Comprehensive Statistical Analysis of Wisconsin Breast Cancer Data: Insights from Regression Analysis and Clustering Techniques

Pedro Ferreira
IPCA, Barcelos, Braga, Portugal
a17029@alunos.ipca.pt

Enmanuel Abilheira
IPCA, Barcelos, Braga, Portugal
a16430@alunos.ipca.pt

Daniel Fernandes
IPCA, Barcelos, Braga, Portugal
a17014@alunos.ipca.pt

***Abstract*—Breast cancer is a significant public health concern, and understanding its predictors and subtypes is crucial for effective diagnosis and treatment. Breast cancer remains one of the most prevalent and formidable challenges in global public health. It is the most commonly diagnosed cancer among women worldwide, with millions of new cases reported annually. Despite advancements in early detection and treatment, breast cancer continues to exact a significant toll on individuals, families, and healthcare systems. Understanding the complex factors contributing to breast cancer development, progression, and response to treatment is essential for improving patient outcomes and reducing mortality rates.**

In this study, we conducted a comprehensive statistical analysis of breast cancer data from the Wisconsin Breast Cancer dataset. Focusing on regression analysis and clustering techniques, we sought to uncover key predictors of breast cancer malignancy and identify distinct subgroups within the dataset. By leveraging clinical and histological features of breast tumors, we aimed to elucidate the multifaceted nature of breast cancer and provide insights that could inform personalized treatment strategies and improve patient care.

***Keywords*—Breast Cancer, Wisconsin Diagnostic Dataset, Regression Analysis, Clustering Techniques**

I. Introduction

Breast cancer is the most commonly diagnosed cancer among women worldwide and is a significant cause of morbidity and mortality. According to the World Health Organization (WHO), breast cancer accounted for an estimated 2.3 million new cases and 685,000 deaths globally in 2020 alone [1].

Having replaced lung cancer as the most commonly diagnosed cancer globally, breast cancer today accounts for 1 in 8 cancer diagnoses and a total of 2.3 million new cases in both sexes combined [2].

Representing a quarter of all cancer cases in females, it was by far the most commonly diagnosed cancer in women in 2020, and its burden has been growing in many parts of the world, particularly in transitioning countries [1], [3].

Statistical analysis plays a crucial role in advancing our understanding of breast cancer and improving patient outcomes. By applying statistical methods to breast cancer research, clinicians and researchers can identify important risk factors associated with disease development and progression. Statistical methods enable researchers to uncover hidden patterns and associations within large-scale genomic and clinical datasets. By analyzing genomic profiles and tumor characteristics, statistical techniques such as clustering and classification algorithms can identify distinct molecular subtypes of breast cancer with unique prognostic and therapeutic implications [4].

A. About the Dataset

The Breast Cancer Wisconsin Diagnostic dataset is widely used in machine learning research and clinical studies. It consists of features computed from digitized images of fine needle aspirates (FNAs) of breast masses, providing valuable information for the analysis and diagnosis of breast tumors.

The dataset contains a total of 569 instances, each with 33 variables. Among these variables, the most crucial one is the diagnosis, which indicates whether the tumor is malignant (M) or benign (B). The remaining variables represent various characteristics of cell nuclei, such as radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry, and fractal dimension.

The features provided in the dataset were computed based on the work described in the paper [5]. This dataset has undergone extensive study and utilization for cancer diagnosis and prognosis prediction.

B. Data Source

The dataset can be obtained from multiple sources, including the UCI Machine Learning Repository [6] and Kaggle [7]. It is also available through the University of Wisconsin's FTP server. The dataset is commonly referred to as the "Wisconsin Breast Cancer dataset" or "WDBC" and is publicly accessible for research purposes.

C. Attributes Description

The dataset contains several attributes, including ID numbers, diagnosis labels, and ten real-valued features computed for each cell nucleus. These features provide valuable insights into the physical dimensions, texture, and composition of the tumor cells. The mean, standard error, and "worst" (or largest) values of these features were computed for each image, resulting in a total of 30 features.

The class distribution within the dataset is imbalanced, with 357 instances labeled as benign and 212 instances labeled as malignant.

