**Wroclaw University of Science and Technology**

**Faculty of Pure and Applied Mathematics**

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

***Working Title:***

**BREAST CANCER PREDICTION USING ROC BASED FEATURE FILTERING: A COMPARATIVE STUDY OF CLASSIFICATION METHODS**

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**Course Title: Data Mining**

**Lecturer: Ph.D. Adam Zagdanski**

**Authors:**

**SEGUN LIGHT JEGEDE – 257389**

**ISAAC AKOJI PAUL – 257388**

**November 25, 2020**

1. **INTRODUCTION**

\*\*\*\*\*\*\*\*\*\*Remember to state research objectives\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

1. **DATA CHARACTERISTICS**

The dataset is a life data, a multivariate with attribute character, integer. The data consists of 699 rows of observation and twelve (12) columns of variables. The first column is the sample id and the last is the classification observation for the cancer cell type. Thus, we can say that there are ten (10) number of attributes and 699 number of instances. Nine (9) of the attributes are the features which will be used as the independent variables, with 6291 observations. The nine (9) features are the characteristics of the breast cancer cell and are described in order of appearance in the data column as (a) Clump Thickness (b) Uniformity of Cell Size (c) Uniformity of Cell Shape (d) Marginal Adhesion (e) Single Epithelial Cell Size (f) Bare Nuclei (g) Bland Chromatin (h) Normal Nucleoli (i) Mitoses. The measurement of these features are scaled from 1 – 10 where value 1 is closet to Benign and 10 is closest to Malignant. The last column is the grouping variable named “class”, which categorize the cell as either “Benign” or “Malignant” with index number “2” and “4” respectively.

**Table 1. Features Description of the Breast Cancer Cell Data**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **S/No** | **Attributes** | **Kind** | **Description** | **Variables of Attribute** |
| 1 | Sample ID | Categorical | Sample Identification |  |
| 2 | Clump Thickness | Numeric (integers) | Clump Thickness | Value between 1 - 10 |
| 3 | Uniformity of Cell Size | Numeric (integers) | Uniformity of Cell Size | Value between 1 - 10 |
| 4 | Uniformity of Cell Shape | Numeric (integers) | Uniformity of Cell Shape | Value between 1 - 10 |
| 5 | Marginal Adhesion | Numeric (integers) | Marginal Adhesion | Value between 1 - 10 |
| 6 | Single Epithelial Cell Size | Numeric (integers) | Single Epithelial Cell Size | Value between 1 - 10 |
| 7 | Bare Nuclei | Numeric (integers) | Bare Nuclei | Value between 1 - 10 |
| 8 | Bland Chromatin | Numeric (integers) | Bland Chromatin | Value between 1 - 10 |
| 9 | Normal Nucleoli | Numeric (integers) | Normal Nucleoli | Value between 1 - 10 |
| 10 | Mitoses | Numeric (integers) | Mitoses | Value between 1 - 10 |
| 11 | Class | Binary | Class of Tumor | ‘benign’, ‘malignant’ |

Key: for attribute values, 1 – 10, value 1 is closet to Benign and 10 is closest to Malignant.

We had to format the datatype of the attributes. We transformed the “sample id” from integer to a string since it is just an identification tag. The “Bare Nuclei” measurement was initially in a string format and we had to convert it to numeric. The grouping variable was also in integer form; we converted it to factors and named the factors appropriately as either “Benign” or “Malignant”. We observed that 16 observations are missing from the “Bare Nuclei” column, which is about 2.29% of the observation in the column. We also observed that the missing value was due to the conversion of this feature from string to numeric and we decided to handle this event by removing the missing observations’ row. Thus, we have 16 less rows of observation, that is, 683 number of instances. The data description is summarized in Table 1.

1. **METHODOLOGY**

The Breast Cancer Wisconsin dataset is a popular medical dataset for a classification task. Thus, we considered two approaches of analysis – the exploratory data analysis and classification. The task pose by the dataset used is a binary classification task, that is, we have two classes to be classified from the nine (9) features. However, we first visualize the data with some plots such as bar plot, histogram and boxplot. These plots are used to verify the distribution of the data, check for outlier and we construct a correlation matrix to check for correlation between features. For a more robust analysis due to the non-normal of the data, we employed a normalization procedure.

