DSC 532 - Statistical Learning Breast Cancer Classification Analysis

Andriani Panagi[1], Panayiota Damianou[2], and Kypriani Paraskevopoulou[3]
[1]Department of Physics
[2]Department of Mathematics and Statistics
[3]Department of Mathematics and Statistics
{apanag21, pdamia01, kparas05}@ucy.ac.cy

Abstract—**The purpose of this project is to analyse the data set "Breast Cancer". We aim to process the data set, namely perform an explanatory preprocessing analysis on the features of the data set, and build a classification model so that at the end the diagnosis will be predicted either to a cancerous or a non-cancerous type. Eventually, a conclusion on the whole data analysis is made at the end.**

I. Introduction

Breast cancer is a cancer that forms in the cells of the breast and is the second most common cancer diagnosed in women in the United States. It can occur in both men and women, but it is far more common in women. Breast cells start to grow uncontrollably with the onset of breast cancer. Typically, these cells develop tumors that can be seen via X-ray or felt as lumps or masses in the breast area. The main obstacle is to identify the distinguish benign (non-cancerous) from the malignant (cancerous) tumors [1] [2]).

From a digital image of a fine needle aspirate (FNA) of a breast mass, features that describe the traits of the cell nuclei in the image are computed. The analysis will try to classify these tumors in benign or malignant, using machine learning algorithms. Thus, the target variable will be the diagnosis column. In addition, through exploratory data analysis, correlations between the variables will be identified where at the end the diagnosis will be predicted with a classification model.

II. Information about the data set

A. Features of the data set

The data set used, consists of 569 observations and 32 variables. The data set includes the identification ('id') of each patient and their diagnosis('Malignant' or 'Benign'). The remaining 30 variables contain information about the mean, standard error, and worst values of various tumor characteristics. These predictors are:

radius_mean, texture_mean, perimeter_mean, area_mean, smoothness_mean, compactness_mean, concavity_mean,concave.points_mean,symmetry_mean, fractal_dimension_mean, radius_se, texture_se, perimeter_se, area_se, smoothness_se, compactness_se, concavity_se, concave.points_se,symmetry_se, fractal_dimension_se, radius_worst, texture_worst, perimeter_worst, area_worst, smoothness_worst, compactness_worst,concavity_worst, concave.points_worst, symmetry_worst, fractal_dimension_worst

Explaining further some of the predictors, firstly, the texture gives the patterns and variations in the pixel intensities within an image. The mean smoothness is a measure of the variation in the radius of the tumor contour and the mean concavity shows the severity of concave portions of the shape. The mean concave points is the number of concave portions of the contour and lastly the fractal dimension is an objective and reproducible measure of the complexity of the tissue architecture of the biopsy specimen. The higher the number, the more abnormal the tissue is. To distinguish between cancerous and non-cancerous tumors, there are certain characteristics, such as smoothness, that differ. The benign (non-cancerous) tumor has distinct, smooth, regular borders where as the malignant (cancerous) tumor has irregular borders and grows faster than a benign tumor. A malignant tumor can also spread to other parts of the body.

B. Data source

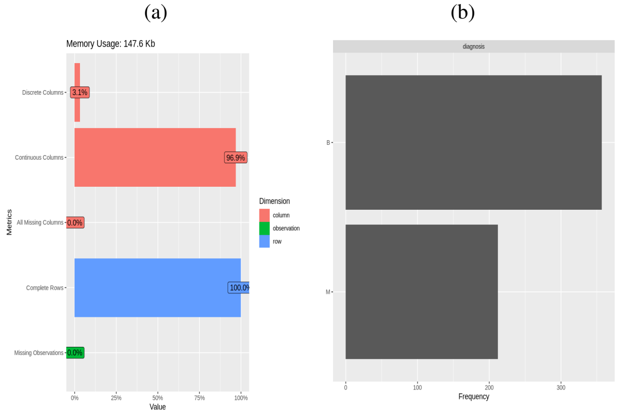
The data set used for this project is provided by Kaggle and contains information about cancerous and non-cancerous tumors as well as the diagnosis of them [3].

III. Data Pre-Processing

A. Missing and irrelevant values, duplicates

Since the 'id' column does not provide any relevant information to the analysis, it was decided to drop it.Additionally, there were no duplicates or missing values in the data set. Figure 1 (a) shows that 99.6% are numerical columns where only 3.1% are categorical. There aren't any missing columns or any missing observations as well as that all the rows are complete. Figure 1 (b) is a visual representation of the column Diagnosis namely the response column, where it contains 357 values for Malignant (cancerous) and 212 values for Benign (non-cancerous).

Fig. 1: Visualization of the missing rows, missing observations and complete columns and of the Diagnosis column



B. Distribution of the features

For all the numerical columns the histograms are made. As it can be seen, none of the features are normally distributed. Most of the distributions of the features (Figure 2) are heavy tailed and right skewed.

