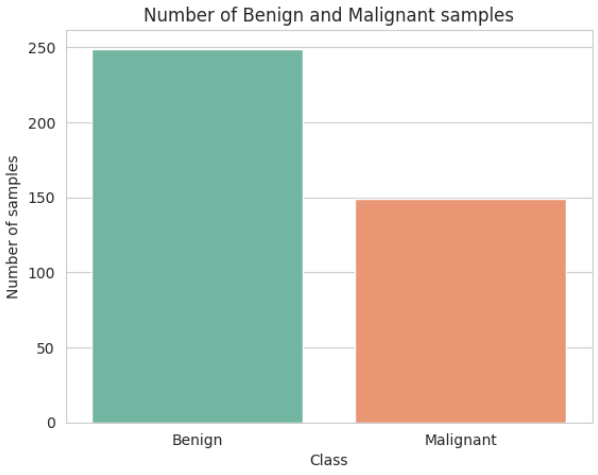


Figure 4.1: Distribution of the classes in the dataset.

We can see the distribution of the classes in Figure 4.1, where the benign class predominates over the malignant class with 62.56 % of the data being benign and 37.44 % being malignant. Although it is not a 50-50 distribution, it is not a heavily unbalanced dataset where the minority class is less than 20 % of the data, so we will not take any action to balance the data, but we will keep this in mind when we evaluate the classifiers since there could be a bias towards the majority class (benign).

The correlation matrix is the next step in the data visualization, this will allow us to see the relationships between the features, this is important because some classifiers assume that the features are independent, and if they are not, the classifier will not work as well. When looking at the correlation matrix we have to understand that values can range from -1 to 1, where -1 means that the features are negatively correlated, 0 means that the features are not correlated, and 1 means that the features are positively correlated. It is important to study the correlation matrix because if the features are highly correlated it can lead to multicollinearity, which can make the model not work as well [43].

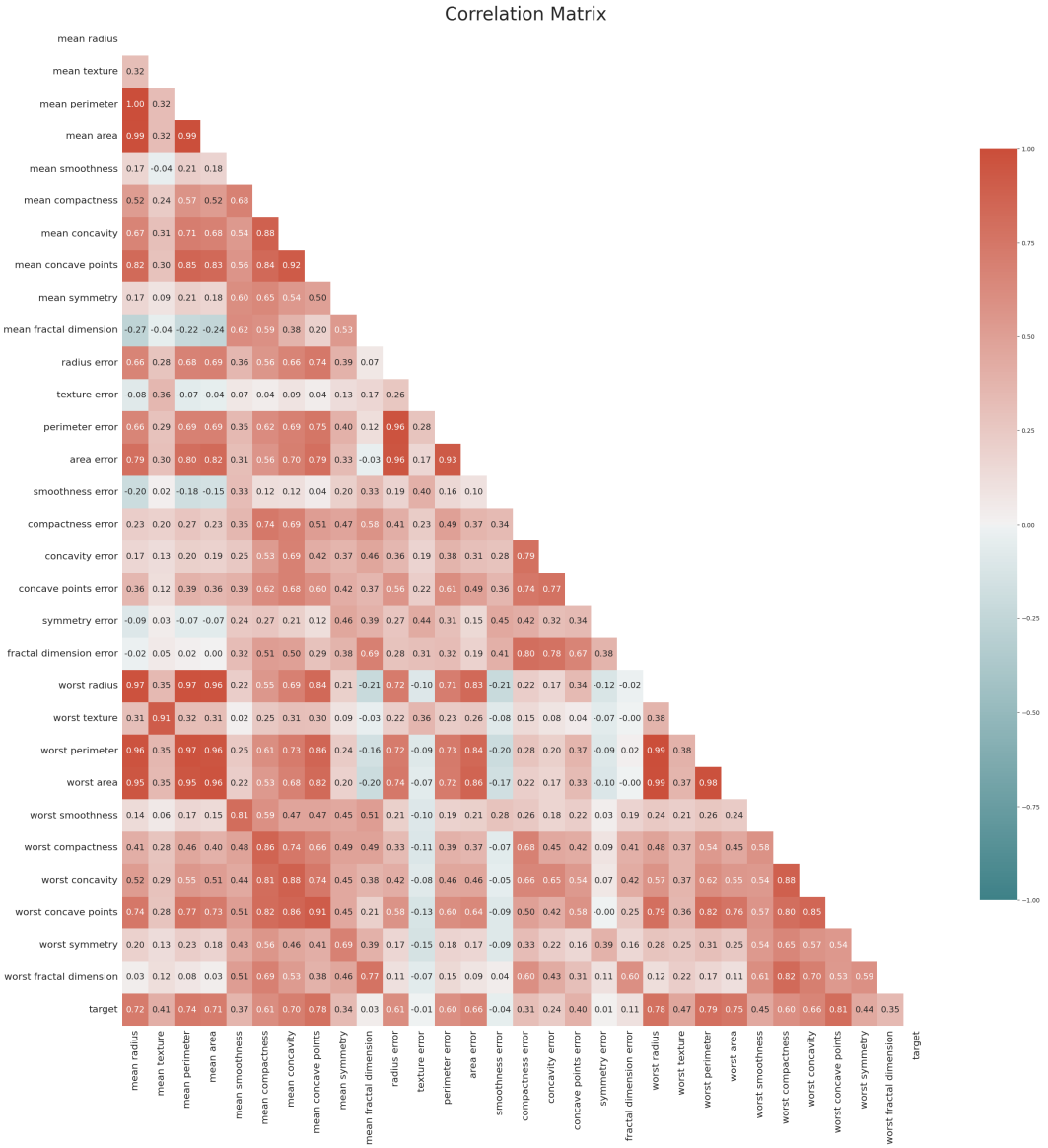


Figure 4.2: Correlation matrix of the features in the dataset.