Data normalization is carried out in order to standardize the selection of features that individually contribute to the classification of the test sample. It was done to decrease the dominance effect of a particular feature in the output of the classifier. We used the min-max normalization, an alternative method to standardization that normalize all the features within [0, 1]. We consider the fact that the features are distance mark between two groups and a transformation that affects the direction of the observation may change important characteristics of the data, thus, our reason for selecting the min-max approach. The transformation is done with the mathematical formula below:

where and are the lower and upper boundaries of .

We proceed to feature selection using the Receiver Operating Characteristic (ROC)-based feature selection procedure, the features were ranked in the order of importance and we construct the classification procedure for two different sets. The first set involves all the nine (9) features while the second set involves the top five (5) most important features.

For classification, we explore and compare the accuracy of five (5) different methods. These classification methods are Linear Regression, Logistic Regression, K Nearest Neighbours (K-NN) Algorithm, Linear Discriminant Analysis (LDA) and Quadratic Discriminant Analysis (QDA). We explore the confusion matrix and consider different statistics for accuracy measurement and verification. We define this statistics in terms of the two classes of tumor, we consider “Benign” as “Positive” and “Malignant” as “Negative”. Thus, we spell the definition as:

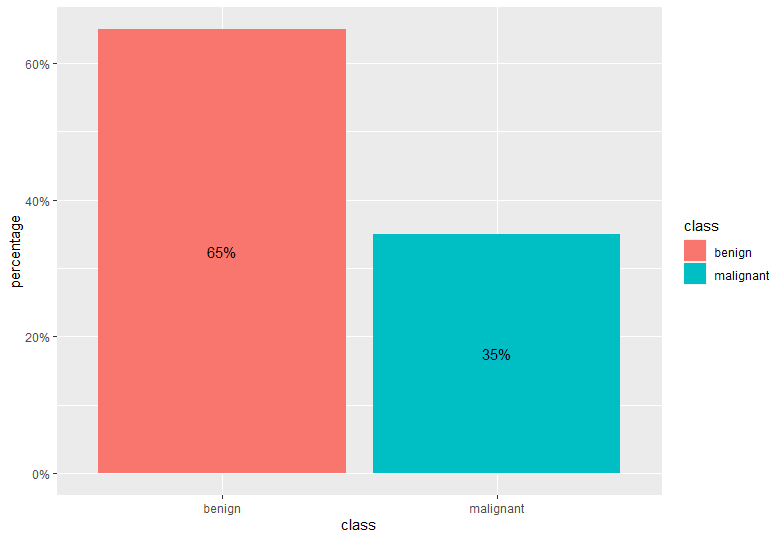
* Accuracy – the ratio measure of correct predictions to the total number of input samples.
* Misclassification error – the ratio measure of wrong predictions to the total number of input samples.
* Sensitivity – measure of benign class that are correctly classified as benign, with respect to all benign class. It is also called the True Positive Rate – “True Benign Rate”.
* Specificity – measure of malignant class that are correctly classified as malignant, with respect to all malignant. It is also called the True Malignant Rate – “True Malignant Rate”.
* Precision – measure of part of predicted benign class that are truly benign. It is also called the Positive Predicted Value – “Benign Predicted Value”.
* False Positive Rate – measure of malignant class that are mistakenly classified as malignant, with respect to all malignant. In accordance to the tumor class, it can be called False Benign Rate.

Lastly, we use the ROC curve to further check the accuracy of the classification methods by calculating the Area Under Curve (AUC); the closer the value of the area to 1, the better the classification method.

1. **RESULTS**

**4.1 Exploratory Data Analysis**

The grouping variable as earlier mention contains two groups, that is, two classes of cancer cells - “Benign” and “Malignant”. 65% of the cells are “Benign” and 35% are “Malignant” – Figure 1.



