MACT 4233 - Applied Multivariate Analysis

Project 2 & 3

Discriminant and Clustering Analysis

Dana Salem 900191664

Nour Montasser 900191101

1. Introduction

1.1 Statement of Problem

The increasing risk of breast cancer severity is more for the women in their entire lifetime (SR, S. C., & Rajaguru, H., 2019). In general, the identification of breast cancer is found either by perceiving a lump in the breast or through mammogram screening (Falk et al., 2018). This lump is classified as either benign or malignant tumors. The clinical physicians use microscopic study for the detection and classification of breast cancer conventionally (Documet et al., 2015). But nowadays, machine learning algorithms that utilize the computational techniques pave them an easier way for the analysis and classification of cancer. To elaborate, the use of discriminate analysis and cluster analysis would be effective tools in identifying the type of breast cancer lung among women based on their test results. The sped-up procedure of diagnosis with these algorithms would result in lower mortality rates because the treatment process would begin immediately (Toğaçar, M., Ergen, B., & Cömert, Z., 2020). However, it is important not to compromise the accuracy of the outcome. For that reason, the objective of this project is to perform discriminant analysis and clustering analysis on a breast cancer dataset to validate the reliability of these tools in classifying the patients to their assigned groups and to group patients with similar characteristics.

1.2 Literature Review

Discriminant analysis is a statistical method that can be used to predict the class membership of an observation based on its set of predictor variables. In the context of the breast cancer dataset, we can use this technique to determine which features are more significant in distinguishing among tumor types. Therefore, it can be concluded that the goal of discriminant analysis for breast cancer data is to find a linear combination of the features that best separates the two classes. This can be done using either Fisher Linear Discriminant Analysis (FLDA) or Multinomial Log-Linear Model. Previously, research has been conducted on breast cancer data using the FLDA method (Jessica, 2021). In fact, even though FLDA makes unrealistic assumpitons about the underlying dataset, it has proven in the aforementioned research its effectiveness in detecting whether a set of features is worthwhile in predicting breast cancer. On the contrary, the Multinomial Log-Linear model which is relatively more flexible in itself assumptions has resulted in slightly higher correct classification rates in comparison to the FLDA method (El-Habil & El-Jazzar, 2013).

Besides discriminant analysis, clustering can also be used to explore patterns in the data. To elaborate, it is an unsupervised learning technique that groups similar observations together based on the similarity of their features. In the breast cancer dataset, we can use clustering to identify subgroups of patients that have similar features. This can be useful for identifying potential risk factors for breast cancer, as well as for identifying subtypes of the disease. The clustering techniques used in this project are hierarchical clustering and k-means clustering. The former groups observations into a hierarchy of nested groups. While the latter groups observations into k clusters based on their similarity. Moreover, studies have shown that the k-means method present effective analyses in identifying breast cancer patients (Azevedo et al., 2022). In addition, it helped health professionals in speeding up the diagnosis stage. However, hierarchical clustering can provide more information about the structure of the data than k-means clustering, but it can also be more computationally intensive.

In conclusion, the Breast Cancer Wisconsin (Diagnostic) dataset is a useful dataset for exploring the relationship between breast cancer diagnosis and various features. Discriminant analysis and clustering are two powerful techniques that can be used to gain insights from this dataset. Discriminant analysis can help us identify the most important features for distinguishing between malignant and benign cases, while clustering can help us identify subgroups of patients with similar features. Together, these techniques can help us better understand the factors that contribute to breast cancer and improve our ability to diagnose and treat the disease.

2. Data Description

2.1 Data Source

Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

https://archive.ics.uci.edu/ml/citation_policy.html

2.2 Summary of Data

The Breast Cancer Wisconsin (Diagnostic) Dataset is a publicly available dataset created by Dr. William H. Wolberg from the University of Wisconsin and was donated to the UCI Machine Learning Repository. It contains measurements of the characteristics of breast cancer tumors with no missing values. The 569 observations were gathered from a digitized image of a fine needle aspirate (FNA) of breast masses of female patients diagnosed with breast cancer. There are 12 variables in this dataset 2 of which are nominal and categorical and the remaining are quantitative.

This dataset will be divided into two classes: malignant patients and benign patients. The description of each variable is given in the below Table 1.0.

2.3 Variables Description

Variable Name	Type	Unit of Measurement	Description
id	Nominal	N/A	This is a unique identifier assigned to each patient.
diagnosis	Categorical	N/A	This is a qualitative variable that identifies the type of the tumor whether it is malignant (M) or benign (B).
radius_mean	Quantitative	Millimeters	This feature measures the mean of distances from the center to points on the perimeter of the tumor. It represents the average size of the tumor.
texture_mean	Quantitative	N/A	This feature measures the mean of gray-scale values in the image of the tumor. It represents the variation in the pixel intensities in the image. It is measured on a scale from 0 to 100.

perimeter_mean	Quantitative	Millimeters	This feature measures the mean perimeter of the tumor, which is the length of its boundary.
area_mean	Quantitative	Millimeters Squared	This feature measures the area of the tumor, which is the total number of pixels inside the boundary. It is measured in square mm.
smoothness_mean	Quantitative	N/A	This feature measures the mean local variation in radius lengths of the tumor. It represents how much the radius of the tumor changes at different points along its boundary. It is measured on a scale from 0 to 1.
compactness_mean	Quantitative	N/A	This feature measures the ratio of the perimeter squared to the area of the

			tumor, minus 1.0. It represents how tightly the tumor is packed together. It is measured on a scale from 0 to 1.
concavity_mean	Quantitative	N/A	This feature measures the mean severity of concave portions of the contour of the tumor. It represents the amount of concavity in the boundary of the tumor. It is measured on a scale from 0 to 1.
concavepoints_mean	Quantitative	N/A	This feature measures the number of concave portions of the contour of the tumor. It represents the number of inwardly-curved sections in the boundary of the tumor. It is measured on a scale from 0 to 1.
symmetry_mean	Quantitative	N/A	This feature measures how symmetric the tumor is. It represents how similar the left and right halves of the tumor are. It is measured on a scale from 0 to 1.

fractaldimension_mean	Quantitative	N/A	This feature measures the mean "coastline approximation" of the tumor. It represents how much the boundary of the tumor is convoluted. It is measured on a scale from 0 to 1.