In Figure 4.2 we can see the correlation matrix of the features in the dataset, where we can see that some features are highly correlated with each other. This is expected since some of the features are mathematically related to each other, for example, the area is related to the radius and the perimeter. There is also a tendency for features that are already correlated in their mean version of the feature to

be correlated in their error and worst version of the feature, e.g., the mean concavity is correlated with the mean compactness (0.88), then the error concavity is correlated with the error compactness (0.79) and the worst concavity is correlated with the worst compactness (0.88).

Furthermore, to understand if the data is separable, we plotted the pair plots of the features. For the sake of space, since we have 30 features we will only show the pair plots of the mean features.

Figure 4.3: Pair plot of the mean features in the dataset.

In Figure 4.3 we can see the pair plot of the mean features in the dataset, where we can see that some features are more separable than others; for example, the mean area and the mean radius are more separable than the mean fractal dimension and the mean symmetry. Along with this, we can also plot the class distribution for each feature to see how separable the classes are.

To see how different the distributions are, we plotted the radius in Figure 4.4 (since we already saw that it was easily separable in the pair plot) and for the fractal dimension in Figure 4.5 (we saw that it was not easily separable in the pair plot). To see the distribution of the other features, we encourage the reader to look at Figures C.1 and C.2 in the Appendix C.

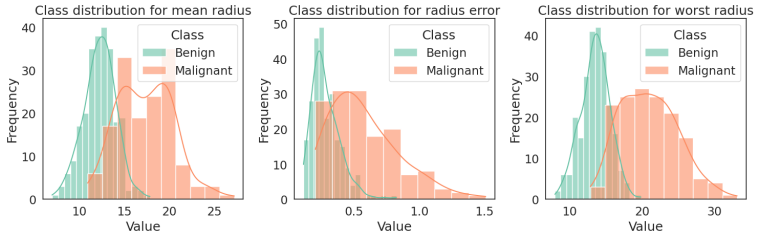


Figure 4.4: Class distribution of the radius in the dataset, from left to right we have the mean, the error, and the worst.

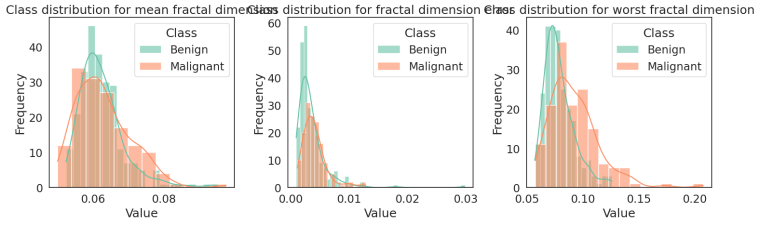


Figure 4.5: Class distribution of the fractal dimension in the dataset, from left to right we have the mean, the error, and the worst.

Once we knew the data was distributed, we proceeded with the preprocessing. As it was explained in Section 3.4, we scaled the data using the *StandardScaler* from the *Scikit-learn* library, and we also performed a PCA to reduce the dimensionality and collinearity of the data retaining 95 % of the variance, which resulted in 10 features made up from a linear combination of the original 30 features with 0 collinearity between any of them. Lastly, this process that we just mentioned was not done to all the training data, but rather implemented in a pipeline so that for each fold in the cross-validation the data is scaled and transformed using only the training data and not the validation one, to avoid data leakage as already discussed.

4.2. Classifier Comparison

This section discusses the results obtained from optimizing the classifiers' hyperparameters to obtain the best possible performance, as it was explained in Section 3.5, we first used *RandomizedSearchCV* to bound the hyperparameters' search space and then used *GridSearchCV* to fine-tune them; the metrics used were: F1-Score, Recall, F2-Score, and Precision.

We first present the results of the base classifiers, followed by the ensemble classifiers that, when possible, were optimized using the same hyperparameters as the base classifiers. Finally, we compare the best base and ensemble classifiers to determine which one is the most suitable for the problem at hand.

4.2.1. Base Classifiers Comparison

The different models were profoundly optimized to obtain the best possible performance. Once all the models were trained, we compared them using the F1-Score, Recall, F2-Score, and Precision. The results are shown in Table 4.3.

Model	F1 Score	Recall	F2 Score	Balanced Accuracy
Logistic Regression	0.9865	0.9846	0.9853	0.9846
Multi-Layer Perceptron	0.9839	0.9839	0.9838	0.9839
Stochastic Gradient Descent	0.9839	0.9826	0.9831	0.9826
Support Vector Machine	0.9839	0.9826	0.9831	0.9826
K-Nearest Neighbors	0.9728	0.9691	0.9705	0.9691
Linear Discriminant Analysis	0.9703	0.9686	0.9691	0.9686
Quadratic Discriminant Analysis	0.9627	0.9652	0.9641	0.9652
Decision Tree	0.9459	0.9424	0.9436	0.9424
Gaussian Naive Bayes	0.9333	0.9364	0.9348	0.9364

Table 4.3: Comparison of all the base classifiers, showing their metrics sorted by the F1-Score, Recall, F2-Score, and Precision.