**Figure 1. Distribution of the Class of Cell**

We summarize the data with selected statistics in Table 3.

**Table 2a. Descriptive Statistics of Breast Cancer Cell Features**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Statistics** | **Clump Thickness** | **Uniformity of Cell Size** | **Uniformity of Cell Shape** | **Marginal Adhesion** |
| Min. | 1.0000 | 1.0000 | 1.0000 | 1.0000 |
| Q1 | 2.0000 | 1.0000 | 1.0000 | 1.0000 |
| Median | 4.0000 | 1.0000 | 1.0000 | 1.0000 |
| Mean | 4.4422 | 3.1508 | 3.2152 | 2.8302 |
| Q3 | 6.0000 | 5.0000 | 5.0000 | 4.0000 |
| Max. | 10.0000 | 10.0000 | 10.0000 | 10.0000 |
| Var. | 7.9567 | 9.3951 | 8.9316 | 8.2057 |
| SD | 2.8208 | 3.0651 | 2.9886 | 2.8646 |
| IQR | 4.0000 | 4.0000 | 4.0000 | 3.0000 |

**Key:** Min. – Minimum, Q1 – First Quantile, Q3 – Third Quartile, Max. – Maximum, Var. – Variance,

SD – Standard Deviation, IQR – Inter Quantile Range.

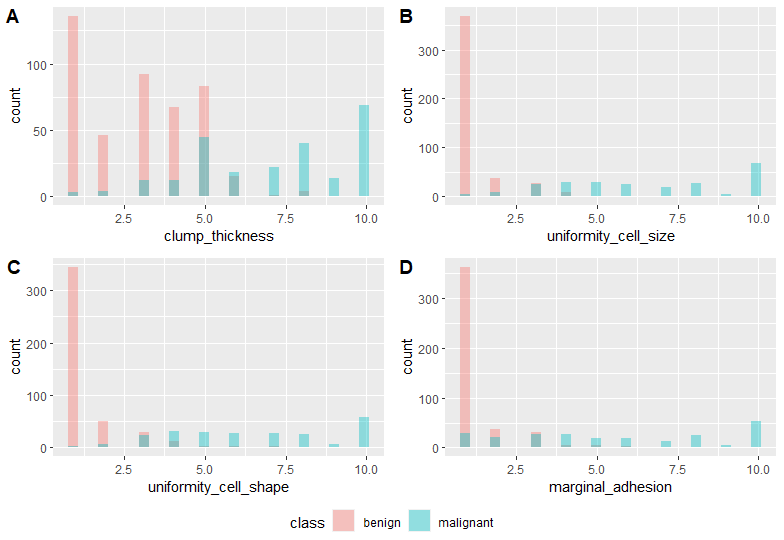
**Table 2b. Descriptive Statistics of Breast Cancer Cell Features**