Description:

1) ID number
2) Diagnosis (M = malignant, B = benign)
3) Ten real-valued features are computed for each cell nucleus:
   a) Radius
   b) Texture (standard deviation of gray-scale values)
   c) Perimeter
   d) Area
   e) Smoothness (local variation in radius lengths)
   f) Compactness
   g) Concavity (severity of concave portions of the contour)
   h) Concave points (number of concave portions of the contour)
   i) Symmetry
   j) Fractal dimension ("coastline approximation")

This dataset serves as a valuable resource for researchers and practitioners in the field of oncology, facilitating the development of predictive models and diagnostic tools for breast cancer detection and treatment.

D. Goals

The primary aim of this comprehensive analysis is to rigorously examine and identify potential predictors for breast cancer through the application of various statistical methodologies. In this particular study, three distinct statistical techniques were utilized: linear regression, logistic regression and clustering methodologies.

1) Linear Regression: The primary objective of applying linear regression is to model the relationship between one or more independent variables (predictors) and a dependent quantitative variable (response). Linear regression aims to find the best-fitting linear equation that describes the relationship between these variables.

Linear regression serves multiple purposes in statistical analysis. Primarily, it facilitates the comprehension of the relationship existing between independent variables and the dependent variable by quantifying this association through the estimation of slope and intercept parameters inherent in the linear equation.

Moreover, it enables the execution of statistical inference, permitting hypothesis testing on model coefficients and the evaluation of overall fit to the data. This functionality serves to ascertain the significance of relationships and the reliability of the model under examination.

2) Logistic Regression: The primary objective of applying logistic regression to a dataset in statistical analysis is to model the relationship between one or more independent variables (predictors) and a binary dependent variable (response). Unlike linear regression, which is suitable for continuous outcomes, logistic regression is used when the dependent variable is categorical with two levels.

Logistic regression serves multiple key purposes in statistical analysis. It models the probability that an observation belongs to a specific category or class, estimating this probability based on predictor variables. Furthermore, it helps understanding how changes in independent variables relate to changes in the probability of the outcome, achieved through estimation of model coefficients.

3) Clustering: Clustering techniques aim to partition the dataset into groups or clusters, where observations within the same cluster are more similar to each other than to those in other clusters.

Clustering techniques play a vital role in statistical analysis by offering a multifaceted approach to understanding and analyzing datasets. Moreover, clustering aids in pattern recognition by identifying inherent structures or clusters within the data. These patterns offer valuable insights that can inform subsequent analysis or decision-making processes. Additionally, clustering serves as a powerful tool for anomaly detection, enabling the identification of observations that deviate significantly from established patterns.

II. Methodology

Before delineating the methodology for each of the techniques, it is imperative to note that the feature "Id" was not considered in the analysis, as it solely serves as an identifier for each record.

A. Linear Regression

Process:

1) For the multiple linear regression (MLR) analysis, emphasis was placed on the mean attributes of the dataset. The selection of mean features was based on their capacity to represent the average value across multiple measurements, thereby offering a robust characterization of the central tendency of each attribute.
2) Subsequently, the correlation between these features was examined, and the feature demonstrating the highest absolute correlation with the others was designated as the response variable.
3) With the selected features, we proceeded to fit the linear model in order to assess the statistical significance of the coefficients/predictors.
4) Following the linear model fitting, we executed a sequential deletion algorithm to ascertain whether a submodel exhibiting comparable or superior explanatory power with fewer predictors, in accordance with the principle of parsimony, could be attained.
5) Furthermore, we conducted an F-test using ANOVA to compare the full model with the submodel.

6) In the final submodel, the summary of the fit revealed that the coefficients of all utilized predictors possessed associated p-values below the conventional significance level of $\alpha = 0.05$. Thus, we infer that all these predictors have statistically significant coefficients.

B. Logistic Regression

Process:

1) For the logistic regression analysis, we initiated by designating "diagnosis" as the response variable, with all remaining features serving as predictors.
2) When attempting to fit the full model (encompassing all features), it was observed that the model failed to converge. This occurrence could be attributed to several factors, including elevated multicollinearity among predictors, data separation (wherein certain predictor values solely correspond to one of the possible outcome variable values, thus leading to perfect prediction), or numerical instability.
3) Following this, an analysis of the correlation between these features was conducted, revealing the presence of numerous highly correlated predictor features.
4) Moreover, a Variance Inflation Factor (VIF) analysis was performed, which quantifies the degree to which the variance of an estimated regression coefficient is inflated due to collinearity. Subsequently, it was noted that the predictors exhibited notably high VIF values.
5) To mitigate this issue, feature selection was conducted utilizing Lasso regression. Cross-validation techniques were employed to determine the optimal lambda (regularization parameter) that minimizes cross-validation error.
6) Subsequently, the significant features identified through Lasso regression were utilized to construct a logistic regression model.
7) Following the logistic model fitting, we executed a sequential deletion algorithm to ascertain whether a submodel exhibiting comparable or superior explanatory power with fewer predictors, in accordance with the principle of parsimony, could be attained.
8) Furthermore, we conducted an F-test using ANOVA to compare the full model with the submodel.
9) In the final submodel, the summary of the fit revealed that the coefficients of all utilized predictors possessed associated p-values below the conventional significance level of $\alpha = 0.05$. Thus, we infer that all these predictors have statistically significant coefficients.
10) Finally, a likelihood ratio test statistic was conducted to ascertain whether removing any of these predictors would lead to a significant decrease in the model fit. It was concluded that all the predictors employed remained statistically significant in the final model.

C. Clustering

Process:

1) In the clustering analysis, all features were utilized except for the qualitative feature 'diagnosis'.
2) Prior to implementing the clustering techniques, a check was conducted to determine if data scaling was necessary. It was observed that the input variables exhibited varying units of measurement and significantly different variances. Consequently, it was concluded that scaling was required.
3) Two clustering techniques were employed: hierarchical clustering and k-means.
4) Regarding hierarchical clustering, the Euclidean distance metric was utilized. An analysis was conducted using clValid to determine the most appropriate linkage method for our data. Following testing and analysis, it was determined that Ward's method was the most suitable linkage method for our purposes. Additionally, through examination of the dendrogram and optimal scores obtained from clValid, it was concluded that the appropriate number of clusters for our data was 2.
5) Regarding k-means clustering, we assessed the optimal starting number of clusters (k) using the elbow method, silhouette method, and clValid. Following this analysis, it was determined that the most appropriate number of initial clusters (k) to yield meaningful results was 2.
6) Furthermore, principal component analysis (PCA) was conducted to identify the minimum number of principal components necessary to explain more than 80% of the variance in the data. It was determined that this minimum number of components was 5. Subsequently, hierarchical clustering and k-means were executed using this number of components.

III. Results

A. Linear Regression

In this subsection, we will present and discuss the results obtained for the final linear regression model.

The model summary output is as follows:

```
Call:
glm(formula = radius_mean ~ perimeter_mean + area_mean + smoothness_mean +
    compactness_mean + concavity_mean + fractal_dimension_mean,
    family = gaussian(), data = mean_features)

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)           1.669e-01  1.225e-01   1.363 0.173547
perimeter_mean        1.566e-01  1.308e-03 119.715  < 2e-16 ***
area_mean            -2.844e-04  7.805e-05  -3.643 0.000294 ***
smoothness_mean       1.214e+00  3.756e-01   3.232 0.001302 **
compactness_mean     -4.780e+00  2.569e-01 -18.602  < 2e-16 ***
concavity_mean       -8.079e-01  1.263e-01  -6.394  3.4e-10 ***
fractal_dimension_mean  3.204e+00  1.323e+00   2.421 0.015789 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for gaussian family taken to be 0.007840178)

    Null deviance: 7053.9466  on 568  degrees of freedom
Residual deviance:    4.4062  on 562  degrees of freedom
AIC: -1135.1

Number of Fisher Scoring iterations: 2
```

Fig. 1. Summary of the linear regression model

The overall model suggests that the coefficients for all predictors have associated p-values that are less than the

conventional significance level of 0.05. This indicates that all predictor variables are statistically significant in explaining the variability in the response variable (radius_mean). Therefore, we can infer that all predictors have statistically significant coefficients.

The model's goodness of fit can be evaluated by comparing the null deviance (before adding predictors) and the residual deviance (after adding predictors). A significant reduction in deviance suggests that the predictors explain a significant amount of variability in the response variable.

Coefficients Interpretation:

1) **Perimeter Mean:** For every one-unit increase in perimeter_mean, the radius_mean is expected to increase by 0.1566 units, on average, holding other predictors constant.
2) **Area Mean:** For every one-unit increase in area_mean, the radius_mean is expected to decrease by 0.0002844 units, on average, holding other predictors constant.
3) **Smoothness Mean:** For every one-unit increase in smoothness_mean, the radius_mean is expected to increase by 1.214 units, on average, holding other predictors constant.
4) **Compactness Mean:** For every one-unit increase in compactness_mean, the radius_mean is expected to decrease by 4.780 units, on average, holding other predictors constant.
5) **Concavity Mean:** For every one-unit increase in concavity_mean, the radius_mean is expected to decrease by 0.8079 units, on average, holding other predictors constant.
6) **Fractal Dimension Mean:** For every one-unit increase in fractal_dimension_mean, the radius_mean is expected to increase by 3.204 units, on average, holding other predictors constant.

