Multivariate Analysis Report: Breast Cancer Diagnosis With Supervised and Unsupervised Learning Techniques (Wisconsin Dataset)

Catarina Rodrigues, Diogo Monteiro, Filipa Costa, Mariana Lopes

ist192434, ist1102093, ist192626, ist1102094

Abstract—Breast cancer is a disease in which cells in the breast grow out of control and is one of the most common diseases in women. To explore this disease and construct predictive models, the Wisconsin Breast Cancer (Diagnostic) Dataset was acquired from Kaggle and the main objective of this work is to distinguish whether the tumour is malignant or benign. In order to avoid the overfitting phenomena, the dataset was splitted into training and test sets, and the explanatory variables were pre-processed through different techniques such as normalisation, standardization, correlation between independent features, correlation between independent features and diagnosis variable, and principal component analysis. The performance measures selected to evaluate our classifiers were sensitivity, specificity, accuracy, balanced accuracy, precision, F-Measure and confusion matrix. In order to perform more complex processing tasks, it was used unsupervised methods, such as K-medoids, K-means and Hierarquical Clustering. Using the standardized dataset the best unsupervised method was the K-means with an accuracy of 91.46% and sensivity of 97.56%. After discovering the best unsupervised method, all the remain datasets were run with K-means and the dataset with the best results were the normalized ones (with all the 30 variables and the PCA with 5 variables). Only the results of the PCA normalized dataset were presented in the report, since this dataset includes only five variables and the results between this one and the dataset with 30 variables were almost the same. Because of this, one realized that not all of the variables are important to predict whether a tumour is malignant or benign. With regard to supervised methods, it was implemented 6 different algorithms namely k-Nearest Neighbors, Naive Bayes, XGBoost, Random Forest and Linear Discriminant Analysis using the training standardized dataset. The most promising models were XGBoost with an accuracy of 98.83% and sensitivity of 99.09%, and kNN with 97.07% of accuracy and sensitivity of 99.09%. Using the clusters obtained with unsupervised learning, the supervised methods were tested again with the 2 best models (XGBoost and kNN). The results were much better, when compared to the ones obtained in unsupervised methods.

Keywords—Breast cancer, Machine Learning Models, Supervised Learning, Unsupurvised Learning, Clustering

I. Introduction

Breast cancer is a disease in which some of cells of the breast tissue grow uncontrollably. It's the most frequent invasive cancer in women in the United States [1]. A tumor can be malignant or benign, the difference between them relies on the velocity of growth and spread. Malignant tumors grow rapidly, invade and destroy nearby normal cell tissues. Benign tumors grow slowly and do not spread [2]. There are several procedures to diagnose breast cancer, which include core needle biopsy, breast magnetic resonance imaging, diagnostic mammogram and breast ultra-sounds [3].

II. Exploratory Data Analysis

The Wisconsin Breast Cancer (Diagnostic) Dataset was acquired from Kaggle [4]. All variables of this dataset were computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. It is done with the help of a graphical computer program named Xcyt, which allows the operator to input the approximate location of a sufficient nucleus to have a representative sample. Once the nucleus is identified by the user, the computer calculates ten nuclear features for each nucleus. The size of the nucleus is expressed by the variables **Area**, **Perimeter** and **Radius**. The nuclear shape is expressed by the variables **Smoothness**, **Concavity**, **Compactness**, **Concave Points**, **Symmetry**, and **Fractal Dimension**. The nuclear **Texture** is expressed by the variance of grey scales intensities in the component pixels. Each image's **mean value**, **worst value** (mean of the three biggest values), and **standard error** (SE) were calculated for each feature, yielding 30 real-valued numerical features. For instance, variable 3 represents Mean Radius, variable 13 provides Radius SE, and variable 23 shows Worst Radius. The dataset has also the information of **ID-Number** of the patient and the **Diagnosis** (B for Benign, M for Malignant) result. If the result is malignant, it means the patient's breast mass contains cancerous cells; otherwise, it is benign. The main purpose of this study is to develop a classifier that can help doctors diagnose patients and determine whether they have benign or malignant breast cancer.

In the first place, the dataset was visually inspected, and it was detected that there are 33 variables. After that, one checked for missing values and identified that the last variable was solely composed of missing values, so it was eliminated. Apart from the ID-Number, which is an irrelevant variable for this statistical analysis, there is only one category vari-

able (Diagnosis) and the remaining variables are real-valued numeric variables. Thus, the dataset is made up of 569 observations over 31 variables. For the variable Diagnosis, Benign (B) and Malignant (M) records were converted into 0 and 1, respectively. The number of malignant and benign cancer cases were plotted to better evaluate and interpret the information for the diagnosis variable, which is the main study variable.

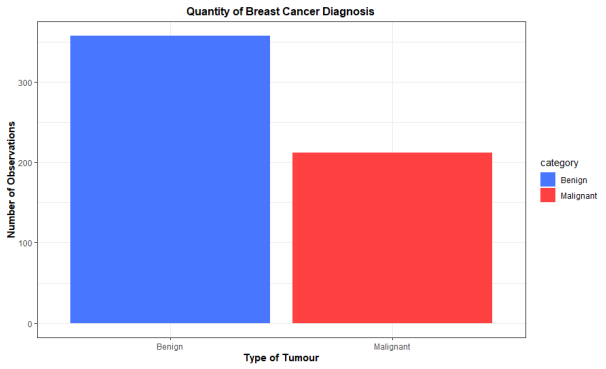

Fig. 1: Quantity of the cancer cases in Breast Cancer Wisconsin (Diagnostic) Dataset

Through the observation of the barplot in Figure 1 it is possible to see that there are more benign tumors (357 observations), than malignant tumors (212 observations), which makes this dataset rather unique when discussing carcinogenic disorders. Indeed, the presence of larger number of benign diagnostic results than their opposites, i.e. malignant outcomes, would have been expected, because it is normal that there are more healthy patients than cancer patients.