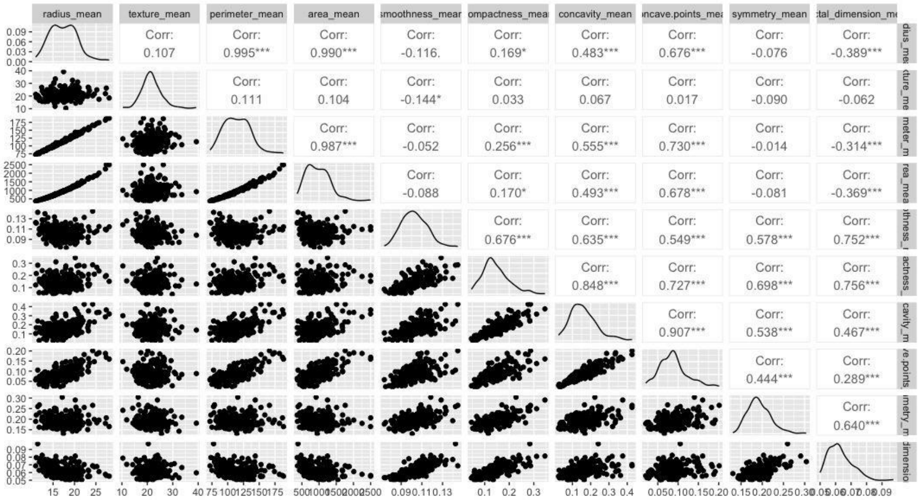
3 Data Analysis

3.1 Linear Discriminant Analysis

3.1.1 Validating Assumptions

A pairwise comparison of the multivariate data is necessary prior to the use of the FLDA algorithm. For that reason, the pairwise scatterplots and the Pearson correlation values. It can be observed from the scatter plots of each group, the benign and malignant, that some variables are not normal. To be more accurate, the QQ-plots for the variables of each group independently are required.

Figure 1.0 Benign group



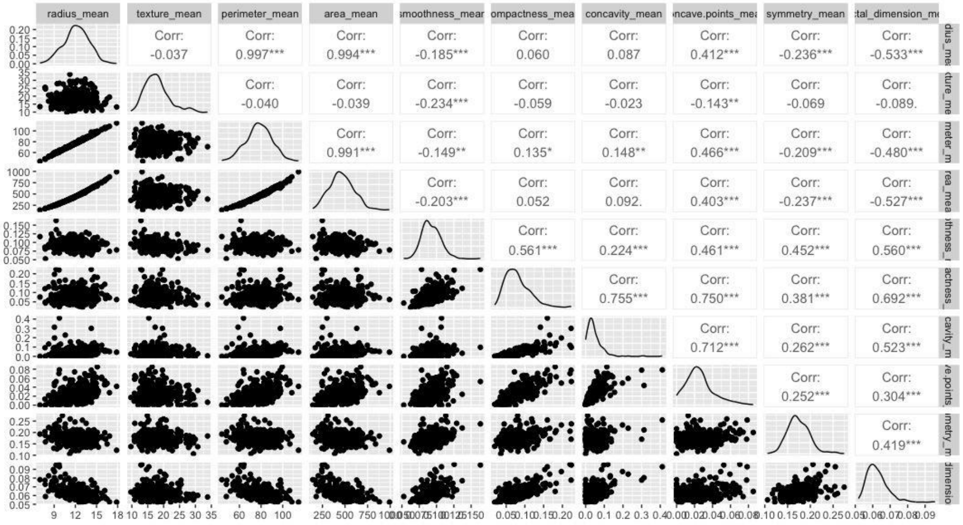
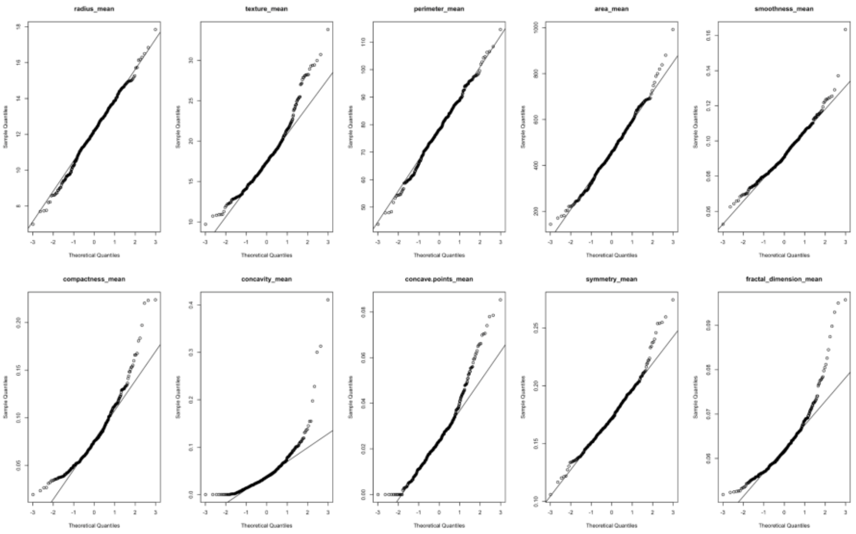


Figure 1.1 Malignant group

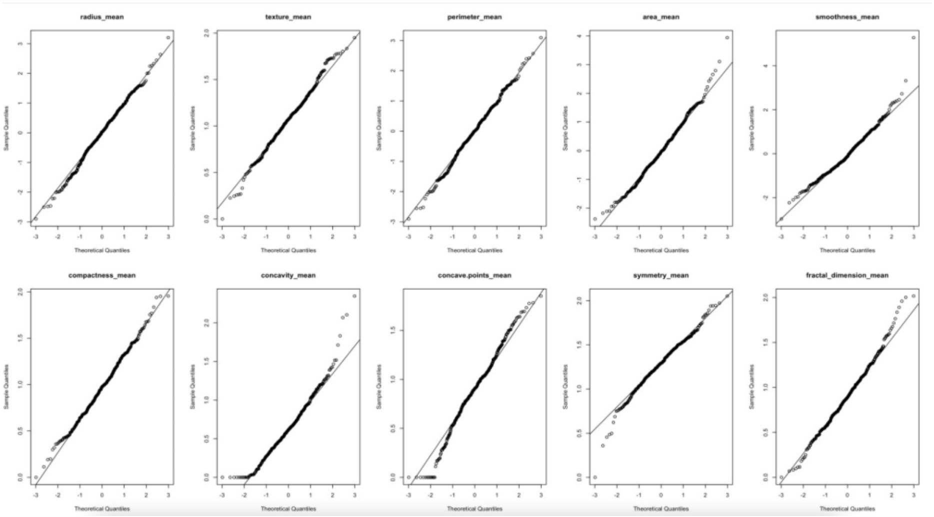
Since normality of the variables in each group is one of the main assumptions of the FLDA method, it had to be examined more closely by plotting the QQ-plots of non-qualitative variables of each group.