The equation for the adjusted hyperplane of radius_mean adjusted, based on the model summary is:

$$\begin{aligned}\text{radius\_mean} = {} & 0.1669 + 0.1566 \times \text{perimeter\_mean} \\ & - 0.0002844 \times \text{area\_mean} \\ & + 1.214 \times \text{smoothness\_mean} \\ & - 4.780 \times \text{compactness\_mean} \\ & - 0.8079 \times \text{concavity\_mean} \\ & + 3.204 \times \text{fractal\_dimension\_mean}\end{aligned} \tag{1}$$

Sum of Squared Residuals:

$$SS_{\text{Residuals}} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{2}$$

For the sum of squared residuals, the calculated value was 4.40618.

Coefficient of Determination (R-squared):

$$R^2 = 1 - \frac{\text{Deviance of MLR model}}{\text{Deviance of null model}} \tag{3}$$

For the coefficient of determination (R-squared), the calculated value was 0.9993754.

Residuals Analysis:

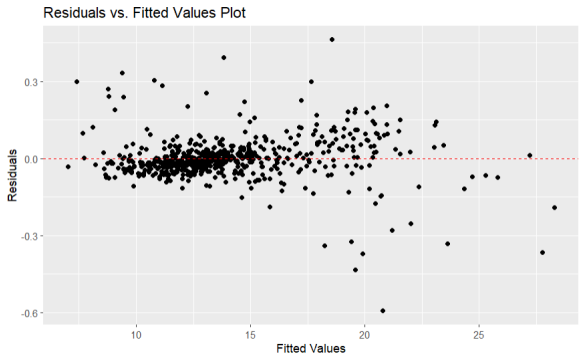

Fig. 2. Residuals vs. Fitted Values Plot

The points depicted in the plot correspond to the residuals associated with each observation within the dataset. Ideally, these residuals should exhibit a random dispersion around the horizontal line defined by y = 0.

From the plot, it is discernible that the residuals exhibit a predominantly random dispersion around the horizontal line at y = 0, exhibiting no discernible pattern as the fitted values vary, save for a number of outliers. Notably, the spread of residuals remains consistently distributed across the spectrum of fitted values, indicating that the linear regression model adequately captures the underlying relationship between the predictors and the response variable.

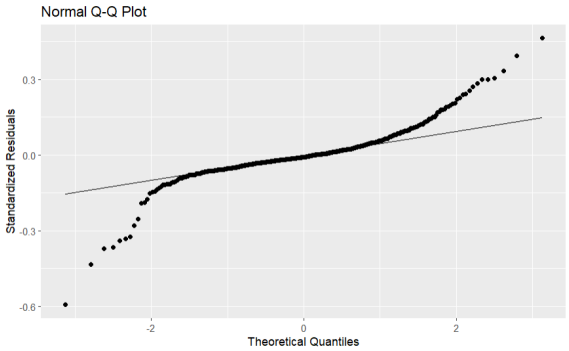

Fig. 3. Normal Q-Q Plot

The normal Q-Q (quantile-quantile) plot is a diagnostic tool used to assess whether the distribution of the residuals from a

regression model follows a normal (Gaussian) distribution.

The x-axis of the plot represents the theoretical quantiles of a standard normal distribution, while the y-axis represents the standardized residuals (residuals divided by their standard deviation), which should ideally follow a standard normal distribution if the regression model assumptions are met.

From figure 3, the observed residuals adhere to the theoretical quantiles for the most part, indicating substantial conformity to the normal distribution assumption. However, some deviations from the expected pattern are discernible at the extremities of the plot.