Next, the 30 variables were grouped into three categories to make them easier to understand: mean values, standard error values, and worst values. Histograms (for each variable mean, standard error, and worst values) were plotted for the benign and malignant instances in an attempt to verify the necessary assumptions for future model usage and better understand the data, by looking at the various qualities and distributions of the different variables. For simplification purposes, it was only decided to plot the histograms for each component mean.

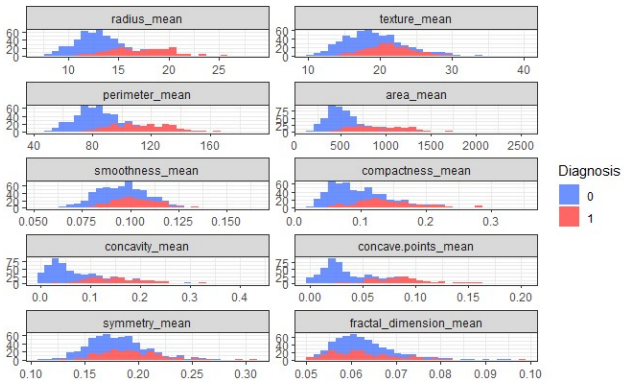

Fig. 2: Histograms of all components mean according to the diagnosis

With the exception of features related to the size of the cell nucleus, which show a slight separation, in Figure 2 it is possible to visualize that for most variables there is no obvious split between the two classes (Benign and Malignant) for each feature. The same thing applies in the "Worst" category, but in the "SE" category, there are no variables that distinguish between groups. It is also worth noting that symmetry in histograms appears in just a small percentage of cases, implying that the data does not follow a normal distribution. To have a better understanding of the data and its atypical values, it is required to check for outliers in every attribute. To accomplish so, the boxplot for all components by diagnosis class was plotted.

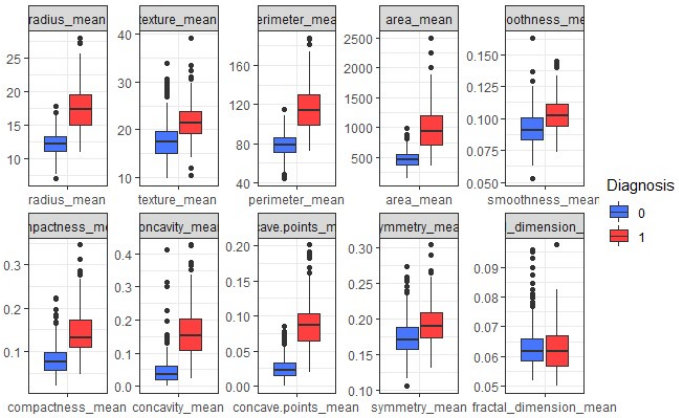

Fig. 3: Boxplot of all variables mean according to the diagnosis

Figure 3 shows a boxplot for each component's mean, with red and blue denoting variables restricted to malignant and benign data, respectively. At first sight, it is observable a few outliers in each variable, whether on the benign subset or the malignant one. However, when taking a closer look, visually the data follow a continuous line. Thus, these observations are called outliers because they are outside the interquartile region, however, as these data have a similar behavior, one just call them atypical observations. It is also worth noting Figure 3 that benign cells tend to have smaller values in all variables than malignant cells, however when looking at the extremes, there are large and small records in both groups. The same conclusions obtained with the mean values can be drawn from the worst and standard deviation boxplots.

With this in mind, it may be inferred that certain of the atypical findings (specifically, extremely low benign values and extremely high malignant values) include useful information concerning cancer diagnosis, particularly the malignant ones, when a malignant tumor is strongly suspected. On the other hand, it was deduced that benign (resp. malignant) observations with higher (resp. lower) values than usual constituted marginal instances in which a patient is equally likely to have breast cancer or not based on each assessed attribute values.

Finally, it was decided not to eliminate these atypical values for three major reasons. Firstly, when used the chosen classifiers without eliminating any outliers, one got extremely good results, which led us to assume that the outliers did not have a detrimental impact on the approaches made. Secondly, one downside of eliminating an observation is that it may be an outlier in one variable but not in another, and therefore information would be lost by removing it.

Last but not least, because all values are unique to the relevant patient, it is important to consider the most severe examples, as they may contain the most useful information