We can see that the Logistic Regression model is the best-performing model, followed closely by the Multilayer Perceptron, Stochastic Gradient Descent, and Support Vector Machine all of them with a F1-Score of 0.9839 but interestingly the MLP has a higher recall, F2-Score, and balanced accuracy; the reason for this is that as shown in the Equation 2.7 the F1-Score is the harmonic mean of the precision and recall, but the F2-Score gives more weight to the recall, which is why the MLP has a higher F2-Score than the SGD and SVM.

On the lower end of the table, we have the GNB and the DT models, with an F1-Score of 0.9333 and 0.9459, respectively. Although the gap between these and the best-performing models is a significant amount of 0.05, the most important thing is that we managed to optimize the models and we will be able to use them to build ensemble models, especially the DT model which is a model that is heavily used in ensemble models [44].

4.2.2. Ensemble Classifiers Comparison

After comparing the base classifiers, we proceeded to build and compare the ensemble classifiers.

Gradient Boosting, AdaBoost and Random Forest only allowed us to work with Decision Tree as the base classifier; we did not use the hyperparameters from the previous DT model because the power of these models lies in the ensemble of weak learners [45], and thus we optimized the hyperparameters from scratch creating a new search space.

On the other hand, the Vote Classifier, Bagging Classifier, and Stacking Classifier allowed the base classifiers to be chosen freely; so we selected the top five base classifiers from the previous comparison and rearranged them into a list with all the possible combinations, going from all the possible couples of base classifiers up to using all five of them in the ensemble, then we fed these combinations to the ensemble classifiers to determine which one was the best. Apart from the base classifiers, we also optimized the other hyperparameters of the ensemble classifiers. The results of the ensemble classifiers are shown in Table 4.4.

Model	F1 Score	Recall	F2 Score	Balanced Accuracy
Voting Classifier (LR, MLP)	0.9892	0.9867	0.9876	0.9867
Stacking Classifier (LR, MLP)	0.9892	0.9867	0.9876	0.9867
Bagging Classifier (MLP)	0.9866	0.9860	0.9861	0.9860
Gradient Boosting (DT)	0.9758	0.9752	0.9754	0.9752
AdaBoost (DT)	0.9729	0.9691	0.9705	0.9691
Random Forest (DT)	0.9594	0.9577	0.9583	0.9577

Table 4.4: Comparison of all the ensemble classifiers, showing their metrics sorted by the F1-Score, Recall, F2-Score, and Precision.

We can see that the models that allowed us to choose the base classifiers ended up using the LR and/or MLP models, which were the best-performing base classifiers, and as a result, they also ended up being the best-performing ensemble classifiers. The VC and SC models ended up with the same metrics, this might be caused because both of them are using the same base classifiers, so the only difference between them is the way they aggregate the predictions, the VC uses a majority vote, while the SC uses a meta-classifier to make the final prediction (see Section 2.1.2 for more information). The models that used DT as the base classifier ended up with a lower performance, but note that they are all outperforming the single DT model from the previous comparison, which is the main advantage of ensemble models, they can outperform the base classifiers [46].

4.2.3. Choosing the Best Classifier

After computing the metrics for all the base and ensemble classifiers we can see from Table 4.3 and Table 4.4 that the best five classifiers overall are Logistic Regression, Multilayer Perceptron, Vote Classifier, Stacking Classifier, and Bagging Classifier; it is important to note that the metrics obtained until now, as it was explained in Section 3.5.3, were obtained using the training set with Cross Validation. To determine which one is the best classifier, we measured the performance of these five classifiers using the test set. Being this the first time that we used the test set, we can be confident that the results are reliable and that there is no data leakage.

Comparing the results obtained from the training set 4.5 and the test set 4.6 we saw that the only

Model	F1 Score	Recall	F2 Score	Balanced Accuracy
Voting Classifier (LR, MLP)	0.9892	0.9867	0.9876	0.9867
Stacking Classifier (LR, MLP)	0.9892	0.9867	0.9876	0.9867
Bagging Classifier (MLP)	0.9866	0.9860	0.9861	0.9860
Logistic Regression	0.9865	0.9846	0.9853	0.9846
Multi-Layer Perceptron	0.9839	0.9839	0.9838	0.9839

Table 4.5: The five best classifiers measured against the training set, showing their metrics sorted by the F1-Score, Recall, F2-Score, and Precision.

Model	F1 Score	Recall	F2 Score	Balanced Accuracy
Logistic Regression	0.9873	0.9841	0.9854	0.9841
Bagging Classifier (MLP)	0.9810	0.9762	0.9780	0.9762
Multi-Layer Perceptron	0.9810	0.9762	0.9780	0.9762
Voting Classifier (LR, MLP)	0.9745	0.9683	0.9706	0.9683
Stacking Classifier (LR, MLP)	0.9745	0.9683	0.9706	0.9683

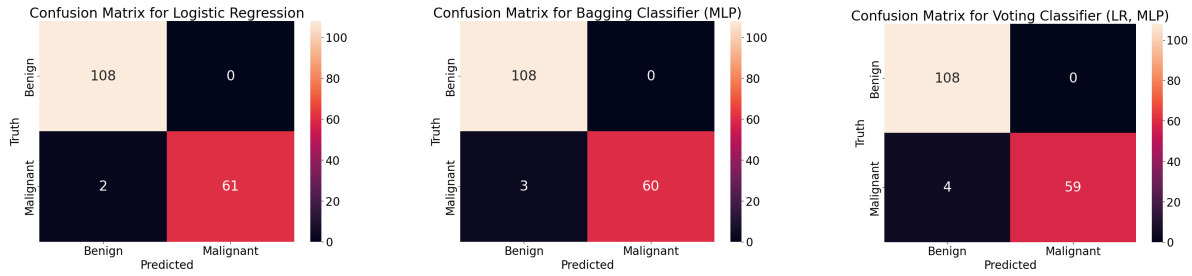
Table 4.6: The five best classifiers measures against the test set, showing their metrics sorted by the F1-Score, Recall, F2-Score, and Precision.

model that performed better in the test set than in the training set was Logistic Regression, the rest diminished their performance, but the most significant decrease was in the Vote Classifier and Stacking Classifier models, which had a decrease of 0.0147 in the F1-Score, and went from being the best-performing models among these five to the worst-performing models. This decrease in performance shows that the Vote Classifier and Stacking Classifier models were overfitting the training set while the Logistic Regression model was generalizing better.