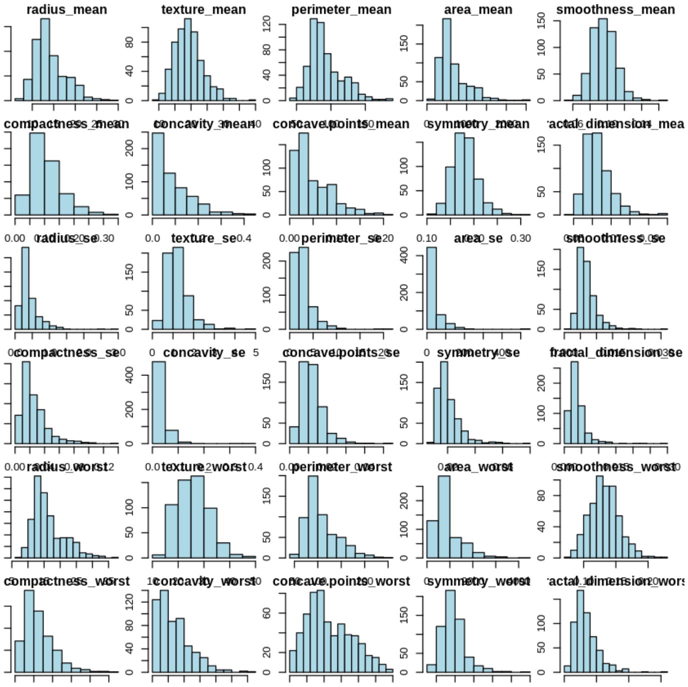


Fig. 2: Distribution of the features

C. Outliers

In order to check if there are any outliers, a box plot for each of the numerical columns is made. As can be seen from Figure 3, all the features contain outliers, with the column 'area_se' having the most. The percentage of outliers in each column was calculated, and it is ranging between 0-11% (Table 1). From the distributions in section 3.B, it is observed that many distributions are right skewed so a number of outliers in the features was expected. Considering this, and also that the data set contains only 569 rows, it was decided, not to remove any values.

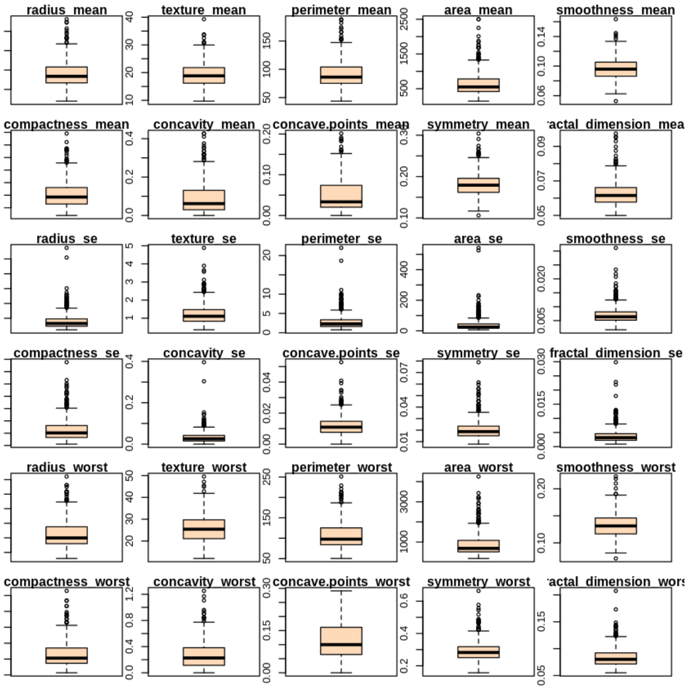


Fig. 3: Outliers of the features

TABLE I: Percentage of outliers in the numerical columns

Outliers in percentage (%)	Columns
2.46	radius_mean
1.23	texture_mean
2.28	perimeter_mean
4.39	area_mean
1.05	smoothness_mean
2.81	compactness_mean
3.16	concavity_mean
1.76	concave.points_mean
2.64	symmetry_mean
2.64	fractal_dimension_mean
6.68	radius_se
3.51	texture_se
6.68	perimeter_se
11.42	area_se
5.27	smoothness_se
4.92	compactness_se
3.87	concavity_se
3.34	concave.points_se
4.75	symmetry_se
4.92	fractal_dimension_se
2.99	radius_worst
0.88	texture_worst
2.64	perimeter_worst
6.15	area_worst
1.23	smoothness_worst
2.81	compactness_worst
2.11	concavity_worst
0.00	concave.points_worst
4.04	symmetry_worst
4.22	fractal_dimension_worst

D. Correlation between the variables

The correlation between the variables is visualized in Figure 4. From here it is evident that there are multi-correlated columns in the data set where most of them are positively correlated. Among them, there are also some negative correlated features, where the correlation range has not so high values. Some of the most correlated features are the perimeter_mean and the area_mean with the radius_worst, the radius_se with the perimeter_se and area_se, the radius_worst with the perimeter_worst and the area_worst and lastly the concave.points_mean with the concave.points_worst. The correlation is self explainable since for example the perimeter and the area can be calculated with the radius. This means that the features contain similar information.