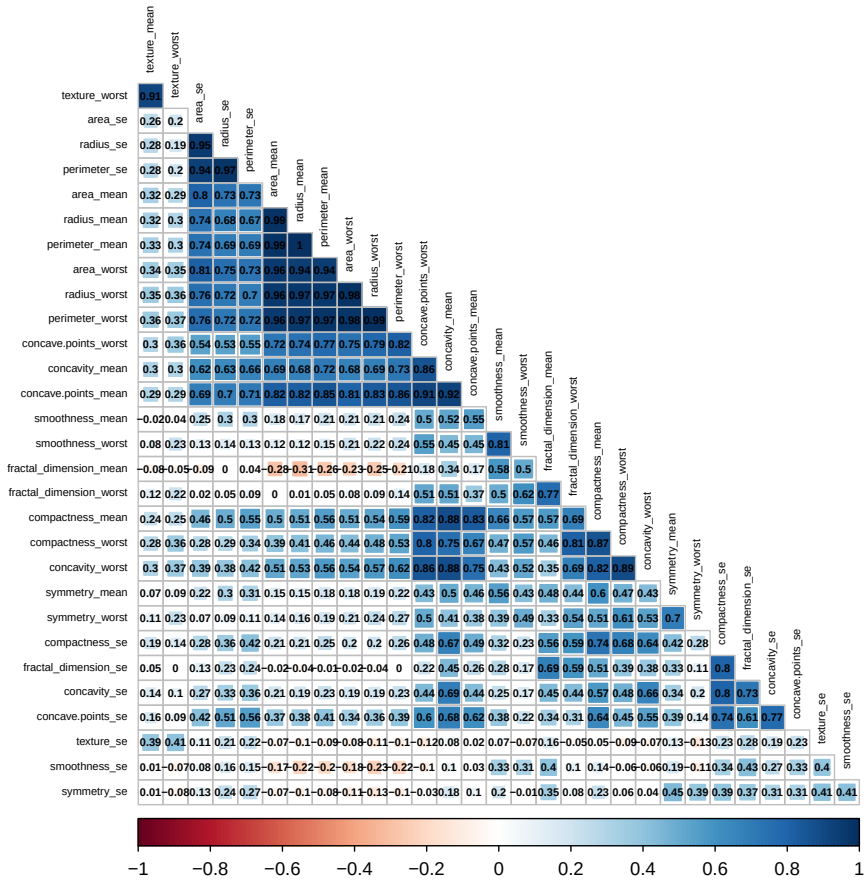

Fig. 4: Correlation plot with all the predictive variables

for discovering unique patterns and qualities in the data that influence medical diagnosis. Indeed, in the original paper [5] for this dataset, the authors mention that "the features are numerically modeled such that larger values will typically indicate a higher likelihood of malignancy".

In addition, a measure must be established to assess the correlations between 30 various features and their impact on cancer cell prediction. As a result, the correlation matrix is computed.

In Figure 4, it is observable that a few pairs of variables are highly positively correlated with each other, such as `area_mean`, `radius_mean`, `perimeter_mean`, `area_worst`, `radius_worst`, which makes sense because area, radius and perimeter are all related metrics. These highly correlated variables will be pre-processed afterwards since they can express some redundant information.

III. Pre-processing

a. Test and Training Data

The training set is composed by 70% of the original data and the test set is composed of the remaining 30%. The splitting of the data was handled carefully, since the same proportion of Benign and Malignant observations of the response variable in both sets needed to be guaranteed. For this reason, 246 (resp. 152) observations of the Benign (resp. Malignant) class appear in the training set, however only 111 (resp. 60) of them are tested.

b. Normalisation

In the explanatory variables, it is necessary to proceed to their normalisation, since they present different scales among themselves. As a result, there is a risk that features with larger magnitudes will be given more weight. Some classification techniques (such as those relying on distances) will suffer as a result, and one clearly don't want the algorithm to be biased towards one variable. Thus, the min-max scaling was used in order to transform the values of the variables to values in the scale [0,1].

$$x_i scaled = \frac{x_i - min(\mathbf{x})}{max(\mathbf{x}) - min(\mathbf{x})} \tag{1}$$

where x_i represents the i-th observation of the variable x. To do this, the scaler was fitted using the training data, and afterwards the training and test data were normalised with that same scaler. This is done because the test data is handled as new, unseen data.

c. Standardization

1. Standardization (mean+sd)

With a unit standard deviation, the standardization scaling procedure centers the results around the mean. This means that the variable's mean becomes zero, and the resultant distribution has a unit standard deviation. Unlike normalization, standardization does not limit the values to a specific range. This manner, even if there are outliers in the data, standardization will not affect them.

$$x_i scaled = \frac{x_i - \mu}{\sigma} \quad (2)$$

where μ represents the sample mean of the variable x and σ the correspondent sample standard deviation. Similarly to the normalization procedure, the scaler was fitted using the training data, and afterwards the training and test data were standardized with that same scaler.

2. Standardization (median+MAD)

To acquire a more robust standardization that is still simple and quick to execute, the mean was replaced by the median, and the standard deviation was replaced by the Median Absolute Deviation (MAD).

d. Correlation between independent features

1. Pearson's Correlation

By analysing the correlation matrix available in Figure 4, it was investigated which explanatory variables had a correlation of more than 80%. Since highly correlated variables explain the same information, then in principle it is not necessary to have both in the model. There may be interference between them, adding complication to the separation process and bias to the prediction models. Among highly correlated variables, the one with the largest mean absolute correlation was eliminated. This approach leads us to keep the variables:

```
texture_mean, concavity_mean,
concave.points_mean, symmetry_mean,
fractal_dimension_mean, perimeter_se,
smoothness_se, compactness_se,
concavity_se, concave.points_se,
fractal_dimension_se,
perimeter_worst, area_worst,
concavity_worst, concave.points_worst,
symmetry_worst, fractal_dimension_worst
```

It is important to note that when correlation analysis is applied to normalized and standardized data the results are the same as for unscaled data.

2. Robust Correlation

Since Pearson's correlation is restricted to linear associations and is overly sensitive to outliers, it was decided to experiment robust methods. Thus, to prevent the instability of classical methods of estimation in the presence of outliers in the data, the robust correlation matrix was also analyzed. To create this matrix, MCD was used to calculate robust covariances, which were then standardized to correlations using the *cov2cor R* function.

The cutoff was again considered 80%, and 18 variables remained after the robust correlation analysis. It should be noted that the variables that remained were the same as before, but instead of `symmetry_mean`, `texture_mean`, area_worst, one has the variables `texture_se`, `perimeter_mean` and `smoothness_worst`. Those results can be explained by the fact that those robust techniques down-weight data points from the samples being drawn.

e. Correlation between independent features and diagnosis variable

Additionally, point-biserial correlation was used to determine the relationship that exists between the independent variables (continuous variables) and the diagnosis variable (dichotomous variable).

All variables with a point-biserial correlation score of more than 0.95 were excluded from this analysis since we only wanted to look at the performance of the classifiers with features that were significantly linked to diagnosis.