Another interesting thing is to note that BC and MLP models had the same performance in the test set, which makes sense since BC was using MLP as the base classifier. And also, VC and SC had the same metrics again.

To better understand the performance of the models, we can see the confusion matrix of all the models in Figure 4.6. Note that there are only three confusion matrices since the VC and SC, and the BC and MLP models had the same metrics (and the same confusion matrix), so we only show one of them.

From the confusion matrices we can see that the only changes reside in the number of samples that were malignant and got classified as benign, which is exactly what we want to avoid, so we can see that the LR model is the best-performing model with only 2 samples that were malignant and were classified as benign, followed by the BC and MLP models with 3 samples, and finally the VC and SC models with 4 samples.



(a) Logistic Regression

(b) Bagging Classifier (MLP) and Multi-Layer Perceptron

(c) Voting Classifier (LR, MLP) and Stacking Classifier (LR, MLP)

Figure 4.6: Confusion matrices of the five best classifiers.

4.3. SHAP Analysis

4.3.1. Global Interpretability

4.3.2. Local Interpretability

5

Conclusions and Future Work

5.1. Conclusions

5.2. Future Work

Bibliography

[1] WCRF International, "Breast cancer statistics | World Cancer Research Fund International."

[2] American Cancer Society, "Breast Cancer Statistics | How Common Is Breast Cancer?."

[3] Y.-H. Yu, W. Wei, and J.-L. Liu, "Diagnostic value of fine-needle aspiration biopsy for breast mass: a systematic review and meta-analysis," *BMC Cancer*, vol. 12, p. 41, Jan. 2012.

[4] S. S. Yadav, V. J. Kadam, and S. M. Jadhav, "Comparative Analysis of Ensemble Classifier and Single Base Classifier in Medical Disease Diagnosis," in *Communication and Intelligent Systems* (J. C. Bansal, M. K. Gupta, H. Sharma, and B. Agarwal, eds.), (Singapore), pp. 475–489, Springer, 2020.

[5] N. C. Oza and K. Tumer, "Classifier ensembles: Select real-world applications," *Information Fusion*, vol. 9, pp. 4–20, Jan. 2008.

[6] D. Gong, "Top 6 Machine Learning Algorithms for Classification," July 2022.

[7] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[8] "sklearn.linear_model.SGDClassifier."

[9] A. Natekin and A. Knoll, "Gradient boosting machines, a tutorial," *Frontiers in Neurorobotics*, vol. 7, Dec. 2013. Publisher: Frontiers.

[10] Hemashreekilari, "Understanding Gradient Boosting," Sept. 2023.

[11] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, pp. 5–32, Oct. 2001.

[12] G. Biau and E. Scornet, "A random forest guided tour," *TEST*, vol. 25, pp. 197–227, June 2016.

[13] Hemashreekilari, "Understanding Random Forest," Aug. 2023.

[14] L. Breiman, "Bagging predictors," *Machine Learning*, vol. 24, pp. 123–140, Aug. 1996.

[15] Scikit Learn, "Bagging Classifier."

[16] Hemashreekilari, "Understanding Bagging," Aug. 2023.

[17] D. H. Wolpert, "STACKED GENERALIZATION,"

[18] F. Ceballos, "Stacking Classifiers for Higher Predictive Performance," Sept. 2019.

[19] H. M and S. M.N, "A Review on Evaluation Metrics for Data Classification Evaluations," *International Journal of Data Mining & Knowledge Management Process*, vol. 5, pp. 01–11, Mar. 2015.

[20] L. Sisters, "Matthews Correlation Coefficient: when to use it and when to avoid it," May 2020.

[21] B. Shmueli, "Matthews Correlation Coefficient is The Best Classification Metric You've Never Heard Of," May 2020.

[22] A. P. Chazhoor, "ROC curve in machine learning," July 2019.

[23] "Cross Validation, Explained - Sharp Sight," Dec. 2023. Section: machine learning.

[24] J. Bergstra and Y. Bengio, “Random Search for Hyper-Parameter Optimization,”

[25] K. Borys, Y. A. Schmitt, M. Nauta, C. Seifert, N. Krämer, C. M. Friedrich, and F. Nensa, “Explainable AI in medical imaging: An overview for clinical practitioners – Beyond saliency-based XAI approaches,” *European Journal of Radiology*, vol. 162, p. 110786, May 2023.

[26] S. M. Lundberg and S.-I. Lee, “A Unified Approach to Interpreting Model Predictions,” in *Advances in Neural Information Processing Systems* (I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, eds.), vol. 30, Curran Associates, Inc., 2017.

[27] A. E. Roth, *The Shapley value: essays in honor of Lloyd S. Shapley*. Cambridge University Press, Oct. 1988. Google-Books-ID: JK7MKu2A9cIC.

[28] “Foreword — Flask Documentation (2.1.x).”

[29] O. M. William Wolberg, “Breast Cancer Wisconsin (Diagnostic),” 1993.