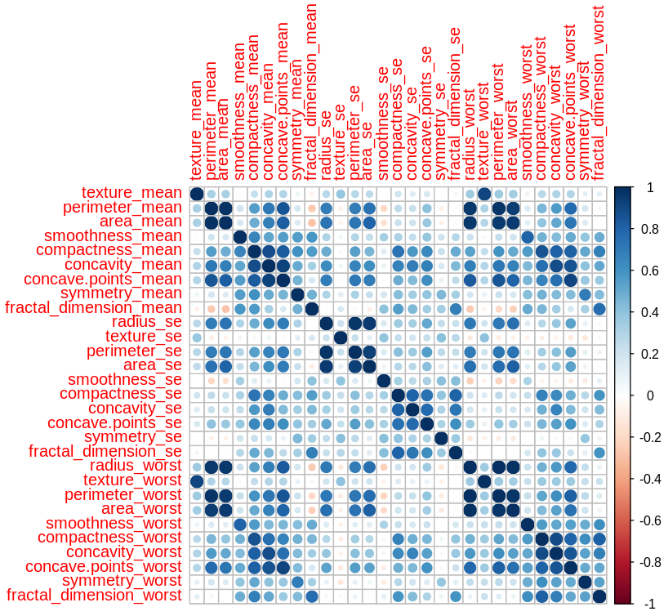


Fig. 4: Correlation between the variables

In addition, in Figure 5, four correlation plots between the features divided in two colors based on the diagnosis are illustrated. In the first two (horizontal) plots there are the scatter plots of the features area_mean vs smoothness_se and radius_mean vs fractal_dimension_mean. Despite the fact that the correlation is not so obvious like the positive correlated features, it can be seen that the correlation between the features is negative. The exact opposite stands for the other two plots. The scatter plots of the columns texture_mean vs texture_worst and concavity_worst vs concave_points_worst indicate a clear positive correlation.

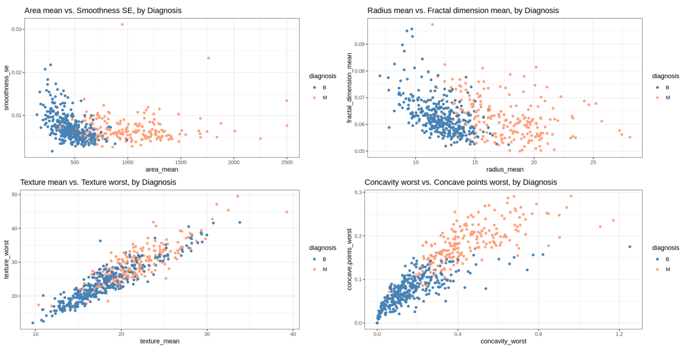


Fig. 5: Correlation plots for highly correlated features

Using the correlation matrix the 8 least correlated features with a cutoff for the correlated columns to be 0.6 were distinguished and can be seen in Figure 6. These features are the texture_mean, area_mean, smoothness_se, symmetry_se, texture_se, fractal_dimension_se, smoothness_worst and fractal_dimension_worst. From these plots one can definitely see the difference between the Malignant and the Benign diagnosis in the mean area while for the other columns the distributions are roughly the same.

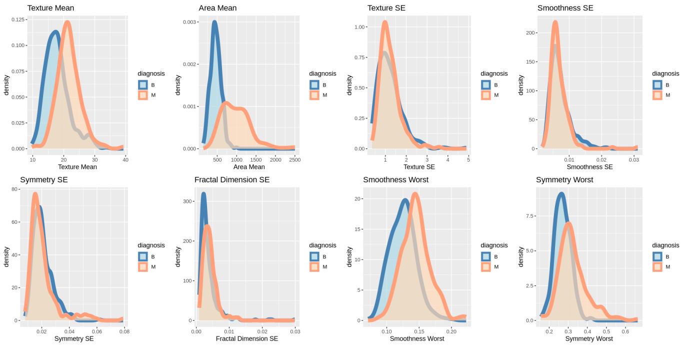


Fig. 6: Distributions of correlated features by the Diagnosis column

E. Pair Plots

In Figure 7, a pair plot involving the diagnosis column and all the features such as the radius, the perimeter, the texture etc, that contain information about the mean is presented. From this plot, it can be especially seen, that some columns are linear correlated, for example the area with the perimeter, the perimeter with the radius and the concavity with the radius. There are some features that seem to have a negative linear relationship, for example the area, radius and perimeter with the fractal dimension mean.