Figure 2.0 Benign group before transformation



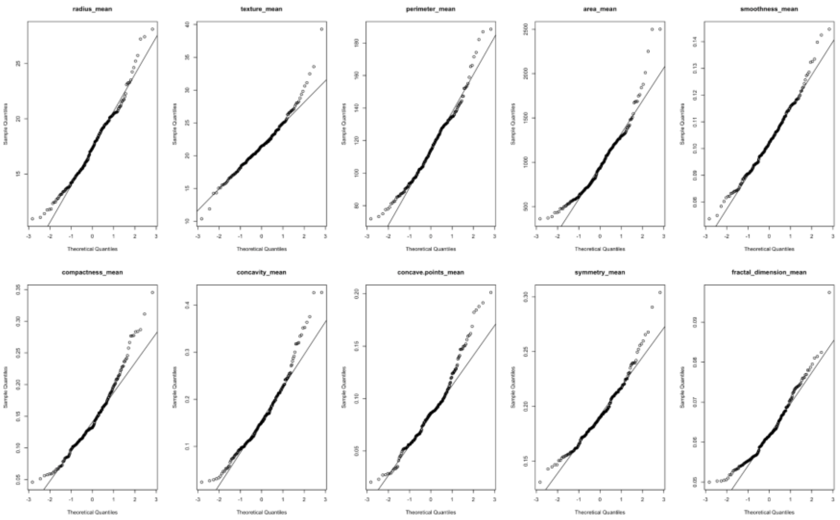
In a benign group, variables are not approximately normally distributed. It appears that only 4 variables which are radius_mean, perimeter_mean, area_mean and smoothness_mean are the most normal variables. The other variables are mostly skewed for that reason the log transformations are required to normalize the other variables.

Figure 2.0 Benign group after transformation



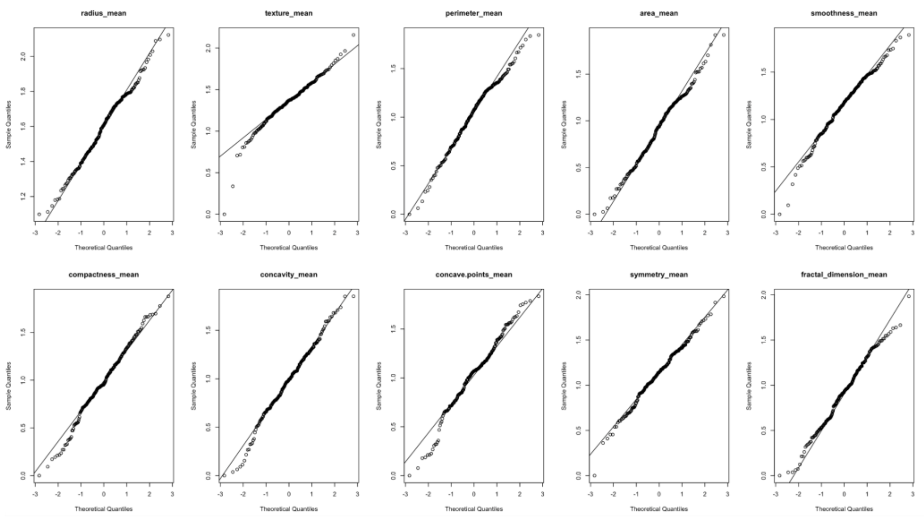
After performing the log-transformation on the benign group, the previously non-normal variables became more normal. For that reason, I would say that the log-transformation was successful.

Figure 3.0 Malignant group before transformation

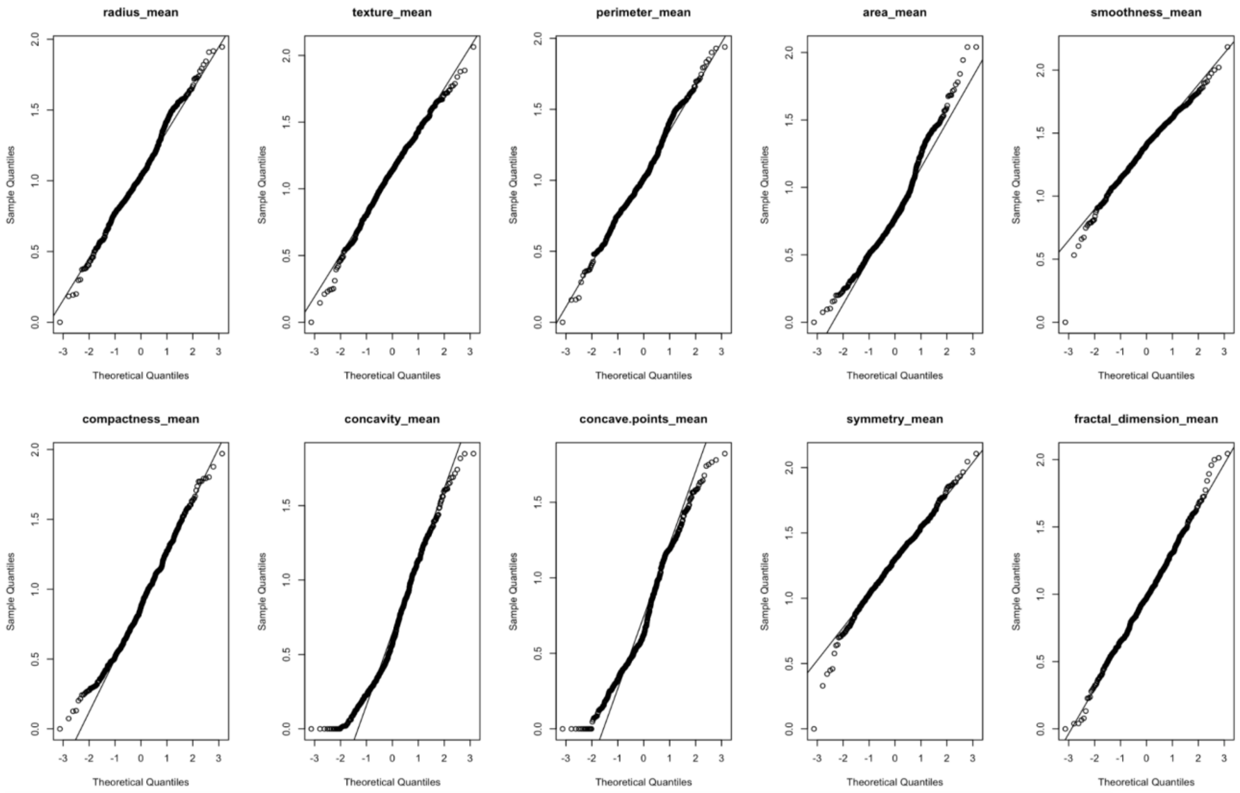


From the above QQ-plot, it is observed that the same 4 variables that were normal in the benign group before log transformation are normal in the malignant group. And the log transformation is required to normalize the other variables.

Figure 3.1 Malignant group after transformation



Once again, I believe that the log transformation was successful in normalizing the skewed variables in the malignant group. Below is a QQ plot of the whole dataset.



The final step prior to performing the discriminant analysis is comparing the covariance matrices of the two groups as they are assumed to be equal in this method. However, this assumption was not validated in the scope of this analysis.