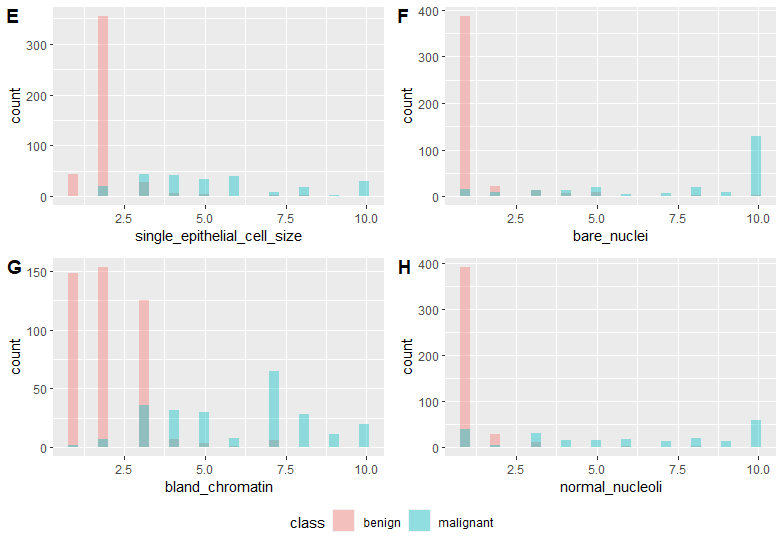
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Statistics** | **Single Epithelial Cell Size** | **Bare Nuclei** | **Bland Chromatin** | **Normal Nucleoli** | **Mitoses** |
| Min. | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 |
| Q1 | 2.0000 | 1.0000 | 2.0000 | 1.0000 | 1.0000 |
| Median | 2.0000 | 1.0000 | 3.0000 | 1.0000 | 1.0000 |
| Mean | 3.2343 | 3.5447 | 3.4451 | 2.8697 | 1.6032 |
| Q3 | 4.0000 | 6.0000 | 5.0000 | 4.0000 | 1.0000 |
| Max. | 10.0000 | 10.0000 | 10.0000 | 10.0000 | 10.0000 |
| Var. | 4.9421 | 13.2777 | 6.0010 | 9.3188 | 3.0022 |
| SD | 2.2231 | 3.6439 | 2.4497 | 3.0527 | 1.7327 |
| IQR | 2.0000 | 5.0000 | 3.0000 | 3.0000 | 0.0000 |

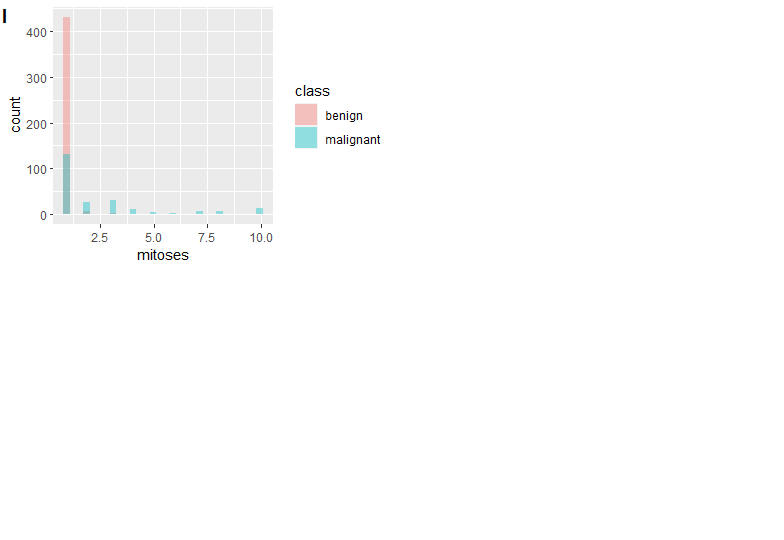
**Key:** Min. – Minimum, Q1 – First Quantile, Q3 – Third Quartile, Max. – Maximum, Var. – Variance,

SD – Standard Deviation, IQR – Inter Quantile Range.

Two of the important statistics to watch out for in Table 2 are the minimum and the maximum. This is because, the presence of an observation lesser than 0 or an observation greater than 1 must be counted as an error based on the earlier stated range of the features, that is, between 1 – 10.