[30] A. Y.-T. Wang, R. J. Murdock, S. K. Kauwe, A. O. Oliynyk, A. Gurlo, J. Brgoch, K. A. Persson, and T. D. Sparks, “Machine Learning for Materials Scientists: An Introductory Guide toward Best Practices,” *Chemistry of Materials*, vol. 32, pp. 4954–4965, June 2020. Publisher: American Chemical Society.

[31] A. Rácz, D. Bajusz, and K. Héberger, “Effect of Dataset Size and Train/Test Split Ratios in QSAR/QSPR Multiclass Classification,” *Molecules*, vol. 26, p. 1111, Jan. 2021. Number: 4 Publisher: Multidisciplinary Digital Publishing Institute.

[32] J. A. M. Sidey-Gibbons and C. J. Sidey-Gibbons, “Machine learning in medicine: a practical introduction,” *BMC Medical Research Methodology*, vol. 19, pp. 1–18, Dec. 2019. Number: 1 Publisher: BioMed Central.

[33] J. D. Hunter, “Matplotlib: A 2D graphics environment,” *Computing in Science & Engineering*, vol. 9, no. 3, pp. 90–95, 2007. Publisher: IEEE COMPUTER SOC.

[34] M. L. Waskom, “seaborn: statistical data visualization,” *Journal of Open Source Software*, vol. 6, no. 60, p. 3021, 2021. Publisher: The Open Journal.

[35] I. Dinç, M. Sigdel, S. Dinç, M. S. Sigdel, M. L. Pusey, and R. S. Aygün, “Evaluation of Normalization and PCA on the Performance of Classifiers for Protein Crystallization Images,” *Proceedings of IEEE Southeastcon / IEEE Southeastcon. IEEE Southeastcon*, vol. 2014, p. 10.1109/SECON.2014.6950744, Mar. 2014.

[36] K. Zhao, “Pre-Process Data with Pipeline to Prevent Data Leakage during Cross-Validation.,” Sept. 2019.

[37] A. Daffertshofer, C. J. Lamoth, O. G. Meijer, and P. J. Beek, “PCA in studying coordination and variability: a tutorial,” *Clinical Biomechanics*, vol. 19, pp. 415–428, May 2004.

[38] M. Rutecki, “Best techniques and metrics for Imbalanced Dataset.”

[39] R. Massafra, A. Fanizzi, N. Amoroso, S. Bove, M. C. Comes, D. Pomarico, V. Didonna, S. Diotaiuti, L. Galati, F. Giotta, D. La Forgia, A. Latorre, A. Lombardi, A. Nardone, M. I. Pastena, C. M. Ressa, L. Rinaldi, P. Tamborra, A. Zito, A. V. Paradiso, R. Bellotti, and V. Lorusso, “Analyzing breast cancer invasive disease event classification through explainable artificial intelligence,” *Frontiers in*

Medicine, vol. 10, 2023.

[40] Scikit Learn, “sklearn.datasets.load_breast_cancer.”

[41] T. p. d. team, “pandas-dev/pandas: Pandas,” Feb. 2020.

[42] H. He and E. A. Garcia, “Learning from Imbalanced Data,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 21, no. 9, pp. 1263–1284, 2009.

[43] D. A. Belsley, E. Kuh, and R. E. Welsch, *Regression diagnostics: Identifying influential data and sources of collinearity*. John Wiley & Sons, 2005.

[44] R. E. Banfield, L. O. Hall, K. W. Bowyer, and W. Kegelmeyer, “A Comparison of Decision Tree Ensemble Creation Techniques,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 1, pp. 173–180, 2007.

[45] S. Mannor and R. Meir, “On the existence of linear weak learners and applications to boosting,” *Machine Learning*, vol. 48, pp. 219–251, 2002. Publisher: Springer.

[46] T. G. Dietterich, “Ensemble methods in machine learning,” in *International workshop on multiple classifier systems*, pp. 1–15, Springer, 2000.

Acronyms

AB AdaBoost.

AI Artificial Intelligence.

BC Bagging Classifier.

CSS Cascading Style Sheets.

CV Cross Validation.

DT Decision Tree.

EDA Exploratory Data Analysis.

FN False Negative.

FNAB Fine Needle Aspiration Biopsy.

FP False Positive.

GB Gradient Boosting.

GNB Gaussian Naive Bayes.

HTML HyperText Markup Language.

JS JavaScript.

KNN K-Nearest Neighbors.

LDA Linear Discriminant Analysis.

LR Logistic Regression.

ML Machine Learning.

MLP Multilayer Perceptron.

PCA Principal Component Analysis.

QDA Quadratic Discriminant Analysis.

RF Random Forest.

SC Stacking Classifier.

SGD Stochastic Gradient Descent.

SHAP SHapley Additive exPlanations.

SVM Support Vector Machine.

TN True Negative.

TP True Positive.

VC Vote Classifier.

XAI Explainable Artificial Intelligence.