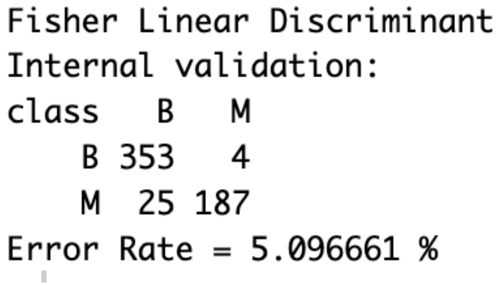
3.1.2 FLDA Results

3.1.2.1 Internal Validation

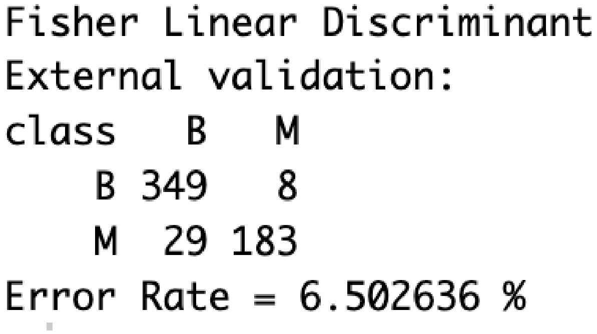
After conducting the Fisher Linear Discriminant analysis, we evaluated the performance of the model on both internal and external validation datasets. The results of the internal validation were promising, with an error rate of 5.096% , indicating that around 95% of the dataset was correctly classified using the FLDA. In addition, metrics such as sensitivity, specificity, precision, and recall showed high performance.

3.1.2.2 External Validation

Using the Leave-One-Out (LOO) cross-validation method, the results of the Fisher Linear Discriminant analysis on the external validation dataset for the Breast Cancer Wisconsin (Diagnostic) Dataset showed an error rate of 6.502%. The results of the external validation were slightly worse relative to the internal validation results, as expected since internal validation usually underestimates the error rate. A closer look at the confusion matrix for the external validation dataset reveals that the model correctly classified 349 out of 357 benign tumors (97.75%) and 183 out of 212 malignant tumors (86.32%). However, the model incorrectly classified 8 benign tumors as malignant and 29 malignant tumors as benign.



```
Fisher Linear Discriminant
Internal validation:
class   B   M
    B 353   4
    M  25 187
Error Rate = 5.096661 %
```



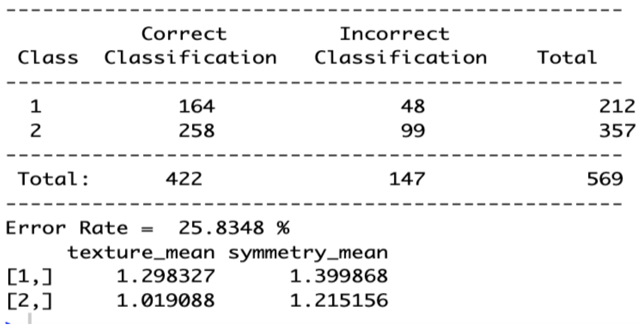
```
Fisher Linear Discriminant
External validation:
class   B   M
    B 349   8
    M  29 183
Error Rate = 6.502636 %
```

Although these error rates might be considered satisfactory, the false positives and false negatives can have serious consequences in medical diagnoses, and thus, it is important to minimize these errors as much as possible.

3.1.3 Fisher Linear Discriminant Analysis for 2 Variables

The FLDA2 is only useful when classification is for two variables. For that reason, the texture_mean and symmetry_mean variables were selected for the implementation of the method. These variables in specific were chosen because they satisfy the assumption of normality which is a requirement for the FLDA. Additionally, these two variables play a scientific role in determining the type of tumor according to numerous studies.

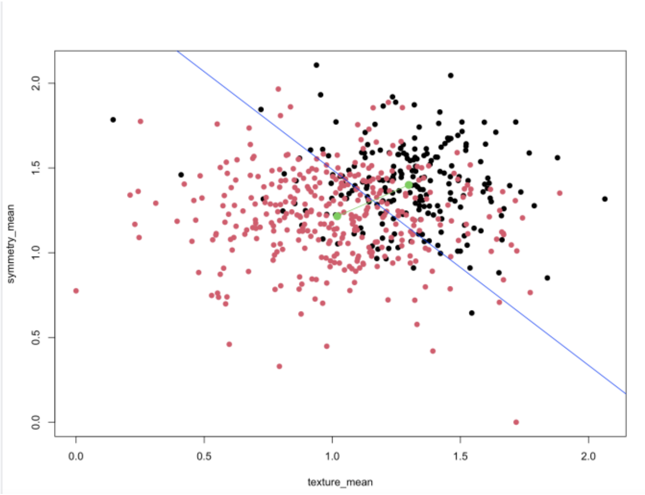
The confusion matrix indicates that a total of 147 observations were correctly classified while 422 observations were correctly classified. In other words, the variables of texture and symmetry produce a rule that results in a 25.834% error rate when internal validation is performed. In comparison to the FLDA, the error rate is extremely higher, as expected, because the FLDA uses



```
--------------------------------------------------
          Correct            Incorrect
 Class  Classification    Classification    Total
--------------------------------------------------
  1            164               48             212
  2            258               99             357
--------------------------------------------------
 Total:      422               147             569
--------------------------------------------------
Error Rate =   25.8348 %
     texture_mean symmetry_mean
[1,]     1.298327      1.399868
[2,]     1.019088      1.215156
```

more variables in identifying the groups into class implying that there is more information to help in classification.

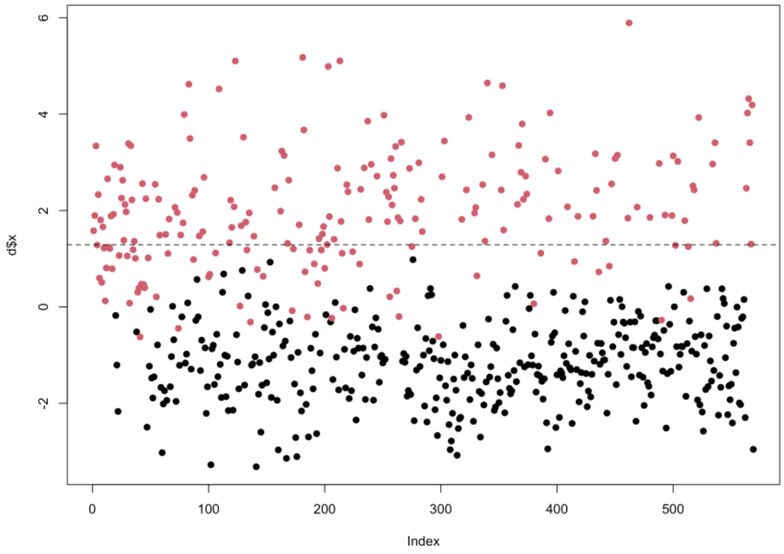
The graph above shows the data separated by two lines. For clarification, the green line connects the means of the two variables. On the other hand, the blue line passes through the midpoint of the means of the two variables. It is noted that the blue line is not orthogonal on the green line. This



implies that the variance covariance matrices of the two groups are not equal. Thus, it can be concluded that the blue line is not a good projection line for the data. In order for it to be a good projection it has to be orthogonal onto the green line.