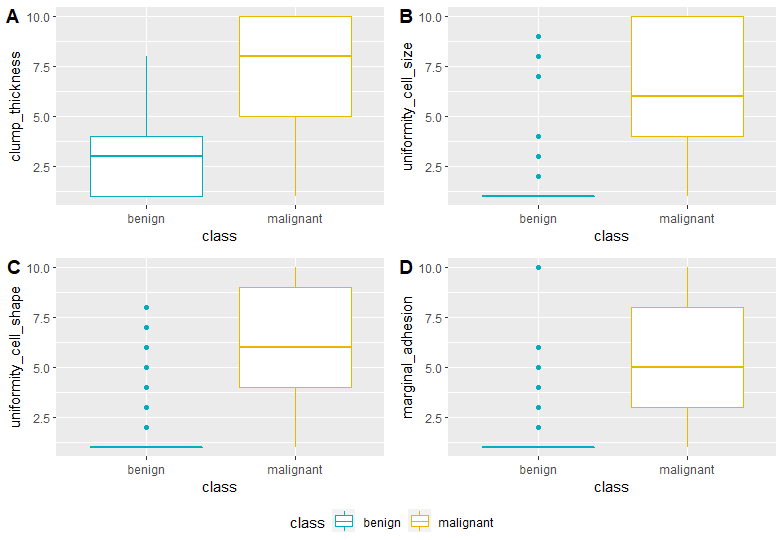


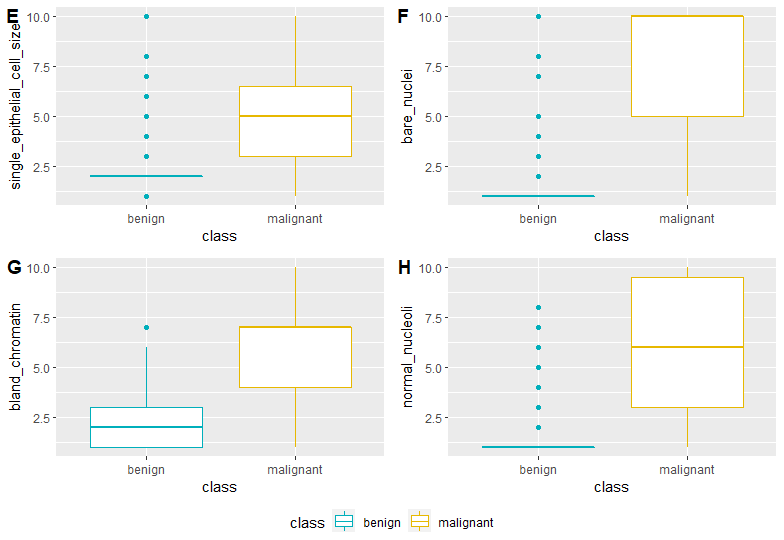


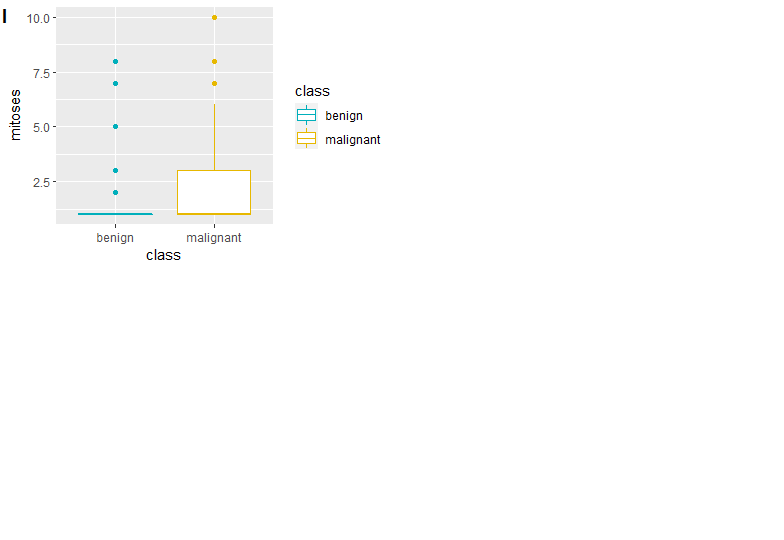


**Figure 2. Distribution of Features of the Breast Cancer Cells**

Based on the distribution of the data in the histogram (Figure 2), we observed that the features are skewed to the right with uniformity of cell size, uniformity of cell shape, marginal adhesion, single epithelial cell size, bare nuclei, normal nucleoli and mitoses are very much skewed to the right. This implies that the observed distances are more distributed to towards the benign class of cancer cell. A look at the boxplot of the features, we can observed the presence of outliers in our data especially for the benign class (Figure 3).