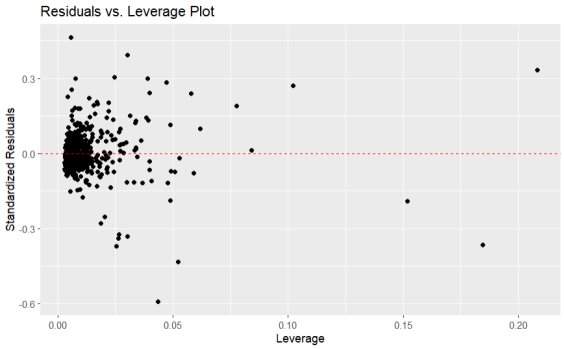

Fig. 4. Standardized Residuals vs. Leverage Plot

The plot displays the standardized residuals against the leverage values. It helps to identify influential observations, particularly those with high leverage and large residuals. In this plot 4, the dashed red line indicates the zero-residual line, helping to identify observations with residuals deviating significantly from zero. High leverage points, combined with large residuals, may have a considerable impact on the model's fit and should be investigated further for potential outliers or influential data points.

We can observe that there is some outliers in terms of residuals or leverage. Outliers in the plot may indicate problematic observations that are exerting undue influence on the regression model.

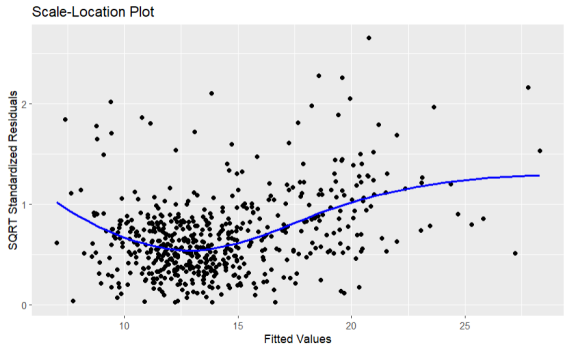

Fig. 5. Scale-Location Plot

The scale-location plot is a diagnostic plot used to assess that the variance of the errors (or residuals) is constant across the range of fitted values.

We can observe from the plot 5 that the smooth trend line, is not approximately horizontal. The slight slope in the line may indicate that the variability of the residuals is not entirely constant but rather increases or decreases as the fitted values change. This could suggest a mild violation of the assumption that the variance of the residuals is constant across the range of fitted values.

B. Logistic Regression

In this subsection, we will present and discuss the results obtained for the final linear regression model.

The model summary output is as follows:

```
Call:
glm(formula = diagnosis_ohe ~ radius_mean + symmetry_mean + texture_worst +
    perimeter_worst, family = binomial(link = "logit"), data = data_logreg)

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)     -35.36245    4.80487  -7.360 1.84e-13 ***
radius_mean      -0.86430    0.35087  -2.463 0.013766 *
symmetry_mean    40.67821   11.05987   3.678 0.000235 ***
texture_worst     0.26246    0.04950   5.302 1.15e-07 ***
perimeter_worst   0.30773    0.05044   6.101 1.05e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 751.44  on 568  degrees of freedom
Residual deviance: 119.71  on 564  degrees of freedom
AIC: 129.71

Number of Fisher Scoring iterations: 8
```

Fig. 6. Summary of the logistic regression model

The overall model suggests that the coefficients for all predictors have associated p-values that are less than the conventional significance level of 0.05. This indicates that all predictor variables are statistically significant in explaining the variability in the response variable (diagnosis). Therefore, we can infer that all predictors have statistically significant coefficients.

Coefficients Interpretation:

1) **Radius Mean:** For every one-unit increase in radius_mean, the log odds of the diagnosis being malignant decrease by 0.86430 units, on average, holding other predictors constant.
2) **Symmetry Mean:** For every one-unit increase in symmetry_mean, the log odds of the diagnosis being malignant increase by 40.67821 units, on average, holding other predictors constant.
3) **Texture Worst:** For every one-unit increase in texture_worst, the log odds of the diagnosis being malignant increase by 0.26246 units, on average, holding other predictors constant.
4) **Perimeter Worst:** For every one-unit increase in perimeter_worst, the log odds of the diagnosis being malignant

increase by 0.30773 units, on average, holding other predictors constant.