3.1.4 First Linear Discriminant Projection

The second method of discriminant analysis used in this project is the FIrst Linear Discriminant Projection. This method tends to project the data onto a vector that maximizes the between class dispersion and to minimize the within class dispersion. The figure below displays the projector (the dotted line) that gives the maximum separation between two groups. Moreover, the data consists of only 2 classes, for that reason, only one projector will be needed to separate the data.

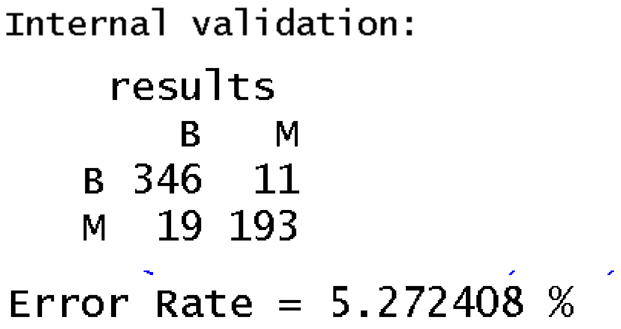


3.1.5 Multinomial

It is observed from the dataset that some of the variables are not quite normally distributed and log-transformations had to be performed in order to make them near normal random variables. However, another model that does not require normal random variables can be implemented to classify the class of each observation. This model is known as the multinomial log-linear model. The R function used to apply this method is called “multinom”.

3.1.5.1 Internal Validation

In comparison to the FLDA internal validation model, the multinomial log-linear internal validation model performed better in terms of correctly classifying the observations. To elaborate, the latter demonstrated a misclassification error rate of 5.272% which is approximately 0.8791% less than the misclassification error under the FLDA model. Moreover, it can be concluded that at least 5% of the data points are susceptible to misclassification depending on the method of discriminant analysis used.

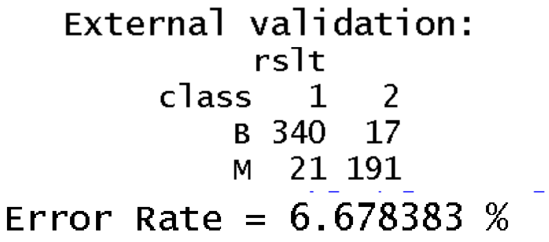


```
Internal validation:
       results
          B   M
    B 346  11
    M  19 193

Error Rate = 5.272408 %
```

3.1.5.2 External Validation

The leave-one-out method was used for external validation. The misclassification rate under the multinomial model is slightly higher by 0.17% compared to the error rate under the FLDA model. In other words, 17 benign cancer patients were falsely classified as malignant and 21 malignant cancer patients were incorrectly classified as benign cancer patients. When comparing the performance of both models, it is noted that the multinomial performs better in the internal validation, while the FLDA slightly performs better in the external validation. Even though, the results from the internal validation are usually underestimated thus affecting its reliability, I would still recommend using the Multinomial Log-Linear model over the FLDA model since it is not limited by unrealistic assumptions such as normality and equal covariance matrices which are not accurately met in this dataset. Additionally, it is noted that the log-transformation did not produce exactly normal variables which is a requirement for the FLDA, thus impacting the accuracy of the results.

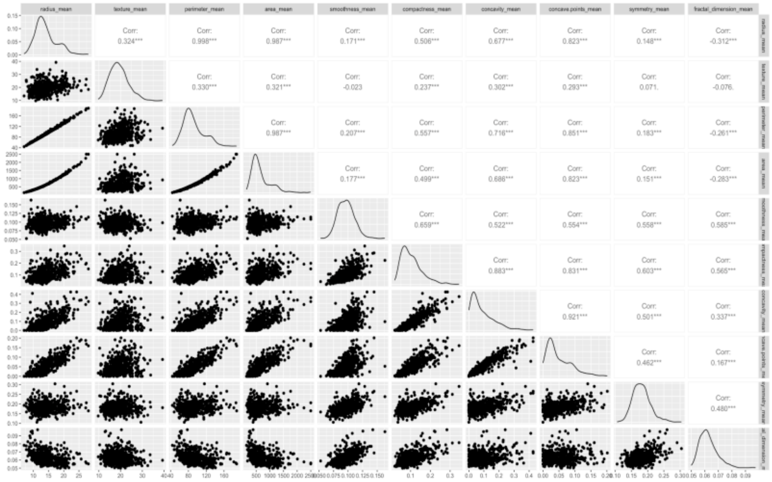


```
External validation:
            rslt
     class    1    2
         B  340   17
         M   21  191
Error Rate = 6.678383 %
```

3.2 Clustering Analysis

3.2.1 Hierarchical Algorithm

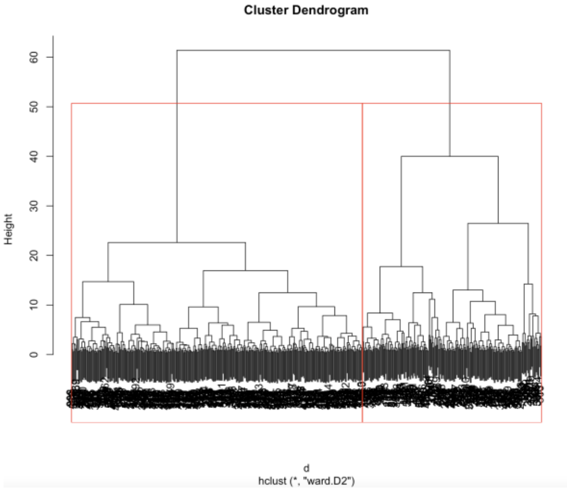
Prior to implementing the clustering algorithm, it was necessary to plot the data in order to determine the shape, structure and direction of the dataset.



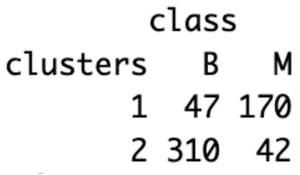
The graph above indicates that the shape varies with variables. To elaborate, some graphs show a spherical shape, and others present an elliptical structure.

3.2.1.1 Euclidean Distance and Ward Method

The first clustering algorithm to be implemented is hierarchical clustering with Euclidean distances as a dissimilarity measure between observations and Ward’s method as a measure of distance between clusters. The reason for this choice is that the ggpair plot of the whole dataset displays both spherical and elliptical shapes. Since the Euclidean distance is not invariant to scaling, the data will be standardized before proceeding with the algorithm. The output is a cluster dendrogram showing how the similar points are put together in the same cluster. According to the output, it was decided that the most suitable number of clusters is k=2.



The confusion matrix was constructed as well to display how many observations were classified in each group. It can be seen that most of the observation were wrongly classified resulting in a huge error rate of 84..35852%. This suggests that this clustering algorithm is not successful.