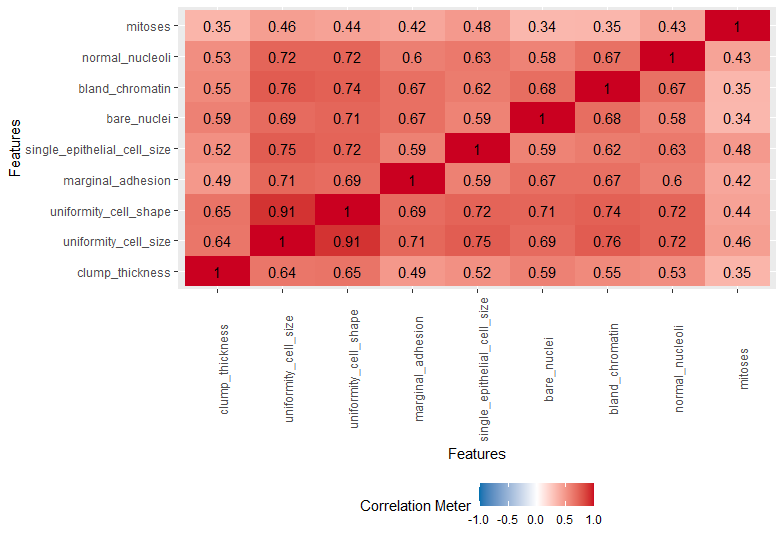






**Figure 3. Boxplot of Features of the Breast Cancer Cells by Class**

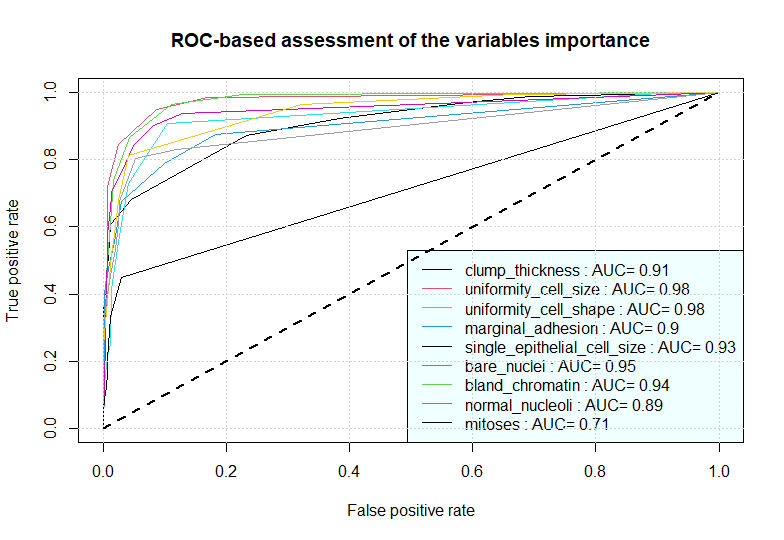
We checked if there is correlation between the variables that we are going to consider for the classification. We discovered a strong positive correlation (> 0.5) between the variables except for mitoses that has a weak positive correlation with the other features. Due to all the non-normal distribution of the data noticed in the features through visualization, we employed the min-max normalization procedure.



**Figure 4. Correlation Matrix of Cancer Cell Features**

* 1. **Feature Selection**

Based on feature rating according to importance using the ROC-based selection method, we observed uniformity of cell size and uniformity of cell shape to be very important, followed by clump thickness, bare nuclei, bland chromatin, single epithelial cell size, clump thickness and normal nucleoli respectively (Figure 5). Moreover, we form subset of variable in the order of importance (Table 3) but we use all the features (set number 8) as “Subset 1” and the top five as subset (set number 4) as “Subset 2”.