As a result, one decided to preserve the variables: `radius_worst`, `perimeter_worst`, `concave.points_worst`, `concave.points_mean`. Then, using a pearson correlation analysis, since there were two variables of the same measure in the dataset (`concave.points_mean` and `concave.points_worst`), the variable with the highest mean correlation, `concave.points_mean`, was then removed, leaving only three explanatory variables.

It is important to mention that both point-bisserial analysis and correlation analysis were performed since one intended to examine the associations between the dependent and independent variables. However, because the dependent variable is not continuous, pearson correlation could not be used.

f. Principal Component Analysis

Lastly, Principal Component Analysis (PCA) was performed on the datasets obtained with the previously techniques. For the normalization, standardization, and correlation datasets the classic pca was used. As for the robust methods, such as the robust standardization and robust correlation, the robust PCA method proposed by Hubert was used. It was established that each dataset should have enough principal components to explain at least 80% of the variability in the data. Moreover, this decision was reinforced by analysing the plot of the percentage of variance explained as a function of number of principal components, following the elbow method.

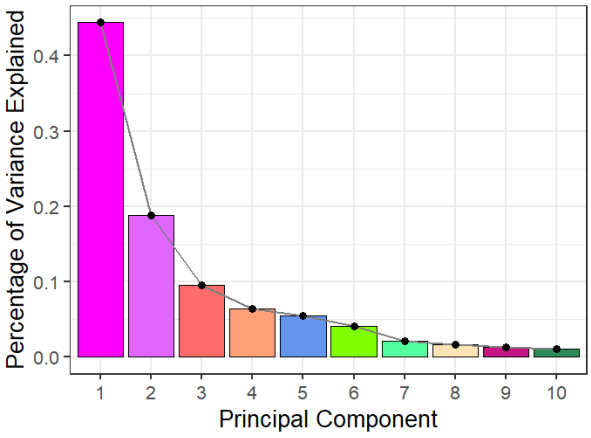

Fig. 5: Elbow method for optimal number of principal components in the standardized datatset

In Figure 5,it is possible to observe that for the standardized dataset the percentage of variance explained drops after 5 principal components. Accordingly, the cumulative proportion of 5 principal components is 0.8482 and so it was

decided to keep this amount.

IV. Classifiers

a. Validation Indices and Confusion Matrix

Having in hand a binary classification problem, here are the measures used to evaluate the quality of the resulting models of the classifiers:

- **Sensitivity** - relative frequency of true positive observations and all the positive observations;
- **Specificity** - proportion of false negatives in relation to all observations that are really negative;
- **Accuracy** - relative frequency of correctly classified observations in the sample;
- **Balanced Accuracy** - due to the possibility that there are unbalances in the classification, the balanced accuracy can be obtained through the average of sensivity and specificity;
- **Precision** - rate of true positives among all the observativons that are classified as positive;
- **F-Measure** - the F-measure, or F1-score, is built on the basis of accuracy and sensitivity, and this can be interpreted as a harmonic mean of these measurements.
- **Confusion Matrix** - The confusion matrix describes the performance of the classifier. If $i = j$, the input a_{ij} corresponds to the number of correct classifications of an observation of a class i; if $i \neq j$, then it corresponds to the number of times class i (j) is incorrectly identified as j (i). From the confusion matrix, it is possible to determine other performance measures, described above.

b. Unsupervised Learning - Clustering

It was decided to use three techniques for clustering: Hierarchical Clustering, K-means, and K-medoids.

Several approaches were used to determine the best number of clusters (k), including the elbow method, the average silhouette width, and the between and within sum of squares. Some indices, such as the Davies Bouldin-index, the Dunn-index, and the C-index were also investigated. Using the standardized dataset, the algorithms were computed for the values of k=2,...,7.

In general, the above approaches showed that $k = 2$ was the best one for Hierarquical Clustering, K-means and K-medoids, , but $k = 3$ also performed well. However, because the diagnosis variable has two alternative outcomes (malignant and benign), the three algorithms were computed with $k = 2$ so that the clustering results could be externally validated.

Because one is working with continuous real variables, one attempted to conduct K-medoids with the Euclidian and Manhattan distances. Since the methods for determining the ideal k yielded superior results when the Euclidian distance was used, the Euclidian distance was chosen to perform the clustering algorithms.

1. Hierarquical Clustering

For the Hierarquical Clustering, four different approaches were experimented, with the Euclidian distance: Single Linkage, Complete Linkage, Average Linkage and Ward's Method. These four hierarquical clustering techniques were performed in the standardized dataset to see which one better describes the data.

Table 1: Confusion Matrices using Hierarchical Clustering(Single Linkage, Complete Linkage, Average Linkage and Ward's Method) performed in the train set, for the standardized dataset

Hierarquical Clustering								
Method	Single Linkage		Complete Linkage		Average Linkage		Ward's Method	
Classes	*B*	*M*	*B*	*M*	*B*	*M*	*B*	*M*
Cluster 1	245	152	246	150	246	150	228	28
Cluster 2	1	0	0	2	0	2	18	124

As it is noticeable in the confusion matrices displayed in Table 1, for Single Linkage, Complete Linkage, and Average Linkage one obtained very poor dendograms that were unable to classify Malignant data. The Ward's Method, on the other hand, only miss-classified 46 observations and performed exceptionally well across all performance metrics (for example, it got an accuracy of 0.8844 on the train set and 0.9123 on the test set), hence it was chosen to be the best one.

Finally, when considering the standardized dataset and choosing the ward's method, a dendogram was constructed as shown in Figure 6.