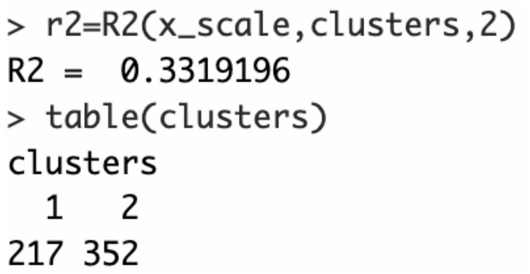
```
        class
clusters   B   M
       1  47 170
       2 310  42
```

Error Rate = 84.35852 %

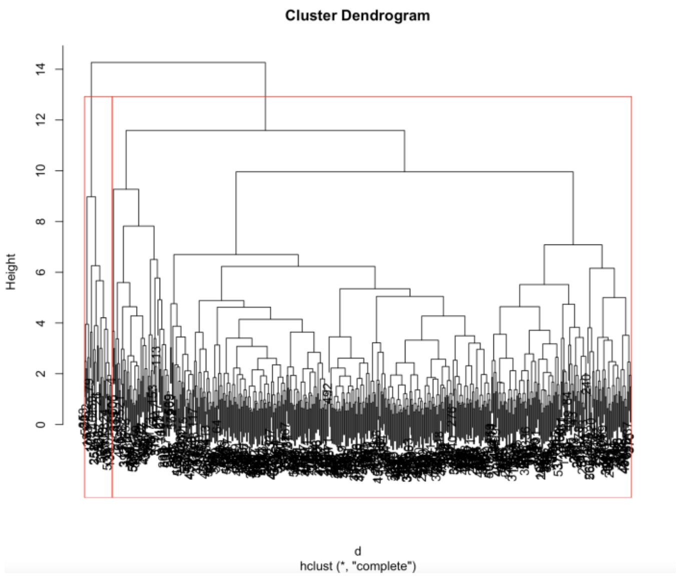
Another way to ensure the goodness of fit of the chosen model for clustering the R2 measure is calculated. It is concluded that the R2 for this model is low at 0.3319 implying that this clustering method was not quite successful because the R2 is far from 1. Both the error rate as well as the R2 value allow us to conclude that this method is not a good fit.

3.2.1.2 Euclidean Distance and Complete Linkage

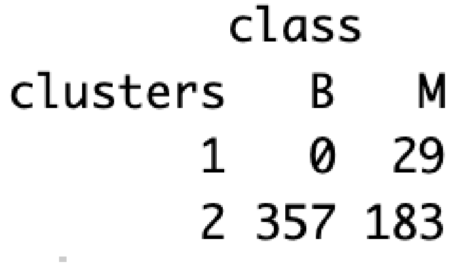
Another combination of distances between observations and clusters were used for the implementation of hierarchical clustering. For the distance between observations the euclidean distance was chosen and for the distances between clusters the complete linkage was chosen. It is obvious from the cluster dendrogram that the most suitable k=2. Additionally, it is observed from the dendrogram that the clusters are more compact with less variance within the clusters relative to the previous dendrogram.



```
> r2=R2(x_scale,clusters,2)
R2 =  0.3319196
> table(clusters)
clusters
  1   2
217 352
```



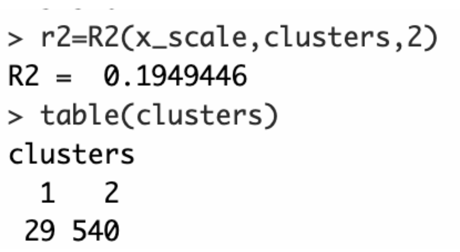
The confusion matrix indicates that 0 obeservations were correctly classified in the benign group and 183 observations only were correctly classified in the malignant group. Consequently, a large error rate of 67.8383% were produced.



```
         class
 clusters   B   M
        1   0  29
        2 357 183
```

Error Rate = 67.83831 %

Furthermore, it is necessary to assess the quality of the results produced by this combination of distances using the R2 measure. Clearly, the R2 measure of goodness of fit is 0.1949 which is a relatively small value given that values for this metric range from 0 to 1. This indicates that this combination of distances for the hierarchical clustering is not a good fit.

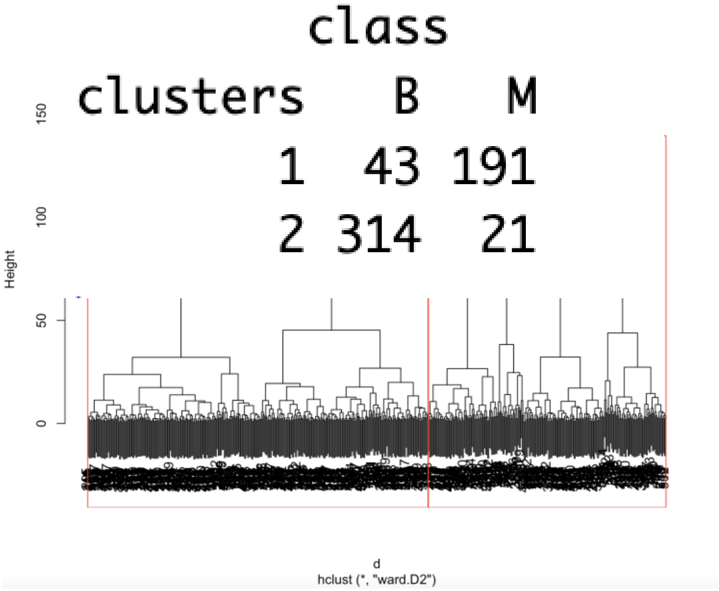


```
> r2=R2(x_scale,clusters,2)
R2 =  0.1949446
> table(clusters)
clusters
  1   2 
 29 540 
```

3.2.1.3 Manhattan Distance and Ward Method

The third clustering algorithm implemented was hierarchical clustering with Manhattan distances as a dissimilarity measure between observations and Ward's method as a measure of distance between clusters. Manhattan distance was chosen as the dissimilarity measure between observations because it is more suitable for non-spherical shapes in the data. The output is a dendrogram that shows how similar points are grouped together in the same cluster. Based on the dendrogram below, it was concluded that the most suitable number of clusters is k=2.

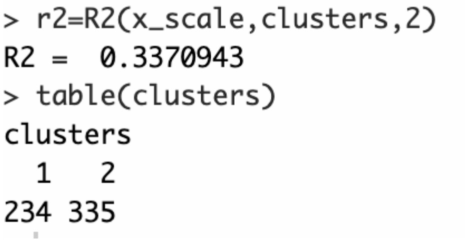
A total of 505 observations were wrongly classified and only 64 observations were correctly classified. The resulting error rate is equal to 88.7522% which is expected since more thna half of the data points are wrongly classified.



```
          class
clusters    B   M
       1   43 191
       2  314  21
```

Error Rate = 88.7522 %

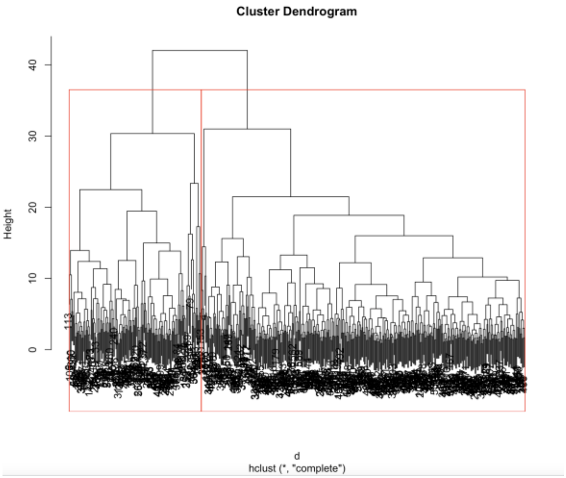
Additionally, the resulting R2 as shown in the output attached, was 0.3370943, backs up the error rate indicating that this clustering method was not entirely successful in explaining the variability in the data.