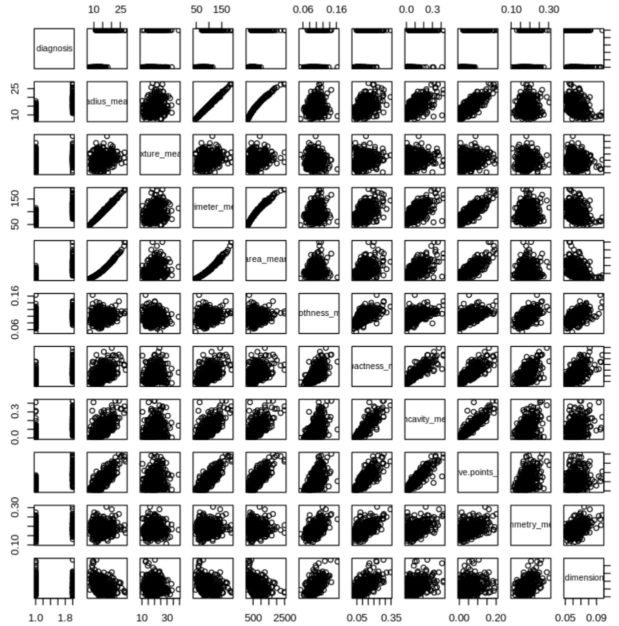


Fig. 7: Pair Plot of diagnosis column with the features that contain information about the mean

In Figure 8, the diagnosis column is plotted with all the features that contain information about the standard error. Once again the perimeter, area and radius have a strong linear relationship. The concavity and compactness have also a clear positive linear relationship. However, there is not any obvious negative linear relationship between the features.

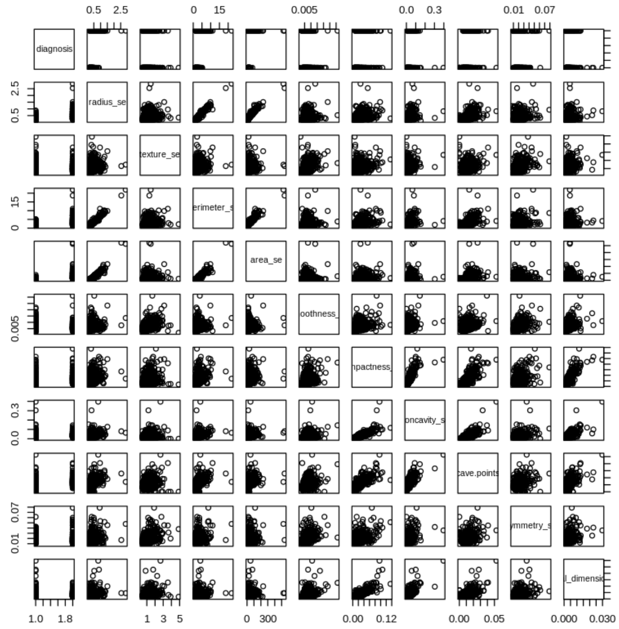


Fig. 8: Pair Plot of diagnosis column with the features that contain information about the standard error

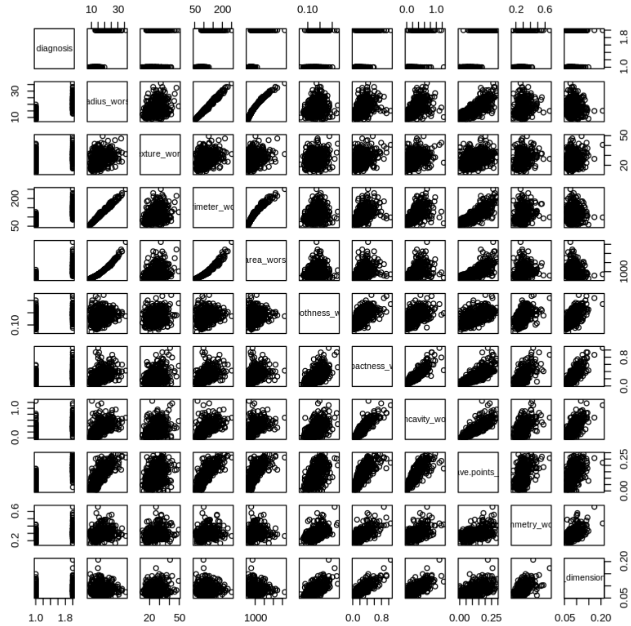


Fig. 9: Pair Plot of diagnosis column with the features that contain information about the worst

Figure 9 showcases a pair plot, which includes the diagnosis column and all the features containing information about the worst values of the measurement. Most of the features seem to have a positive linear relationship. As expected the most clear linear relationships are between the features area, radius and perimeter. However, there are also other high positive linear relationships between features involving the compactness and the concavity, or the concave points and the concavity. On the other hand, again there are not any strong negative relationships between the features.

F. Scaling of the data set

Due to the presence of columns with a wide range of values, including both small and large values, it is generally recommended to standardize the data to ensure that all predictor variables are given equal weight in the models and to improve their performance. The data will be scaled using the preProcess = c(”scale”,”center”) attribute in the train() function in R during a later stage (see section IV.D).