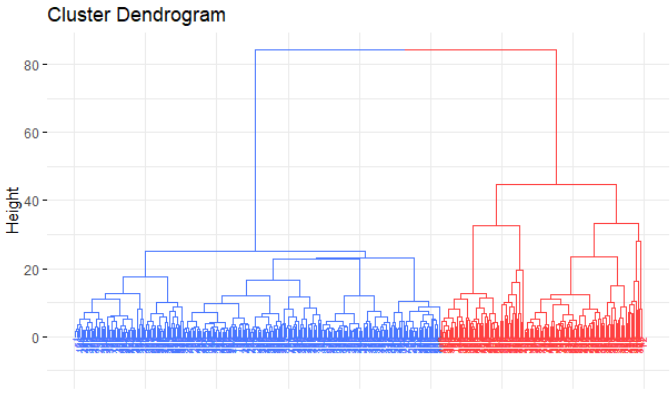

Fig. 6: Dendogram for the Standardized Dataset with Ward's Method

In this dendogram, for amusement, all the observations were correctly classified.

2. Selection of the "Best" Method

To find the best unsupervised method, the three clustering techniques (Ward's Method, K-Means and K-Medoids) were run one time for the standardized dataset. Table 2 shows the resultant confusion matrices and the correspondent accuracies and sensitivites:

As it can be observed, 46 observations (for Hierarchical Clustering), 34 observations (for K-means), and 42 observations (for K-medoids) were assigned improperly to the cluster where they should be. One also analysed the accuracy values for each clustering technique and the accuracy for the K-means method is the highest value between the three methods, with a value of 0.9146. Finally, it is critical to have a

Table 2: Hierarquical clustering, K-means and K-medoids for the training standardized dataset

Best Clustering Method - train data						
Method	**Hierarquical Clustering**		**K-Means**		**K-Medoids**	
Classes	*B*	*M*	*B*	*M*	*B*	*M*
Cluster 1	228	28	240	28	238	33
Cluster 2	18	124	6	124	8	119
Accuracy	0.8844		0.9146		0.8970	
Sensitivity	0.9268		0.9756		0.9674	

high sensitivity (i.e. the amount of malignant observations that were categorized as benign) when determining whether a tumor is benign or malignant. Since classifying a malignant tumor as benign is worse than diagnosing a malignant tumor as benign, K-Means was the algorithm chosen as the best classifier to perform with the remaining datasets.

After applying K-means to the remaining datasets, the best results are achieved when the dataset was normalized, which would indicate that the min-max normalization scaling is a very good indicator of the importance and influence of the explanatory variables to the diagnosis determination.

On the other hand, when applying PCA to the normalized dataset, the results were almost the same, although this time there were only 5 variables. This way, comparing these two normalized datasets, one can conclude that not all variables are useful for the model. Because of this, in Table 3, one will only present the performance measures when the PCA normalized dataset was used.

It is worth noting that the method used to predict K-means clustering results was designed with the goal of assigning each new observation to the cluster whose center has the lowest euclidian distance from it.

Table 3: Performance Measures for the normalized PCA dataset, with 5 variables

Performance Measures - 5 variables				
Normalization+PCA	**Train data**		**Test data**	
Classes	*B*	*M*	*B*	*M*
Cluster 1	237	29	109	8
Cluster 2	9	123	2	52
Accuracy	0.9045		0.9415	
Sensitivity	0.9634		0.9820	
Specificity	0.8092		0.8667	
Balanced Accuracy	0.8863		0.9243	
Precision	0.8910		0.9316	
F1-Score	0.9258		0.9561	

These performance measurements reveal that the sensitivity was quite high, implying that practically all malignant observations were accurately diagnosed and only 8 benign observations were incorrectly classified. It is also worth remembering that the PCA results were better than some of the outcomes that employed datasets with more variables since the datasets obtained from PCA have substantially fewer variables, which reduces the likelihood of having low variance variables, which makes the clustering process much more difficult.

As it be can seen in the Graph 7 below, the elbow method is represented with the normalized PCA dataset. Through the graph one can conclude that the optimal k value is indeed two (as was said at the beginning of this chapter).

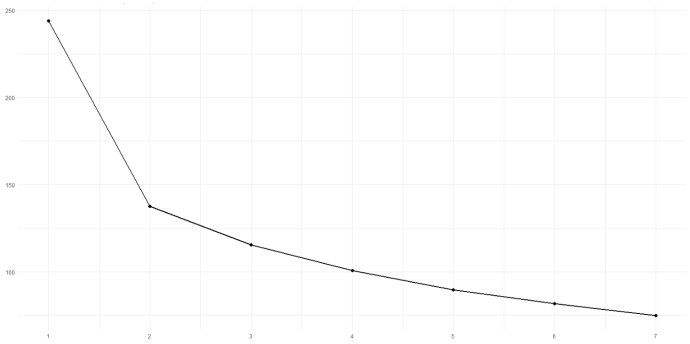

Fig. 7: Elbow Method for optimal value of k in K-Means for the Normalized+PCA dataset

The Figure 8 contains a scatter plot of data points colored by cluster numbers. In the resulting plot, observations are represented by points, using the 2 principal components (Dim1 and Dim2) that represent the highest variability.

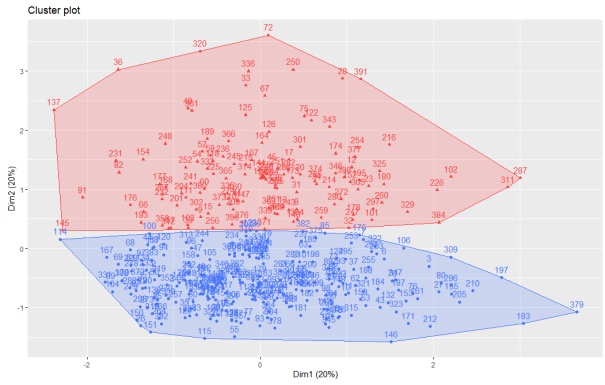

Fig. 8: Cluster Plot for the Normalized+PCA dataset

The red points in this figure represent cancerous cases, while the blue points represent benign cases. One can see that there is some little overlap between the two categories of diagnosis that correspond to three observations, which suggests that only three observations are misclassified if just the two main components with the largest variability are evaluated. However, if plotting in five dimensions were allowed, there would only be an overlap of the 38 observations that are incorrectly labeled while clustering.