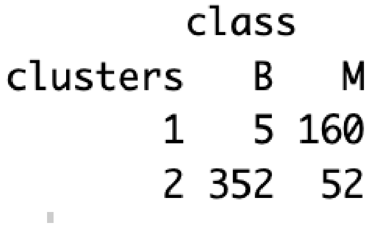
```
> r2=R2(x_scale,clusters,2)
R2 =  0.3370943
> table(clusters)
clusters
  1   2
234 335
```

3.1.2.4 Manhattan Distance and Complete Linkage

Hierarchical clustering with Manhattan distance and Complete Linkage as a measure of distance between clusters was performed on the data as the last hierarchical clustering algorithm. After examining the dendrogram, it was concluded that the most suitable number of clusters is k=2.



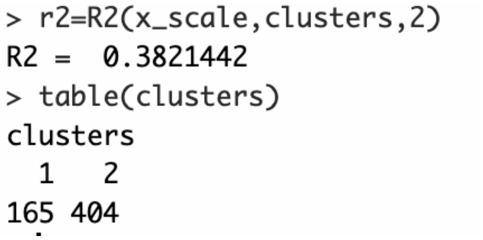
This combination of distances between the observations and clusters used in producing this clustering method resulted in the highest error rate of all the previous methods which is equal to 89.98243%.



```
            class
clusters    B   M
       1    5 160
       2  352  52
```

Error Rate = 89.98243 %

Based on the output attached, hierarchical clustering with Manhattan distance and Complete Linkage did not produce highly successful clustering results for this dataset, as indicated by the relatively low R2 value. This suggests that the data may not be well-suited to clustering using this method and that a different approach may be more appropriate.



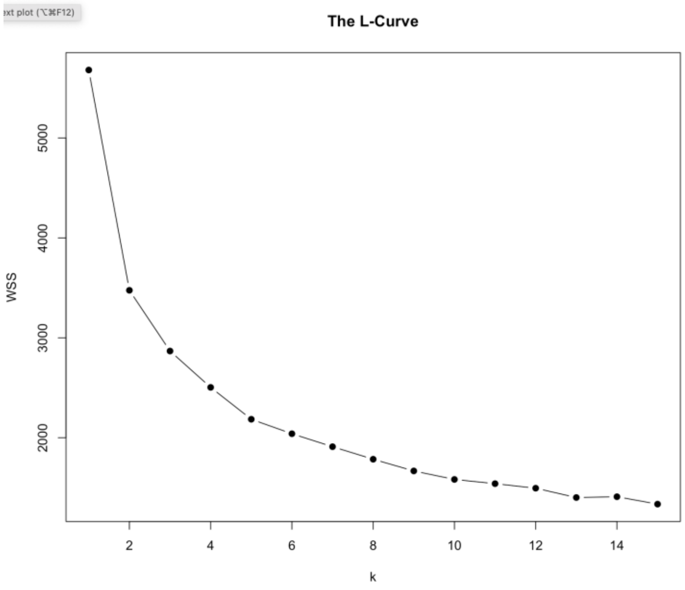
```
> r2=R2(x_scale,clusters,2)
R2 =  0.3821442
> table(clusters)
clusters
  1   2
165 404
```

3.1.2.5 Conclusion on Hierarchical Clustering

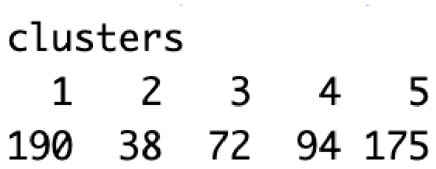
The four hierarchical clustering combinations produced error rates and R2 values that are within the same range. To elaborate, the error rates ranged between 67% to 89%, which is relatively unsatisfactory because it implies that the clustering methods wrongly classify more than half of the data most of the time. As for the R2, the values are relatively low ranging between 0.19 to 0.33 implying that these clustering methods were not successful in grouping the data. Given the high error rates and the low R2 values, it can be concluded that these clustering methods are not appropriate for the dataset. However, out of the four methods, I would recommend the Manhattan and Ward method because it produces the highest R2 of 0.38214.

3.2.2 K-means

This clustering method requires determining the number of clusters in advance. For that reason, the L-curve is plotted. This curve shows the value of the within sum of squares for each value of k. The goal is to choose a k that minimizes the within sum of squares. Therefore, we choose the k at which the within sum of squares stabilizes, in my judgment, it starts to stabilize at k=5 as seen in the graph below.



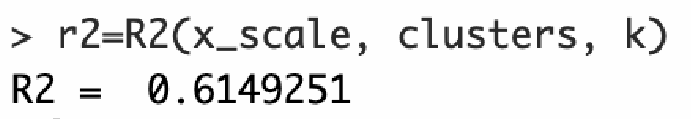
Also, the distribution of the data points in each cluster is seen in the figure below. It is observed that most of the data points are either grouped in cluster 1 or cluster 5.



```
clusters
  1   2   3   4   5 
190  38  72  94 175 
```

The k-means method produced a relatively lower error rate in comparison to the hierarchical clustering method, of 63.09%. However, it is important to note that it is still a high error rate, because it suggests that more than 60% of the data points are classified in the wrong cluster group. As for the R2, it has a value of 0.61492. In my judgment, it is not an extremely high value for R2, meaning that this model is not the best fit, but it is better than the ones produced by the hierarchical clustering.

Error Rate = 63.09315 %



```
> r2=R2(x_scale, clusters, k)
R2 =  0.6149251
```

3.3 Comparison of Clustering Methods

Both the hierarchical cand k-means clustering methods were performed. From the results discussed above, it is clear that the hierarchical clustering method is not suitable for this dataset as it produced very low R2 values and high error rates. On the other hand, the K-means clustering produced much more satisfactory results, to be specific, the R2 value is higher and the error rate is slightly lower. Therefore, I would recommend the use of k-means over the hierarchical. However, given my knowledge on the dataset, I know that the true number of clusters in the dataset is 2 and the k-means suggests the data to have 5 classes. This implies that the k-means clustering as well is not suitable for this dataset.