IV. Classification Analysis

A. Feature Selection

The goal of feature selection techniques in machine learning is to find the best set of features that allows one to build optimized models of studied phenomena [4].

1) Feature Importance

Feature Importance refers to techniques that calculate a score for all the input features for a given model — the scores simply represent the “importance” of each feature [5]. Feature importance scores can be calculated for problems that involve predicting a class label, e.g. classification. A higher score means that the specific feature will have a larger effect on the model that is being used to predict a certain variable. Also, it helps to better understand the data and to reduce the number of input features which helps with dimensionality reduction and thus it improves the model. After the application of the importance-based feature selection method (Figure 8), it was discovered that 12 out of all the features are considered the most significant. These features were identified as follows: area_worst, perimeter_worst, radius_worst, concave_points_worst, concave.points_mean, texture_worst, area_se, smoothness_worst, concavity_worst, texture_mean, area_mean and concavity_mean. A threshold of 7.0 is set for choosing the most important ones. Some other features that lie really close to this threshold and may be also important will be cross checked with another method later.

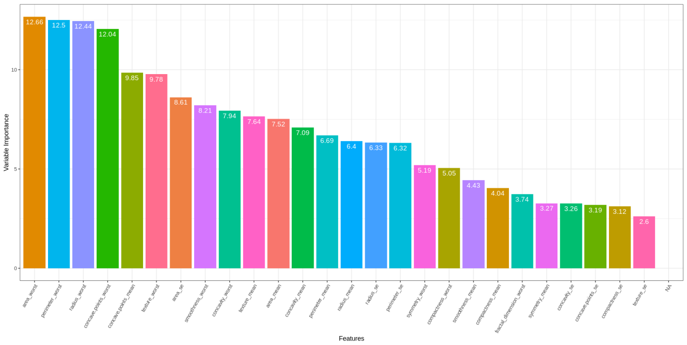


Fig. 10: Variance Importance of the features with the feature importance method

2) *Recursive Feature Elimination*

Recursive Feature Elimination (RFE) is a feature selection method commonly used in machine learning. The basic idea behind RFE is to iteratively eliminate features from a data set and select the subset of features that produces the best performance on a given metric. By recursively eliminating the least important features, RFE aims to identify the most informative subset of features for a given machine learning task. RFE is often used in combination with a wrapper method, where a model is trained and evaluated on different subsets of features to find the best performing subset.

Here, the RFE method is used in combination with a random forest algorithm as the selection function. The rfeControl function defines the settings for the RFE algorithm, such as the selection function, the method of cross-validation, and the number of folds, while the rfe function performs the actual feature selection process. The predictor function is used to print out the selected features that were chosen by the RFE algorithm. The RFE method calculated the number of features based on accuracy and kappa metrics to be 24 out of the 31 columns, as shown below in Figures 11 and 12.

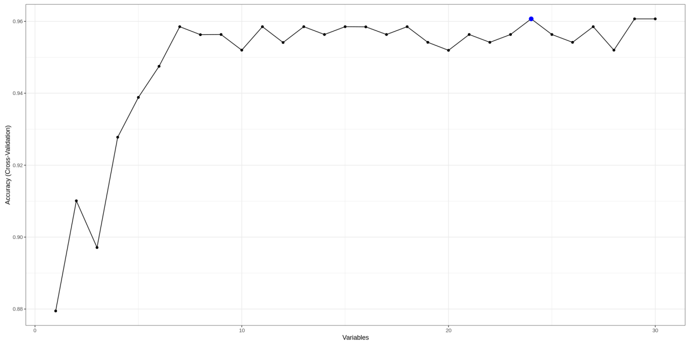


Fig. 11: RFE Method - Accuracy metric

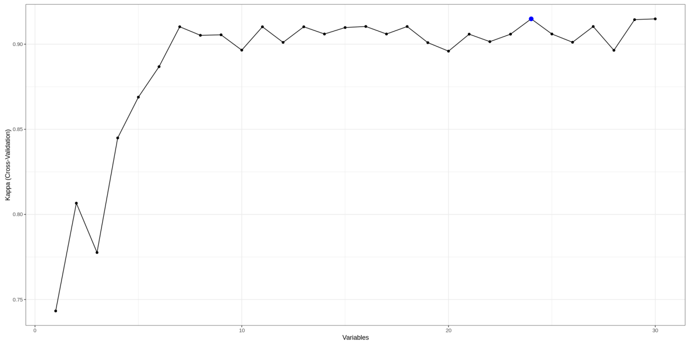


Fig. 12: RFE Method - Kappa metric

The features that were selected are the area_worst, perimeter_worst, radius_worst, concave.points_worst, concave.points_mean area_worst, perimeter_worst, radius_worst, concave.points_worst, concave.points_mean, texture_worst, area_se. smoothness_worst, concavity_worst, texture_mean, area_mean, concavity_mean, perimeter_mean, radius_mean, radius_se, perimeter_se, symmetry_worst, compactness_worst, smoothness_mean, compactness_mean, concavity_se, fractal_dimension_worst, compactness_se, concave.points_se.