The logistic regression model's adjusted hyperplane equation can be represented as follows:

$$\begin{aligned}\text{logit}(\hat{p}) = \quad & -35.36245 \\ & -0.86430 \times \text{radius\_mean} \\ & +40.67821 \times \text{symmetry\_mean} \\ & +0.26246 \times \text{texture\_worst} \\ & +0.30773 \times \text{perimeter\_worst}\end{aligned} \tag{4}$$

The expression for p(x), which represents the probability of the outcome given the predictor variables x, can be derived from the logistic regression model using the sigmoid function:

$$p(x)_{\text{logistic}} = \frac{1}{1 + e^{-(-35.36245 - 0.86430 \times \text{radius\_mean} + \cdots)}} \tag{5}$$

Odds ratios

$$OddsRatio_{x_i} = e^{\beta_i} \tag{6}$$

	OddsRatio	Std. Err.	z	P>|z|
radius_mean	4.213484e-01	1.478377e-01	-2.463306	1.376624e-02
symmetry_mean	4.637909e+17	5.129465e+18	3.678002	2.350681e-04
texture_worst	1.300125e+00	6.435971e-02	5.301933	1.145827e-07
perimeter_worst	1.360333e+00	6.861536e-02	6.100882	1.054846e-09

Fig. 7. Odds ratios

The odds ratio indicates how the odds of the outcome change with a one-unit increase in the predictor variable, holding all other variables constant.

Residuals Analysis:

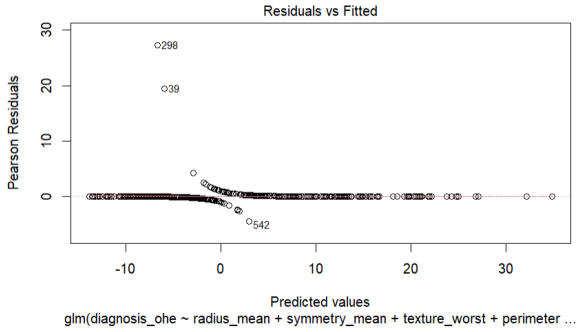

Fig. 8. Residuals vs. Fitted Values Plot

Upon examination of the residuals vs. fitted values plot, it is observed that the residuals exhibit a random dispersion around the horizontal line at y = 0, with some isolated exceptions. Furthermore, there is no discernible pattern evident in the residuals as the fitted values vary. Additionally, the spread of residuals remains relatively consistent across the spectrum of fitted values. These observations collectively suggest that the logistic regression model adequately captures the underlying relationship between the predictors and the response variable.

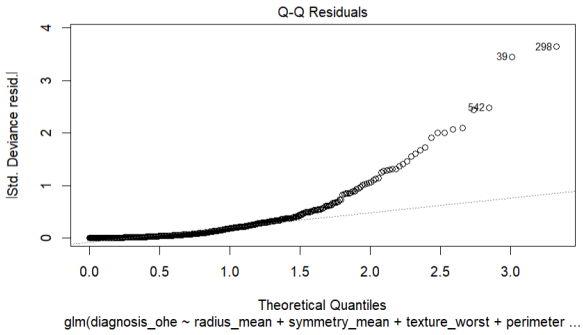

Fig. 9. Normal Q-Q Plot

In the normal quantile-quantile (Q-Q) plot of the residuals, the observed residuals predominantly adhere to the theoretical quantiles, suggesting substantial conformity to the normal distribution assumption. However, a deviation from the expected pattern is noticeable at the right side of the quantiles axis. Despite this deviation, the overall assessment from the normal Q-Q plot suggests a reasonable level of adherence to normality for the residuals.

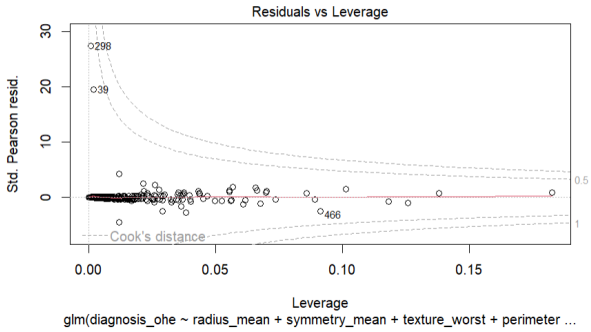

Fig. 10. Standardized Residuals vs. Leverage Plot

In a residuals vs. leverage plot, the leverage values represent the influence that each data point has on the estimation of the regression coefficients.

Overall, the plot exhibits a predominant clustering of points around the y = 0 line, with only minor instances of departure, indicative of a relatively well-behaved dataset with some slight outliers.