c. Supervised Learning

The main purpose of Supervised Learning is to build a model which can learn from a training dataset, and predict outputs given new inputs. The training set is formed by the inputs and their corresponding outputs, allowing to the model to learn progressively with time. The accuracy of the algorithm is determined by the loss function, which suffers adjustments until the error, between the new and old outputs, has been minimized [6]. One choose to implement 6 different models to predict the outputs of the test dataset namely K-Nearest Neighbors, Naive-Bayes, XGBoost, Random Forest and Linear Discriminant Analysis. In all of them, one decided to develop them through application of the function train of the package caret with cross-validation (number of folds equals to ten). Cross-validation allows estimate the competence of a model on unseen data (data not used for training the model). The metric implemented to select the optimal model is the

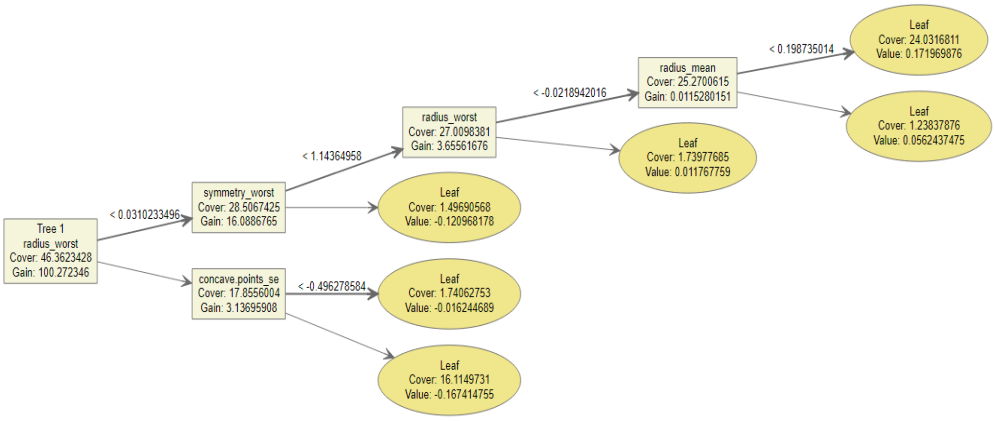

Fig. 9: Resulting tree of the XGBoost model for the standardized dataset

Accuracy. Moreover, all models were developed and trained with standardized dataset in order to select the best supervised method.

1. k-Nearest Neighbours

k-Nearest Neighbors (kNN) method allows to make a prediction based on the distance between the test set and all the training set. kNN classifier only requires a parameter k, a set of labeled training samples, and a metric measure for determining distances [7, 8]. This algorithm works in a simple way. The first step was to determine the parameter k (number of nearest neighbors). There are several ways of tunning, one chooses to use an interval of the number of neighbors from 3 to 19, in steps of 2. It is better if k is an odd number, as this avoids statistical ties, which are more probable with even numbers. The value of k will be the hyperparameter to be determined with the k-fold CV. Then, the algorithm calculates the distance between each testing sample and all the training samples. There are several distance measures, but by default in R, the distance measure is the Euclidean one. Next, the algorithm sorts the distance and determine nearest neighbors based on the k-th threshold.

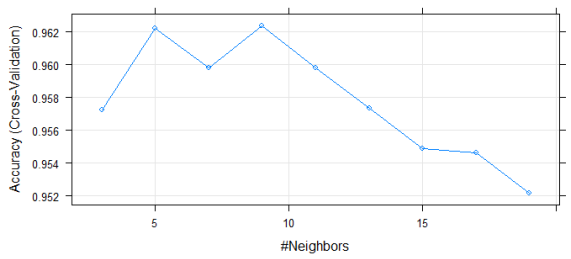

Fig. 10: Relation between k values and accuracy of the kNN model

By analysing the Figure 10, it is possible to see that the best number of nearest neighbors was 9 for our dataset. With this, the algorithm is able to determine the category (class) for each of the nearest neighbors and uses simple majority of the category of nearest neighbors as the prediction value of the testing sample classification.

2. Naïve Bayes

Naïve Bayes (NB) Classifier is a probabilistic algorithm that is usually used for classification problems. In this algorithm, the distribution of samples in each class is modelled using a probabilistic model which assumes that all the variables (given the class) are independent from each other. The Naive Bayes classifier combines Naive Bayes model with the maximum a posteriori decision rule, which instructs the model to choose the most probable hypothesis in the end. It finds the probability of a given set of inputs for all possible values of the possible outcomes and pick up the output with maximum probability [9]. However, one disadvantage of this method is if the individual class is missing when the frequency-based probability estimate will be zero, so we will get a zero when all the probabilities are multiplied. To overcome this issue, one choose to apply Laplace Smoothing [9]. Laplace Smoothing is a technique that ensures that each feature has a nonzero probability of occurring for each class. The application of Laplace Smoothing was done through the expand.grid function in which one choose to use parameters usekernel=TRUE, adjust=c(0, 0.5, 1.0) and fL= c(0, 0.5, 1.0). The last mentioned parameter allows to incorporate the Laplace smoother, while the others allow to adjust the bandwidth of kernel density and to use kernel density estimate, respectively.

3. XGBoost

XGBoost classifier is a supervised learning algorithm that is applied for structured and tabular data. It is an implementation of gradient boosted decision trees designed for speed and performance. Furthermore, Gradient boosting is a powerful technique, able to fit non-parametric predictive models and was motivated as being a gradient descent method in a function space that is the reason it is called a “Gradient Boosting Algorithm”. [10]. There are several ways to prevent overfitting in XGBoost, specifically control directly the model complexity and to add randomness to make training robust to noise. Therefore to all booster parameters, except gamma, one chooses to apply an interval of values well established in the literature [11]. The parameter gamma of tree booster, which is minimum loss reduction required to make further partition on a leaf node was set to 0.