**Figure 5. ROC-based Assessment of the Feature Importance**

**Table 3. Formulated Feature Set based on the ROC Assessment**

|  |  |
| --- | --- |
| **Set Number** | **Feature Set** |
| 1 | Uniformity of Cell Size, Uniformity of Cell Shape |
| 2 | Uniformity of Cell Size, Uniformity of Cell Shape, Clump Thickness |
| 3 | Uniformity of Cell Size, Uniformity of Cell Shape, Clump Thickness, Bare Nuclei |
| 4 | Uniformity of Cell Size, Uniformity of Cell Shape, Clump Thickness, Bare Nuclei, Bland Chromatin |
| 5 | Uniformity of Cell Size, Uniformity of Cell Shape, Clump Thickness, Bare Nuclei, Bland Chromatin, Single Epithelial Cell Size |
| 6 | Uniformity of Cell Size, Uniformity of Cell Shape, Clump Thickness, Bare Nuclei, Bland Chromatin, Single Epithelial Cell Size, Clump Thickness |
| 7 | Uniformity of Cell Size, Uniformity of Cell Shape, Clump Thickness, Bare Nuclei, Bland Chromatin, Single Epithelial Cell Size, Clump Thickness, Normal Nucleoli |
| 8 | Uniformity of Cell Size, Uniformity of Cell Shape, Clump Thickness, Bare Nuclei, Bland Chromatin, Single Epithelial Cell Size, Clump Thickness, Normal Nucleoli, Mitoses |

* 1. **Classification Output of Features**

Figure 6 and Figure 7 presents the confusion matrix for all the five (5) classification methods, which are individually indicated by the colour for subset 1 and subset 2. The 6-NN algorithm presents a more rightly predicted class of tumor for subset 1. For subset 2, the 6-NN algorithm and the quadratic discriminant analysis presents a more rightly predicted class of tumor for subset. However, the 6-NN classifies the benign more while the quadratic discriminant analysis classifies the malignant class more. This is not a standard observation because the two classes of tumor do not have equal size.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  | **Actual** | | **Actual** | |
|  |  | Benign | Malignant | Benign | Malignant |
| **Predicted** | Benign | **436** | **8** | **434** | **11** |
| Malignant | **19** | **220** | **10** | **228** |
| **Predicted** | Benign | **153** | **2** | **154** | **4** |
| Malignant | **3** | **70** | **2** | **68** |
| **Predicted** | Benign |  |  | **148** | **1** |
| Malignant |  |  | **8** | **71** |

**Linear Regression Logistic Regression**

**6-NN Method Linear Discriminant Analysis**

**Quadratic Discriminant Analysis**

**Figure 6. Confusion Matrices for the Methods of Classification for Subset 1**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  | **Actual** | | **Actual** | |
|  |  | Benign | Malignant | Benign | Malignant |
| **Predicted** | Benign | **436** | **8** | **434** | **14** |
| Malignant | **26** | **213** | **10** | **225** |
| **Predicted** | Benign | **141** | **7** | **144** | **12** |
| Malignant | **6** | **74** | **3** | **69** |
| **Predicted** | Benign |  |  | **139** | **5** |
| Malignant |  |  | **8** | **76** |

**Linear Regression Logistic Regression**

**6-NN Method Linear Discriminant Analysis**

**Quadratic Discriminant Analysis**

**Figure 7. Confusion Matrices for the Methods of Classification for Subset 2**

**4.4 Classification Accuracy and Statistics**

Table 4(a & b) presents the effectiveness of the classification methods for feature subset 1 and 2. The 6-NN algorithm achieved the highest value of accuracy while the highest value of AUC is the logistic regression and the linear discriminant analysis (Table 4a).

**Table 4a. Classification Accuracy and Statistics for Subset 1**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Models** | **TPR** | **TNR** | **PPV** | **FPR** | **ACC.** | **Mis. Err.** | **AUC** |
| Linear Regression | 0.9820 | 0.9205 | 0.9582 | 0.0795 | 0.9605 | 0.0395 | 0.9957 |
| Logistic Regression | 0.9753 | 0.9580 | 0.9775 | 0.0420 | 0.9693 | 0.0307 | 0.9963 |
| 6-NN | 0.9871 | 0.9589 | 0.9808 | 0.0411 | 0.9781 | 0.0219 | 0.9878 |
| Linear Discriminant Analysis | 0.9747 | 0.9714 | 0.9872 | 0.0286 | 0.9737 | 0.0263 | 0.9963 |
| Quadratic Discriminant Analysis | 0.9933 | 0.8987 | 0.9487 | 0.1013 | 0.9605 | 0.0395 | 0.9932 |