Appendices

A

Code Snippets

Code A.1: Obtaining the predicted probability of the pipeline and calculating the SHAP values

```
def model_predict_proba(data_asarray):
    data_asframe = pd.DataFrame(data_asarray, columns=X_train.columns)
    proba = lr_pipeline.predict_proba(data_asframe)
    return proba[:, 1]

explainer = shap.Explainer(model_predict_proba, X_train)
shap_values = explainer.shap_values(X_test)
explainer_x_test = explainer(X_test)
```

Code A.2: Plotting the SHAP global features, single feature, and local explanation

```
shap.initjs()

# global feature importance
shap.plots.bar(explainer_x_test, max_display=25)
shap.plots.beeswarm(explainer_x_test, max_display=30, alpha=0.75)

# single feature importance
shap.dependence_plot(feature, shap_values, X_test, alpha=0.65, show=False)
ax = plt.gca()
x = ax.collections[0].get_offsets()[:, 0]
y = ax.collections[0].get_offsets()[:, 1]
regression_model = LinearRegression().fit(x.reshape(-1, 1), y)
x_range = np.linspace(x.min(), x.max(), 100)
y_range = regression_model.predict(x_range.reshape(-1, 1))
plt.plot(x_range, y_range, color='red')

# prediction explanation
shap.plots.waterfall(explainers_x_test[prob][idx], show=False)
```

B

Web Application

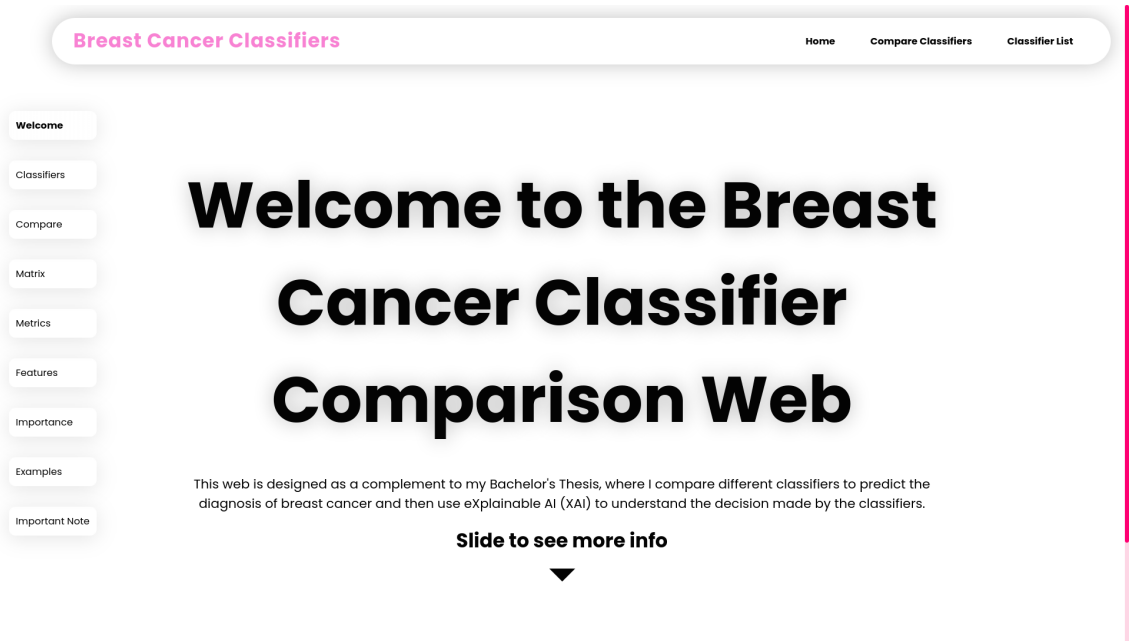


Figure B.1: Home page of the web application.

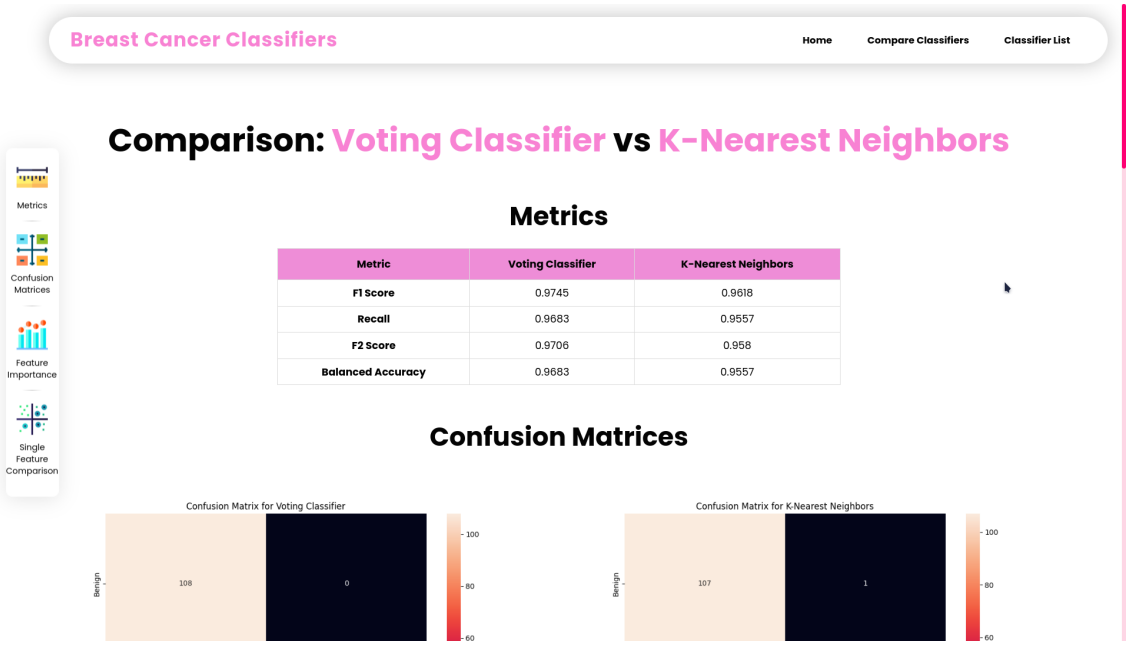


Figure B.2: Comparison of two classifiers.