Figure 9 shows a resulting tree of depth 4 of this method for the standardized dataset. This low value of max_depth makes the model generalize better by learning less likely from noise.

4. Random Forest

Decision Trees are a non-parametric supervised learning method used for classification. They learn from data to approximate a sine curve with a set of if-then-else decision rules. The deeper the tree, the more complex the decision rules and the fitter the model [12]. The Random Forest (RF) method is a combination of individual decision trees, in which each tree depends on the values of a random vector sampled independently, and with equal distribution for all trees involved in the forest [12]. In this case, the square root of the number of variables is used in the input.

5. Linear Discriminant Analysis

Linear Discriminant Analysis (LDA) is a simpler algorithm in terms of preparation and application. However, this method makes some assumptions about the data, which are assuming that the predictor values are normally distributed, and the classes have identical covariance matrices [8, 13]. At the first hand, one can imagine that its performance might be slightly worse when compared to the other methods. For The training of this method no parameter was adjusted.

6. Selection of the "Best" Method

To find the best supervised learning method, the five techniques mentioned above were run one time for the standardized dataset. Table 4 shows the confusion matrices and Table 5 presents the performance measures. Through the observation of the results in Tables 4 and 5, the model with worst performance was NB, assigning 9 observations improperly. NB assumes that all features are independent, which most of the times does not happen in real life specially in the medical field. In addition, it is also possible to see that the best two models are XGboost and Random Forest, which assigned 2 and 4 observations improperly, correspondingly. However, they have methodologies very similar, they both use decision trees. Therefore, one can say that between these 2 models the most promising is XGBoost.

Table 4: Confusion Matrices - Standardized test set

Method	**kNN**		**NB**		**XGBoost**		**RF**		**LDA**	
Classes	*B*	*M*	*B*	*M*	*B*	*M*	*B*	*M*	*B*	*M*
B	110	4	107	5	110	1	110	3	110	4
M	1	56	4	55	1	59	1	57	1	56

Table 5: Performance Measures - Standardized test set

Method	**kNN**	**NB**	**XGBoost**	**RF**	**LDA**
Accuracy	0.9707	0.9415	0.9883	0.9766	0.9707
Sensitivity	0.9909	0.9549	0.9909	1.0000	0.9909
Specificity	0.9333	0.9166	0.9833	0.9333	0.9333
Balanced Accuracy	0.9621	0.9358	0.9871	0.9666	0.9621
Precision	0.9649	0.9549	0.9909	0.9652	0.9649
F1-Measure	0.9777	0.9549	0.9909	0.9823	0.9777

To obtain holistic vision of supervised learning methods, one decided to consider another optimistic model. The performance of kNN with 9 nearest neighbors is not so far from Random Forest, and seem to be promising. Although kNN and LDA present the same results, one preferred kNN because kNN is completely non-parametric approach, which means that there is no need to make assumptions about the shape of decision boundary.

When it comes to choosing the best classifier using the standardized dataset, one may argue that kNN is the best method to pick because its measurements are quite comparable to XGBoost (the method that performed the best), but it is less complex and time consuming. However, since one is dealing with a cancer diagnosis problem, it is not necessary to have quick answers, being preferable to use the classifier with the greatest possible performance measures. As a result, the XGBoost classifier was selected.

In order to see what were the best pre-processing techniques, one applied XGBoost to all datasets. There was a tie between the normalized, the standardized and the standardized robust datasets. They all have 30 variables and an accuracy of 0.9883. The next datasets that presented a better performance were robust correlation with an accuracy of 0.9825 and correlation with an accuracy of 0.9766, with 16 and 15 variables, respectively. In addition, the datasets of Normalized PCA and of Standardized PCA also presented good results with an accuracy of 0.9649 with only 5 variables. Also on the dataset where was performed point-biserial correlation, which had 3 variables, one obtained an accuracy of 0.9532.

V. Supervised Learning with Clustering Results

After collecting the results from the supervised and unsupervised approaches, a supervised classification was performed using only the unsupervised clusters. This way, the clustering results were considered to be the new real values for the diagnosis variable when training the models. For this, the datasets with best results from clustering (normalized (Norm) and normalized PCA) were selected, and the models were trained with the partition clusters obtained as classes. The chosen models were the XGBoost and the kNN and the results are shown in table 7

Table 6: Confusion Matrix of Test Set Using Clusters Results

Method	**XGBoost**				**kNN**			
Dataset	Norm		PCA (Norm)		Norm		PCA (Norm)	
Classes	*B*	*M*	*B*	*M*	*B*	*M*	*B*	*M*
B	110	10	110	7	110	7	110	7
M	1	50	1	53	1	53	1	53

Table 7: Performance Measures of Test Set Using Clusters Results

Method	**XGBoost**		**kNN**	
Dataset	*Norm*	*PCA (Norm)*	*Norm*	*PCA (Norm)*
Accuracy	0.9356	0.9532	0.9532	0.9532
Sensivity	0.9909	0.9909	0.9909	0.9909
Specificty	0.8333	0.8833	0.8833	0.8833

As expected, the results are worse than those obtained by training the supervised models with the true classes, since the data separated by clustering contains some incorrect classifications because the clustering methods didn't produce 100% accuracy. On the other hand, these results are a significant improvement over the ones obtained for the test data through unsupervised methods only. Furthermore, by looking at the

results we see that no difference exists between the performance of the two datasets for the kNN and the normalized PCA for the XGBoost.

VI. Discussion and Conclusion

Two distinct types of analysis can be done: one for supervised methods and one for unsupervised methods. When performing these methods, as previously explained, the original dataset was partitioned into two different subsets: the training set for modelling and the test set to test the model. Because of this, a predictive model in every setting can be used to predict new observations.