However, as it can be seen from the figures above, even if 10 features are chosen, still the accuracy is high, really close to the optimal one.

B. *Principal Components Analysis*

As a part of the feature extraction, Principal Component Analysis (PCA) is applied. PCA is responsible for finding a reduced number of features that would represent the original data set in a compressed way, capturing up to a certain portion of its variance depending on the number of new features that its end up selecting. After the PCA procedure is applied, the explained_variance_ ratio attribute showed how much of the variance in the original data was encapsulated in the new component variables, as shown in Figure 13. Furthermore, in Figure 14 the principal components are plotted with the cumulative variance explained. Based on this, the top principal components that explain at least 90% of the variance in the data are selected. The number of components ended up to be 7. The resulting data set includes the target variable and the selected principal components, which can be used for the predictive modeling.

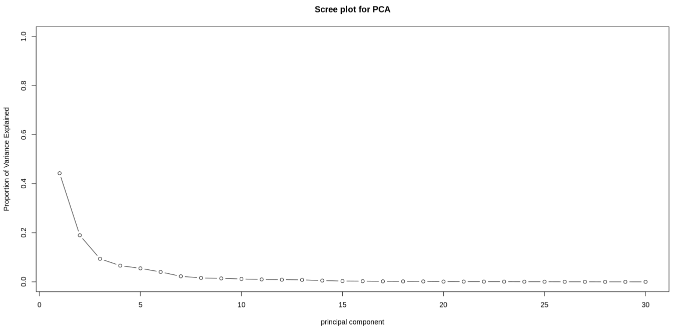


Fig. 13: PCA - Proportion of variance explained

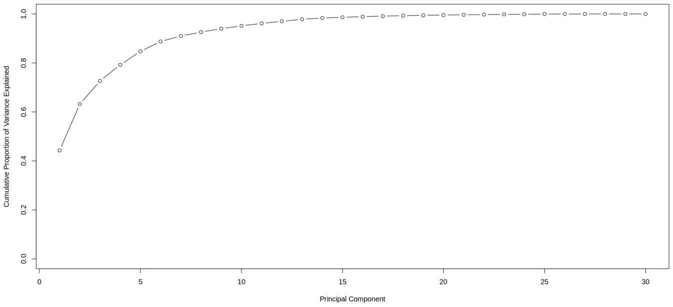


Fig. 14: PCA - Cumulative proportion of variance explained

C. *Evaluation Metrics*

1) Accuracy: The accuracy measures the overall performance of the model in terms of the proportion of correct predictions out of the total predictions. It is computed as the ratio of the number of correct predictions to the total number of predictions. [6]
2) Precision: Precision measures the proportion of true positive predictions out of all positive predictions made by the model. It is computed as the ratio of the number of true positive predictions to the total number of positive predictions made by the model. [7]
3) Recall (Sensitivity): Recall measures the proportion of true positive predictions out of all actual positive instances in the data set. It is computed as the ratio of the number of true positive predictions to the total number of positive instances in the data set. [7]
4) F1 Score: F1 score is the harmonic mean of precision and recall. It is a useful metric when the classes are imbalanced in the dataset. It takes into account both precision and recall and provides a single score that balances the trade-off between them. It is computed as 2 x (precision x recall) / (precision + recall). [8]

These metrics are commonly used to evaluate the performance of classification models. While accuracy is a good metric to use when the classes are balanced, it can be misleading when the classes are imbalanced. In such cases, it is often useful to look at precision, recall and F1-Score to get a better understanding of how the model is performing. In this case, it would be better to look more at recall and F1-Score, whereas the accuracy and precision could help to be more sure for the previous decisions.

D. *Classification Algorithms*

For predicting the Diagnosis, various models are used for the purpose of classification. The models that were applied are the Logistic Regression, the Random Forest (a meta estimator that fits a number of decision tree classifiers on various subsamples of the data set and uses averaging to improve the predictive accuracy and control over-fitting), the k-Nearest Neighbors, Linear Discriminant Analysis, as well as Quadratic Discriminant Analysis, Stepwise Logistic Regression (a step-by-step iterative construction of a regression model that involves the selection of independent variables to be used in the final model) and lastly Naive Bayes. In the first step the models were trained with a 10-fold cross validation on the original data. The same, was done also for the scaled data set and the results were the same therefore it was decided to use the original one.