**Key:** TRP – True Positive Rate (Sensitivity) TNR – True Negative Rate (Specificity)

PPV – Positive Prediction Rate (Precision) FPR – False Positive Rate

ACC – Accuracy Mis. Err. – Misclassification Error

AUC – Area Under Curve

For the case of Subset 2, the highest value of accuracy and AUC were achieved by logistic regression up to 96.5% and 99.4% respectively (Table 4b).

In the case of subset 1, we have a complicated case of selecting between 6-NN Algorithm and the logistic regression based on the two conflicting accuracy measure – Accuracy and Area Under Curve (AUC). Based on literature, the AUC is commonly used for the binary classification task and from the subset 2 with optimal selected features, the logistic regression and the linear discriminant analysis are better classifier for our dataset.

**Table 4b. Classification Accuracy and Statistics for Subset 1**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Models** | **TPR** | **TNR** | **PPV** | **FPR** | **ACC.** | **Mis. Err.** | **AUC** |
| Linear Regression | 0.982 | 0.8912 | 0.9437 | 0.1088 | 0.9502 | 0.0498 | 0.9938 |
| Logistic Regression | 0.9688 | 0.9574 | 0.9775 | 0.0426 | 0.9649 | 0.0351 | 0.9941 |
| 6-NN | 0.9527 | 0.9250 | 0.9592 | 0.0750 | 0.943 | 0.0570 | 0.9839 |
| Linear Discriminant Analysis | 0.9231 | 0.9583 | 0.9796 | 0.0417 | 0.9342 | 0.0658 | 0.9903 |
| Quadratic Discriminant Analysis | 0.9653 | 0.9048 | 0.9456 | 0.0952 | 0.943 | 0.057 | 0.983 |

**Key:** TRP – True Positive Rate (Sensitivity) TNR – True Negative Rate (Specificity)

PPV – Positive Prediction Rate (Precision) FPR – False Positive Rate

ACC – Accuracy Mis. err. – Misclassification Error

AUC – Area Under Curve

**5.0 CONCLUSION**

In this study, we explore and compare five (5) different classification method on the breast cancer data with nine (9) features using the ROC-based feature selection to obtain an optimal subset of five (5) features for our second set of data for classification. For better understanding of the data, we visualize and look at some statistics of the attributes, and we employ min-max normalization for data transformation. We finally apply the classification methods and based on Area Under Curve (AUC), the Logistic Regression was observed to be the best fit.

**6.0 FURTHER RESEARCH SUGGESTION**

The feature selection method used in this study led us to create eight (8) subsets, out of which only two was considered, it may be of interest to explore the other subsets. Much more the number of breast cancer features consider for the classification can be a limitation for this study, the reliability of the study may be further improved by introducing other features to make a larger vector. Secondly, there are various method, which can be used for feature selection such as the filter, wrapper and embedded methods. The dominance-based filtering is another technique that can be considered, and its feature rank is algorithm liberated.

**SOME SOURCES CONSULTED FOR THE STUDY**

[1] <http://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Original)>

[2] <http://www.damienfrancois.be/blog/files/modelperfcheatsheet.pdf>

[3] <https://www.dataschool.io/simple-guide-to-confusion-matrix-terminology/>

[4] Ling C.X., Huang J., Zhang H. (2003) AUC: A Better Measure than Accuracy in Comparing Learning Algorithms. In: Xiang Y., Chaib-draa B. (eds) Advances in Artificial Intelligence. Canadian AI 2003. Lecture Notes in Computer Science (Lecture Notes in Artificial Intelligence), vol 2671. Springer, Berlin, Heidelberg.

[5] <https://datascience.stackexchange.com/questions/806/advantages-of-auc-vs-standard-accuracy>