In general, when using XGBoost for supervised approaches, one can produce predictive models that are over 98% accurate, which is an astounding outcome. Furthermore, while using kNN, one can achieve a comparable result of 97%. When using the normalized, standardized, and standardized robust datasets, these results are improved. This way, despite the preprocessing methods used, all of these supervised learning methods resulted in models with very high prediction accuracy. It is worth noting that one of the greatest results was obtained when the dataset with strongly correlated variables with the diagnosis variable was used, ended up having only three explanatory attributes, namely radius_worst, perimeter_worst and concave.points_worst. As a result, these three variables can be said to be both necessary and sufficient for classifying a tumor as benign or malignant.

It is important to keep in mind that in every variable, the corresponding average values of these variables for the malignant group are greater than the values for the benign subset, according to the preliminary analysis. As a result, the malignant observations have on average higher values for radius_worst, perimeter_worst, and concave.points_worst, indicating that larger and deformed breast cells are more likely to be part of malignant tumors. Furthermore, the type of cell tumor may be determined simply by examining the values of these three properties. As a result, the medical community can be more selective in the data it collects for disease analysis, lowering research expenses and freeing up funding to explore new cancer therapies.

In the end, despite being a more sophisticated algorithm, the best classifier was XGBoost, which presented the best performance measurements. When comparing the outcomes of XGBoost, kNN, and LDA, the results are quite identical; however, XGBoost has a specificity of 98%, whilst the other two methods have a specificity of 93%. This high specificity is also important because the doctor might decide to get a false positive (malignant) rate near zero, meaning that all of the patients who are classified as unhealthy are, in fact, sick and must receive the appropriate treatment.

When performing unsupervised learning methods such as K-means, K-medoids and hierarchical clustering, one was interested in identifying patterns in the data that might be used to separate the observations (Malignant and Benign) and attempting to identify the most important variables or variable transformations that would lead to the best results. For this, since many types of preprocessing techniques were made, and the main goal is to find the best clustering method, hierarchiqual clustering, kmeans and kmedoids were performed in the standardized dataset. Starting with hierarchical clustering, very good results were achieved with the ward's method, and a almost perfect separation of the observations was achieved with a sensitivity of 93%. However, as explained before, the best unsupervised method was the K-means, since while evaluating whether a tumor is benign or malignant it is vital to have a high sensitivity (i.e. the amount of malignant observations that were correctly labeled), because this way the doctor is assured that all the patients that were classified as being healthy are completely safe and thus don't need to be subjected to additional, perhaps painful and definitely costly exams. This way, the k-means approach not only presented the highest sensitivity (98%), but also the best accuracy (91%) and confusion matrix (only 34 observations were miss-classified) values.

Furthermore, the best results were found in datasets that experienced changes, such as dimensionality reduction, showing that some of the variables are not valuable for the problem at hand.

Concerning possible future work, even though the obtained results were good, one could try some other reliable and efficient methods, such as SVM or neural networks. One could also try adding some other variables, such as whether the patient have had cancer before, the patients gender (women are more likely to get breast cancer [14]), the presence of diabetes, and if exists family history of breast cancer.

References

[1] N. C. Institute. (September 2021) Breast cancer - patient version. (accessed: 27.12.2021). [Online]. Available: https://www.cancer.gov/types/breast

[2] C. for Disease Control and Prevention. (September 2021) What is breast cancer? (accessed: 27.12.2021). [Online]. Available: https://www.cdc.gov/cancer/breast/basic_info/what-is-breast-cancer.htm

[3] N. H. Service. (October 2018) Breast cancer in women - diagnosis. (accessed: 27.12.2021). [Online]. Available: https://www.nhs.uk/conditions/breast-cancer/diagnosis/

[4] W. N. S. Dr. William H. Wolberg and O. L. Mangasarian. (September 2016) Breast cancer wisconsin (diagnostic) dataset. (accessed: 27.12.2021). [Online]. Available: https://www.kaggle.com/uciml/breast-cancer-wisconsin-data

[5] W. N. Street, W. H. Wolberg, and O. L. Mangasarian, "Nuclear feature extraction for breast tumor diagnosis," in *Biomedical image processing and biomedical visualization*, vol. 1905. International Society for Optics and Photonics, 1993, pp. 861–870.

[6] B. Shetty. (July 2019) Supervised machine learning classification: An depth in guide | built in. (accessed: 22.01.2022). [Online]. Available: https://builtin.com/data-science/supervised-machine-learning-classification

[7] Z. Zhang, "Introduction to machine learning: k-nearest neighbors," *Annals of Translational Medicine*, vol. 4, no. 11, 2016.

[8] M. Kantardzic, *Data Mining: Concepts, Models, Methods, and Algorithms*. IEEE Press, 2011.

[9] Naive bayes classifier. (accessed: 26.01.2022). [Online]. Available: https://uc-r.github.io/naive_bayes

[10] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," pp. 785–794, 2016.

[11] Xgboost parameters. (accessed: 26.01.2022). [Online]. Available: https://xgboost.readthedocs.io/en/latest/parameter.html

[12] E. Fix and J. L. Hodges, "Discriminatory analysis. nonparametric discrimination: Consistency properties," *International Statistical Review/Revue Internationale de Statistique*, vol. 57, no. 3, pp. 238–247, 1989.

[13] Y. Xiaozhou. (May 2020) Linear discriminant analysis, explained. (accessed: 22.01.2022). [Online]. Available: https://towardsdatascience.com/linear-discriminant-analysis-explained-f88be6c1e00b

[14] K. Mehra, A. Berkowitz, and T. Sanft, "Psychosocial consequences and lifestyle interventions," in *The Breast*. Elsevier, 2018, pp. 1039–1048.