TABLE II: Metrics of the models with 10-fold cross validation

Models	**Accuracy**	**Precision**	**Recall**	**F1-Score**
Logistic Regression	0.9292	0.9474	0.8571	0.9000
Random Forest	0.9469	0.9500	0.9048	0.9268
KNN	0.9292	0.9722	0.8333	0.8974
LDA	0.9204	0.9714	0.8095	0.8831
QDA	0.9646	0.9750	0.9286	0.9512
Stepwise Logistic Regression	0.9558	0.9744	0.9048	0.9383
Naive Bayes	0.9381	0.9268	0.9048	0.9157

The Breast Cancer data set contains information about medical data therefore all the inserted values didn't need any modifications such as cleaning, filling missing values etc., in the preprocessing phase. As a result, all the above models achieve already very good results and even an improvement in the order of 0.001 is still a progress for classifying the cancerous from the non cancerous tumors. Among the models, the best model that gave the highest score based on the F1-Score (Table 2) is the Quadratic discriminant analysis with a score of 0.9512, following the Random Forest, the Stepwise Logistic Regression, the Naive Bayes and last the logistic regression with the LDA and the KNN.

Scaling was not used in these models. However, by using the preProcess = c(”scale”,”center”) attribute in the train() function in R, scaling was applied and the same results were obtained, indicating that scaling did not affect the results.

For a further examination, the data set was split into train and test and applied on the same classification algorithms as before. Here again the same classification metrics were calculated and presented below.

TABLE III: Metrics of the models with a split into train and test set

Models	Accuracy	Precision	Recall	F1-Score
Logistic Regression	0.9182	0.9571	0.9178	0.9371
Random Forest	0.9818	0.9863	0.9863	0.9863
KNN	0.9545	0.9595	0.9726	0.9660
LDA	0.9545	0.9474	0.9863	0.9664
QDA	0.9455	0.9718	0.9452	0.9583
Stepwise Logistic Regression	0.9364	0.9583	0.9452	0.9517
Naive Bayes	0.9364	0.9583	0.9452	0.9517

In comparison with the previous results, using a 10-fold cross validation the F1-Score was increased where the accuracy slightly decreased for all the models. Random forest gave the highest F1 score of 0.9863 where there is a difference of 0.0595 with the results using cross validation. It might seem to be a very small, however since the data set contains information about medical data (if a person has cancer or not), even the slightest difference plays a very important role.

Moreover, Figure 15 visual represents the LDA model of the two predicted values. At some points, it can be seen that the two distributions are overlapping although they are well separated. LDA achieves the second highest F1-Score when splitting the data set into test and train.

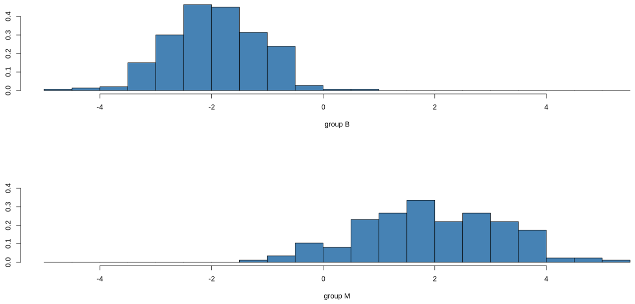


Fig. 15: Fit of the LDA Model with train and test sets

In addition the models were trained again with a 10-fold cross validation with the features that were selected with the RFE method, explained in the previous section. Now the selected columns that were used in the models are:

area_worst, perimeter_worst, radius_worst, concave.points_worst, concave.points_mean, texture_worst, area_se, smoothness_worst, concavity_worst, texture_mean, area_mean, concavity_mean, perimeter_mean, radius_mean, radius_se, perimeter_se, symmetry_worst, compactness_worst, smoothness_mean, compactness_mean, concavity_se, fractal_dimension_worst, compactness_se, concave.points_se

TABLE IV: Metrics of the models with 10-fold cross validation and feature selection

Models	Accuracy	Precision	Recall	F1-Score
Logistic Regression	0.9381	0.9730	0.8571	0.9114
Random Forest	0.9558	0.9744	0.9048	0.9383
KNN	0.9292	0.9474	0.8571	0.9000
LDA	0.9204	0.9714	0.8095	0.8831
QDA	0.9558	0.9744	0.9048	0.9383
Stepwise Logistic Regression	0.9558	0.9512	0.9286	0.9398
Naive Bayes	0.9381	0.9268	0.9048	0.9157

By some models such as the KNN, the LDA and QDA nothing changed or some metrics even decreased with a feature selection. By the the Logistic Regression and the Random Forest there was an increase in all the classification metrics in comparison with the 10-fold cross validation using all of the features. The F1-Score of the Random Forest increased to 0.9383 whereas for the logistic Regression to 0.9114 (in comparison with the initial results, see Table 2). The best model with a feature selection according to F1-Score is the Stepwise Logistic Regression with a value of 0.9398. Here the feature selection only had a positive impact on some models where for others the metrics decreased.