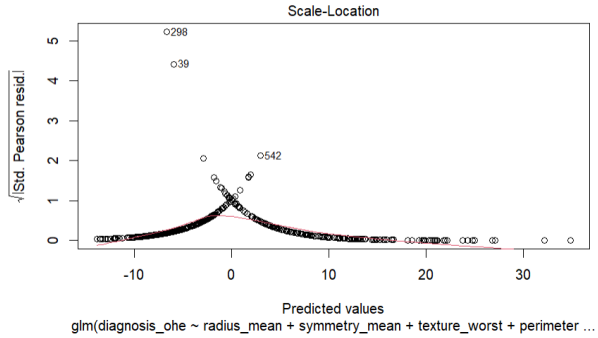

Fig. 11. Scale-Location Plot

We can observe from the plot that the smooth trend line, is not approximately horizontal. The slight slope in the line may indicate that the variability of the residuals is not entirely constant but rather increases or decreases as the fitted values change. This could suggest a mild violation of the assumption that the variance of the residuals is constant across the range of fitted values.

C. Clustering

Prior to presenting the results, it is pertinent to note that the features underwent scaling procedures before performing the clustering analysis.

1) Hierarchical Clustering: To analyze the optimal number of clusters for our data, we conducted a dendrogram analysis.

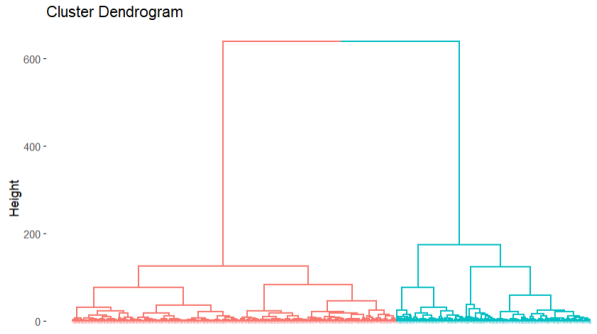

Fig. 12. Dendrogram

Analyzing the dendrogram, we can ascertain that the most suitable number of clusters for our data is 2.

A visual representation of the hierarchical clusters:

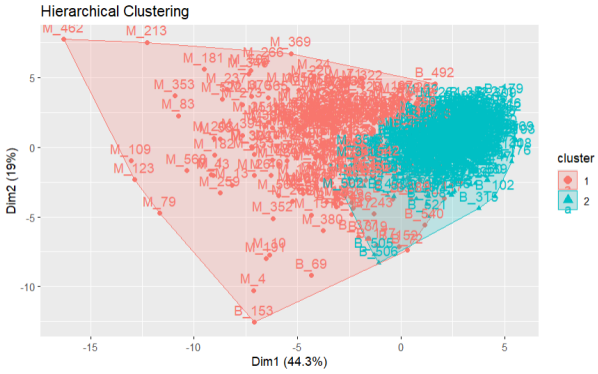

Fig. 13. Hierarchical clusters

Mean and standard deviation for all variables by cluster. The table is truncated, showing only a few features.

hc.clusters	perimeter_mean_m	perimeter_mean_std	perimeter_se_m	perimeter_se_std	perimeter_worst_m
1	0.9049638	0.9959089	0.7293945	1.2489579	0.9508281
2	-0.5414531	0.4715084	-0.4364074	0.3978177	-0.5688944

Fig. 14. Mean and standard deviation by cluster

Comparison of cluster membership with diagnosis:

Comparison of Cluster Membership with Diagnoses

	Actual Benign	Actual Malignant
Cluster 1	29	184
Cluster 2	328	28

Fig. 15. Hierarchical Cluster vs. Actual Diagnosis

2) K-means: Regarding k-means clustering, we assessed the optimal starting number of clusters (k) using the elbow method, silhouette method, and clValid.

For illustrative purposes, we will present the graphical representation of the silhouette method.

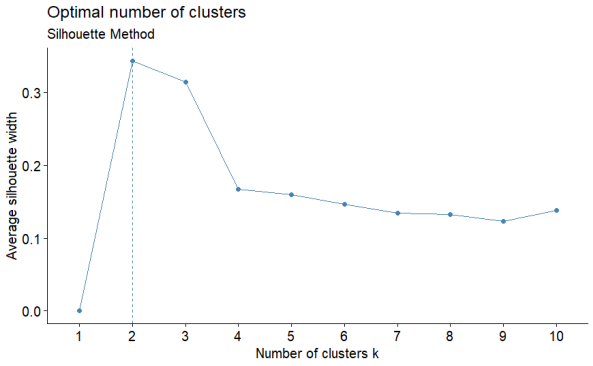

Fig. 16. Silhouette method

A visual representation of the k-means clusters:

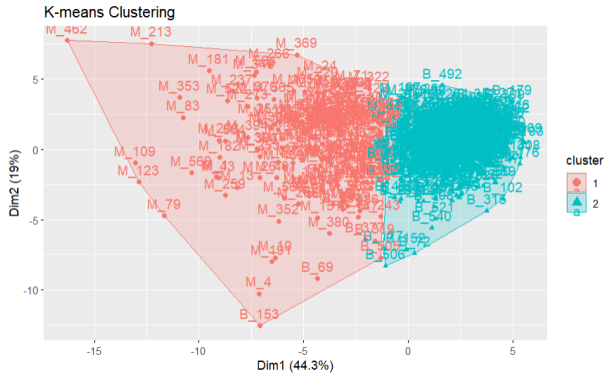

Fig. 17. K-means clusters

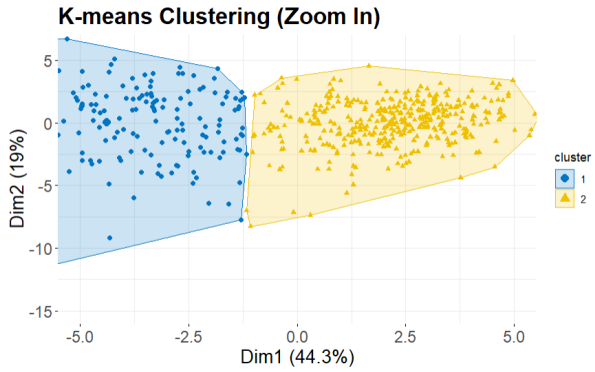

Fig. 18. K-means clusters (zoomed)

Mean and standard deviation for all variables by cluster. The table is truncated, showing only a few features.

km.clusters	perimeter_mean_m	perimeter_mean_std	perimeter_se_m	perimeter_se_std	perimeter_worst_m
<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	1.0057496	0.9636973	0.8595226	1.2673289	1.0650336
2	-0.5002281	0.5316766	-0.4274994	0.3877156	-0.5297141

Fig. 19. Mean and standard deviation by cluster

Comparison of cluster membership with diagnosis:

Comparison of Cluster Membership with Diagnoses

	Actual Benign	Actual Malignant
Cluster 1	14	175
Cluster 2	343	37

Fig. 20. K-means Cluster vs. Actual Diagnosis

IV. Conclusion

In conclusion, the statistical analysis of the Wisconsin Breast Cancer dataset, incorporating linear regression, logistic regression, and clustering techniques, successfully achieved its objectives. Through these methodologies, significant insights were gleaned regarding the predictive factors and underlying structures associated with breast cancer, thereby enhancing our understanding of this complex disease and potentially informing future research and clinical practices.

References

[1] M. Arnold, E. Morgan, H. Rumgay, A. Mafra, D. Singh, M. Laversanne, J. Vignat, J. R. Gralow, F. Cardoso, S. Siesling, *et al.*, "Current and future burden of breast cancer: Global statistics for 2020 and 2040," *The Breast*, vol. 66, pp. 15–23, 2022.

[2] H. Sung, J. Ferlay, R. L. Siegel, M. Laversanne, I. Soerjomataram, A. Jemal, and F. Bray, "Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA: a cancer journal for clinicians*, vol. 71, no. 3, pp. 209–249, 2021.

[3] E. Heer, A. Harper, N. Escandor, H. Sung, V. McCormack, and M. M. Fidler-Benaoudia, "Global burden and trends in premenopausal and postmenopausal breast cancer: a population-based study," *The Lancet Global Health*, vol. 8, no. 8, pp. e1027–e1037, 2020.

[4] C. Curtis, S. P. Shah, S.-F. Chin, G. Turashvili, O. M. Rueda, M. J. Dunning, D. Speed, A. G. Lynch, S. Samarajiwa, Y. Yuan, *et al.*, "The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups," *Nature*, vol. 486, no. 7403, pp. 346–352, 2012.

[5] K. P. Bennett and O. L. Mangasarian, "Robust linear programming discrimination of two linearly inseparable sets," *Optimization methods and software*, vol. 1, no. 1, pp. 23–34, 1992.

[6] "UCI Machine Learning Repository." https://archive.ics.uci.edu/. Accessed: April 8, 2024.

[7] "Kaggle: Your Home for Data Science." https://www.kaggle.com/. Accessed: April 8, 2024.