C

Exploratory Data Analysis

Measure	mean radius	mean texture	mean perimeter	mean area	mean smoothness	mean compactness	mean concavity	mean concave points	mean symmetry	mean fractal dimension
count	398.0	398.0	398.0	398.0	398.0	398.0	398.0	398.0	398.0	398.0
mean	14.18	19.29	92.28	655.88	0.10	0.10	0.09	0.05	0.18	0.06
std	3.41	4.20	23.57	333.48	0.01	0.05	0.08	0.04	0.03	0.01
min	7.98	10.72	43.79	143.50	0.05	0.02	0.00	0.00	0.12	0.05
25%	11.84	16.17	76.16	432.10	0.09	0.06	0.03	0.02	0.16	0.06
50%	13.46	18.89	87.17	557.40	0.10	0.09	0.06	0.04	0.18	0.06
75%	15.78	21.60	104.00	782.68	0.11	0.13	0.13	0.07	0.20	0.07
max	27.22	39.28	182.10	2250.00	0.16	0.35	0.43	0.20	0.30	0.10

Measure	radius error	texture error	perimeter error	area error	smoothness error	compactness error	concavity error	concave points error	symmetry error	fractal dimension error
count	398.0	398.0	398.0	398.0	398.0	398.0	398.0	398.0	398.0	398.0
mean	0.39	1.21	2.79	38.14	0.01	0.03	0.03	0.01	0.02	0.00
std	0.23	0.56	1.69	33.28	0.00	0.02	0.03	0.01	0.01	0.00
min	0.11	0.36	0.76	7.23	0.00	0.00	0.00	0.00	0.01	0.00
25%	0.23	0.83	1.60	18.15	0.01	0.01	0.02	0.01	0.02	0.00
50%	0.33	1.08	2.33	24.61	0.01	0.02	0.03	0.01	0.02	0.00
75%	0.47	1.46	3.37	45.41	0.01	0.03	0.04	0.01	0.02	0.00
max	1.51	4.88	11.07	233.00	0.03	0.14	0.40	0.05	0.08	0.03

Measure	worst radius	worst texture	worst perimeter	worst area	worst smoothness	worst compactness	worst concavity	worst concave points	worst symmetry	worst fractal dimension	target
count	398.0	398.0	398.0	398.0	398.0	398.0	398.0	398.0	398.0	398.0	398.0
mean	16.30	25.70	107.47	876.22	0.13	0.26	0.27	0.12	0.29	0.08	0.37
std	4.63	6.00	32.09	525.51	0.02	0.16	0.21	0.07	0.06	0.02	0.48
min	7.93	12.49	50.41	185.20	0.07	0.03	0.00	0.00	0.16	0.06	0.00
25%	13.13	21.22	84.66	524.05	0.12	0.14	0.11	0.06	0.25	0.07	0.00
50%	15.00	25.45	98.13	688.10	0.13	0.21	0.23	0.10	0.28	0.08	0.00
75%	18.71	29.62	125.05	1063.50	0.15	0.34	0.39	0.16	0.32	0.09	1.00
max	33.12	45.41	220.80	3216.00	0.22	1.06	1.25	0.29	0.66	0.21	1.00

Table C.1: Descriptive statistics of the whole training dataset.