4 Conclusion

After performing discriminant analysis and clustering analysis on The Breast Cancer Wisconsin (Diagnostic) Dataset, the following conclusions were reached. To elaborate, the Fisher Linear Discriminant Analysis and the multinomial log-linear model were used to implement the discriminant analysis. The former method requires certain assumptions, such as the normality of variables of each group and that the groups would have equal variance covariance matrices, in order for it to be performed successfully, unlike the latter method that does not require making any assumptions. However, these assumptions were not quite met with our dataset which impacted the quality of the results. For that reason, the multinomial log-linear model is preferred over the Fisher Linear Discriminant Analysis even though the error rate produced from the external validation of the multinomial log-linear model was slightly higher than that produced by the FLDA method. In addition, the First Linear Discriminant Projection was also performed. The result showed the orthogonal vector that maximizes the separation of the two groups. As for the clustering analysis, two methods were performed: the hierarchical and k-means clustering methods. For the hierarchical method, a combination of different distances between the observations and distances between clusters were used. All the four combinations produced unsatisfactory results of R2 and error rates. In other words, the outputs resulted in low R2 values and high error rates, indicating that the methods were neither a good fit of successful in classifying the groups into their correct clusters. For that reason, it was concluded that the hierarchical methods of clustering were not appropriate for the dataset used. As for the k-means clustering, it produced a much better R2 and error rate in comparison to the hierarchical clustering method. Therefore, it is recommended to perform the k-means clustering over the hierarchical clustering method. However, given that the dataset used already provides the class variable, it is known that there are only two clusters in this dataset which is contradictory to the output produced by the k-means clustering where it was decided to divide the data into 5 clusters. Consequently, this suggests that the k-means as well is not appropriate for this dataset.

In conclusion, there are numerous other discriminant and clustering analysis methods that can be performed and would be more suitable for this dataset.

References

Azevedo, B. F., Alves, F., Rocha, A. M. A., & Pereira, A. I. (2022, January). Cluster analysis for breast cancer patterns identification. In Optimization, Learning Algorithms and Applications: First International Conference, OL2A 2021, Bragança, Portugal, July 19–21, 2021, Revised Selected Papers (pp. 507-514). Cham: Springer International Publishing.

Documet P, Bear TM, Flatt JD, et al. The association of social support and education with breast and cervical cancer screening. Health Edu Behav. 2015;42:55–64.

El-Habil, A., & El-Jazzar, M. (2013). A comparative study between linear discriminant analysis and multinomial logistic regression. An-Najah University Journal for Research-B (Humanities), 28(6), 1525-1548.

Falk D, Cubbin C, Jones B, et al. Increasing breast and cervical cancer screening in rural and border Texas with friend to friend plus patient navigation. J Cancer Edu. 2018;33:798–05.

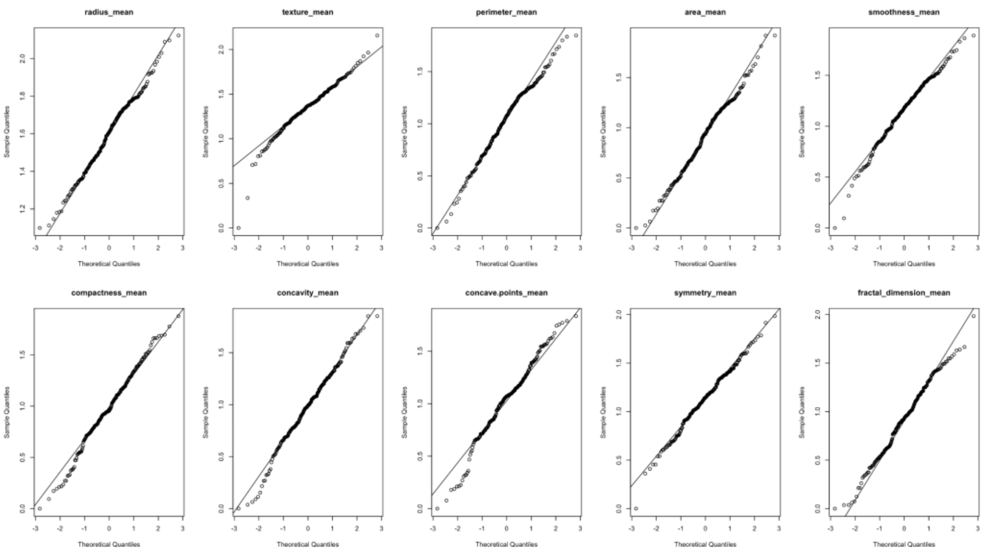
Jessica, E. O., Hamada, M., Yusuf, S. I., & Hassan, M. (2021, December). The Role of Linear Discriminant Analysis for Accurate Prediction of Breast Cancer. In 2021 IEEE 14th International Symposium on Embedded Multicore/Many-core Systems-on-Chip (MCSoC) (pp. 340-344). IEEE.

SR, S. C., & Rajaguru, H. (2019). Comparison analysis of linear discriminant analysis and cuckoo-search algorithm in the classification of breast cancer from digital mammograms. Asian Pacific journal of cancer prevention: APJCP, 20(8), 2333.

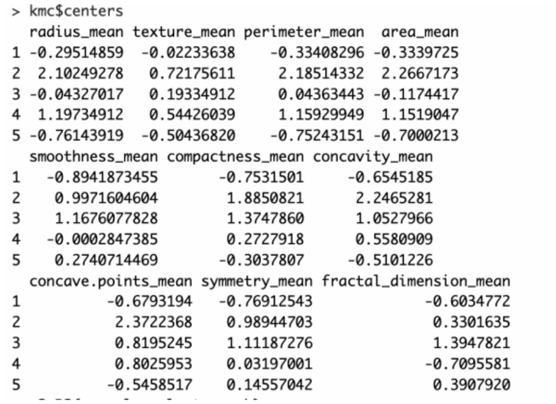
Toğaçar, M., Ergen, B., & Cömert, Z. (2020). Application of breast cancer diagnosis based on a combination of convolutional neural networks, ridge regression and linear discriminant analysis using invasive breast cancer images processed with autoencoders. Medical hypotheses, 135, 109503.

Appendix

The QQ-plots of the whole data set after perfoming the log-transformtion



The output of the k-means clustering method



```
> kmc$centers
  radius_mean texture_mean perimeter_mean  area_mean
1 -0.29514859  -0.02233638    -0.33408296 -0.3339725
2  2.10249278   0.72175611     2.18514332  2.2667173
3 -0.04327017   0.19334912     0.04363443 -0.1174417
4  1.19734912   0.54426039     1.15929949  1.1519047
5 -0.76143919  -0.50436820    -0.75243151 -0.7000213
  smoothness_mean compactness_mean concavity_mean
1   -0.8941873455       -0.7531501     -0.6545185
2    0.9971604604        1.8850821      2.2465281
3    1.1676077828        1.3747860      1.0527966
4   -0.0002847385        0.2727918      0.5580909
5    0.2740714469       -0.3037807     -0.5101226
  concave.points_mean symmetry_mean fractal_dimension_mean
1          -0.6793194   -0.76912543             -0.6034772
2           2.3722368    0.98944703              0.3301635
3           0.8195245    1.11187276              1.3947821
4           0.8025953    0.03197001             -0.7095581
5          -0.5458517    0.14557042              0.3907920
```