Lastly applying the PCA on the data set again the same models with a 10-fold cross validation were trained. The results follow on Table 5:

TABLE V: Metrics of the models with 10-fold cross validation and PCA

Models	Accuracy	Precision	Recall	F1-Score
Logistic Regression	0.9646	0.9535	0.9535	0.9535
Random Forest	0.9469	0.9744	0.8837	0.9268
KNN	0.9381	0.9737	0.8605	0.9136
LDA	0.9469	0.9744	0.8837	0.9268
QDA	0.9469	0.9302	0.9302	0.9302
Stepwise Logistic Regression	0.9646	0.9535	0.9535	0.9535
Naive Bayes	0.9204	0.9250	0.8605	0.8916

By applying the PCA, the F1-Score for some of the models decreased (such as Naive Bayes) whereas for others increased in comparison with the results of Table 2. Now comparing the results of PCA with the results of the feature selection, it is obvious that Stepwise Logistic Regression achieves in both cases the highest F1-Score. For the models such as the Logistic Regression, the KNN, the LDA and the Stepwise Logistic Regression, PCA gave better results. On the other hand, for the rest the F1-Score decreased using PCA.

V. Discussion

A. Hierarchical presentation of results

This section is divided into results of primary and of secondary significance.

1) Primary Significance

TABLE VI: Best two models

Models	Accuracy	Precision	Recall	F1-Score
Random Forest with train and test data sets	0.9818	0.9863	0.9863	0.9863
LDA with train and test data sets	0.9545	0.9474	0.9863	0.9664

As it can be seen from the above table, the best model used on this data set is Random Forest with a split on the data set to train and test where it has achieved a F1-Score of 0.9863. Moving on, the 2nd best model is LDA with a split on the data set to train and test where it has achieved a score equal to 0.9535.

2) Secondary Significance

Firstly, seeing the distributions of the features, it was noticed that most of them were heavy tailed right skewed distributions. The percentage of outliers ranges between 0-11% and as it was detected some of the features carry the similar information such as the perimeter, the area or the radius which is self-explainable. After performing a feature selection using the variance importance method, it is obvious that, the most important features are area_worst, perimeter_worst, radius_worst, concave.points_worst, concave.points_mean, area_worst, perimeter_worst, radius_worst, concave.points_worst, concave.points_mean, texture_worst, area_se. smoothness_worst, concavity_worst, texture_mean, area_mean, concavity_mean, perimeter_mean, radius_mean, radius_se, perimeter_se, symmetry_worst, compactness_worst, smoothness_mean, compactness_mean, concavity_se, fractal_dimension_worst, compactness_se, concave.points_se. On the other hand, using PCA seven components are selected, with a combination of the features, as the most important. The best model according to PCA was both the Stepwise Logistic Regression and the Logistic Regression with a F1-Score of 0.9535 resulting to the second best models as mentioned above.

B. Comparison of results with other analysis from Kaggle

Comparing with the results in Kaggle (Table 7), it is evident that Random Forest with a feature selection achieves a higher accuracy (Table 6) of 0.9818. The results have a very small difference of order 0.01. KNN and Naive Bayes are performing better in Kaggle than the results obtained with a 10-fold cross validation (see Table 2). Logistic Regression and Random Forest on the other hand as they were calculated in Kaggle, give higher scores in terms of accuracy.

TABLE VII: Results from Kaggle

Models	Accuracy
Logistic Regression	0.9883
KNN	0.9591
Naive Bayes	0.9357
Random Forest	0.9708

VI. Conclusion

To conclude, through the analysis, it was able to classify the tumors in benign or malignant, using machine learning algorithms. Through the exploratory data analysis, correlations between the variables where identified that helped understanding the data and give interpretations to the results. The diagnosis was predicted with various classification models, with the Random Forest model achieving the highest F1-Score of 0.9863. Overall, the performance of the final model gave very good results thus, the aim to achieve a high score and make accurate predictions was accomplished.

References

[1] "Breast cancer symptoms and causes-mayo clinic," https://www.mayoclinic.org/diseases-conditions/breast-cancer/symptoms-causes/syc-20352470.

[2] "Breast cancer overview: Causes, symptoms, signs, stages types," https://my.clevelandclinic.org/health/diseases/3986-breast-cancer .

[3] "Breast cancer data set from kaggle," https://www.kaggle.com/code/hsniyesakmak/breast-cancer-wisconsin-diagnostic/data?select=data.csv .

[4] "Feature selection techniques in machine learning," https://www.analyticsvidhya.com/blog/2020/10/feature-selection-techniques-in-machine-learning/ .

[5] "Understanding feature importance and how to implement it in python," https://towardsdatascience.com/understanding-feature-importance-and-how-to-implement-it-in-python-ff0287b20285 .

[6] "Classification metric accuracy," https://en.wikipedia.org/wiki/Accuracy_and_precision.

[7] "Classification metric precision and recall," https://en.wikipedia.org/wiki/Precision_and_recall.

[8] "Classification metric f1 score," https://en.wikipedia.org/wiki/F-score.