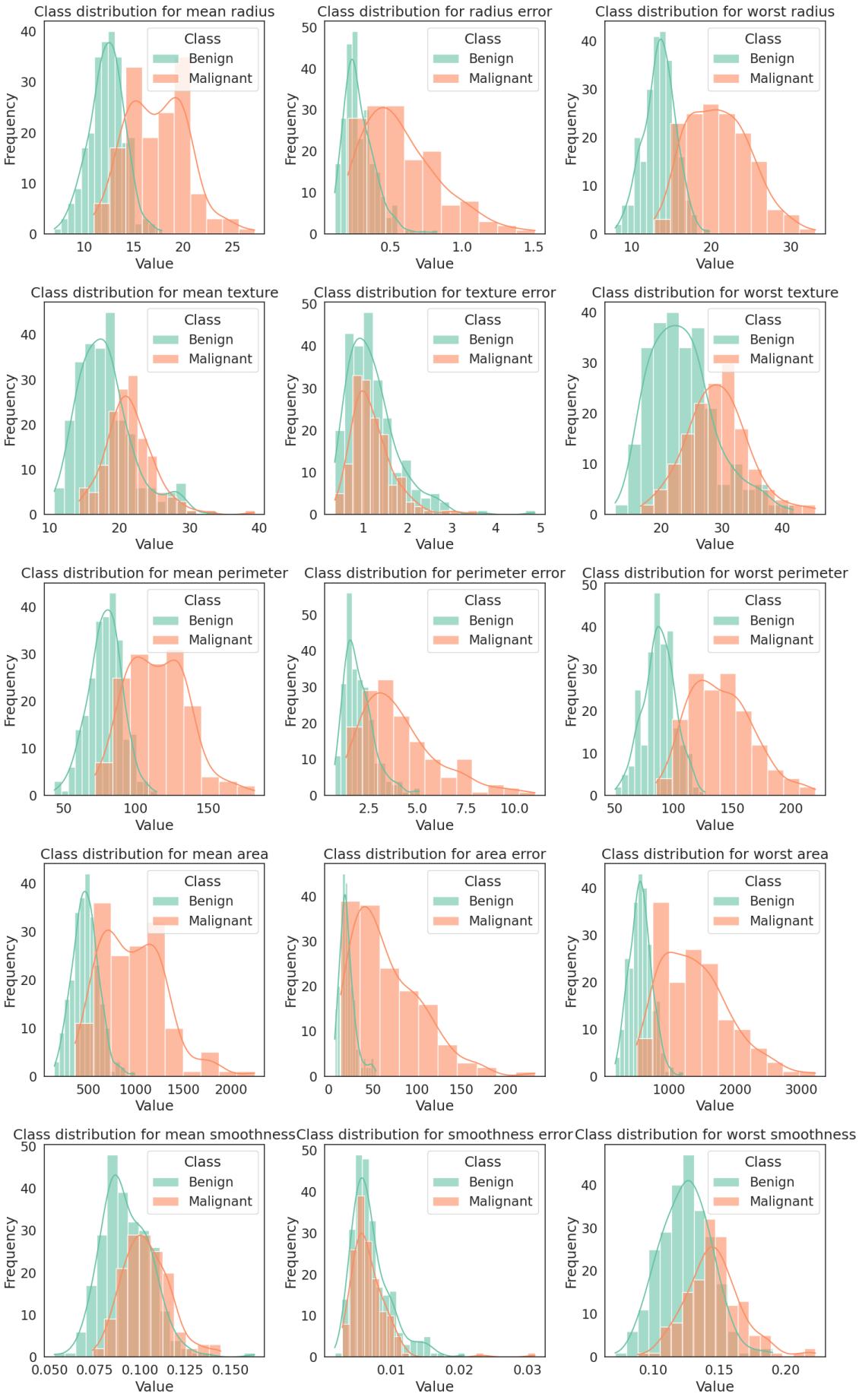


Figure C.1: First half of the features' distribution.

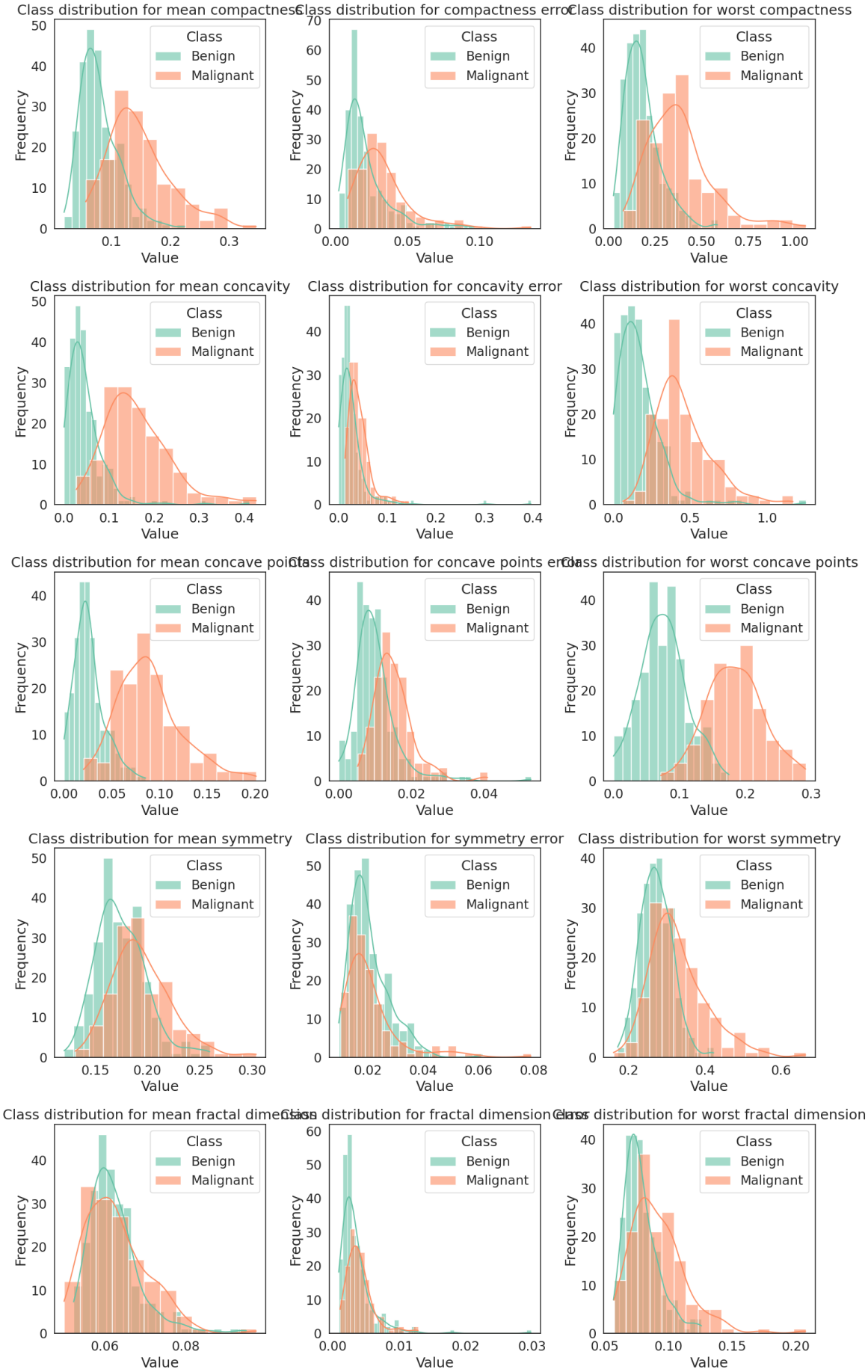


Figure C.2: Second half of the features' distribution.


UÁM
Universidad Autónoma
de Madrid