Logistics Report

Kilian Wan

July 14, 2024

1 Introduction

This report aims to develop a predictive model using logistic regression to classify tumors as benign or malignant based on several cellular attributes extracted from breast mass samples. The data have been collected from the Wisconsin Breast Cancer Dataset (Original). To explain what does these variables do, we will use the article of [5].

1. **Clump Thickness**: This is used to assess if cells are mono-layered or multi-layered. Benign cells tend to be grouped in mono-layers, while cancerous cells are often grouped in multi-layer. We will call it X_1 for the logistic regression.
2. **Uniformity of Cell Size**: It is used to evaluate the consistency in the size of cells in the sample. Cancer cells tend to vary in size. That is why this parameter is very valuable in determining whether the cells are cancerous or not. We will call it X_2.
3. **Uniformity of Cell Shape**: It is used to estimate the equality of cell shapes and identifies marginal variances, because cancer cells tend to vary in shape. We will call it X_3.
4. **Marginal Adhesion**: Normal cells tend to stick together. Cancer cells tend to loose this ability. So loss of adhesion is a sign of malignancy. We will call it X_4.
5. **Single Epithelial Cell Size**: It is related to the uniformity. Epithelial cells that are significantly enlarged may be a malignant cell. We will call it X_5.
6. **Bare Nuclei**: This is a term used for nuclei that is not surrounded by cytoplasm. Those are typically seen in benign tumors. We will call it X_6.
7. **Bland Chromatin**: Describes a uniform "texture" of the nucleus seen in benign cells. In cancer cells, the chromatin tends to be coarser. We will call it X_7.
8. **Normal Nucleoli**: Nucleoli are small structures seen in the nucleus. In normal cells the nucleolus is usually very small if visible at all. In cancer cells the nucleoli become much more prominent, and sometimes there are more of them. We will call it X_8.
9. **Mitoses**: It is an estimate of the number of mitosis that has taken place. Larger the value, greater is the chance of malignancy. We will call it X_9.

And *Class*: Indicates whether the tumor is benign (2) or malignant (4). The primary question addressed in this report is: *Can we accurately predict the malignancy of breast tumors using logistic regression?*

2 EDA

Univariate Numerical Analysis

In this subsection we will explore each variable in the dataset thanks to the univariate numerical analysis. We will use the fact that the difference between the mean and median shows a skewed distribution. So we will not mention it every time for simplicity, that is when the mean and median are considerably different, we will assume directly the skewness. The summary is given on the Table 1 below.

Variable	Min	1st Qu.	Median	Mean	3rd Qu.	Max
`Clump Thickness`	1	2	4	4.42	6	10
`Uniformity of Cell Size`	1	1	1	3.13	5	10
`Uniformity of Cell Shape`	1	1	1	3.20	5	10
`Marginal Adhesion`	1	1	1	2.84	4	10
`Single Epithelial Cell Size`	1	2	2	3.22	4	10
`Bare Nuclei`	1	1	1	3.54	6	10
`Bland Chromatin`	1	2	3	3.44	5	10
`Normal Nucleoli`	1	1	1	2.87	4	10
`Mitoses`	1	1	1	1.60	2	10

TABLE 1: Summary statistics for each variable in the Breast Cancer Wisconsin dataset.

Remarks:

- *Clump Thickness*: The mean and median values indicate that while most cell clumps are thin, there are instances of significantly thicker clumps, as reflected by the higher maximum value. This variation can be indicative of malignancy.
- *Uniformity of Cell Size*: The low median value and right-skewed distribution suggest that most cells are uniform in size, with some showing significant variation. Higher uniformity values can indicate malignancy.
- *Marginal Adhesion*: The majority of samples exhibit low adhesion values, but a few samples have much higher values, and it can be again an indicative of malignancy.
- *Bare Nuclei*: The presence of bare nuclei is a strong indicator of malignancy. The distribution shows that while most samples have few bare nuclei, some have significantly more, indicating potential malignancy.
- *Mitoses*: Most samples have low mitotic activity, but some exhibit higher rates of cell division, which can be associated with malignancy.

Univariate Graphical

For the univariate graphical analysis, we examined the distribution of each predictor variable using histograms. The x-axis of each histogram represents the values of the predictor variable, while the y-axis represents the frequency of those values in the dataset. Due to similarity in the distributions, we provide a generalized explanation for all histograms and apply this reasoning to each of the nine histograms.

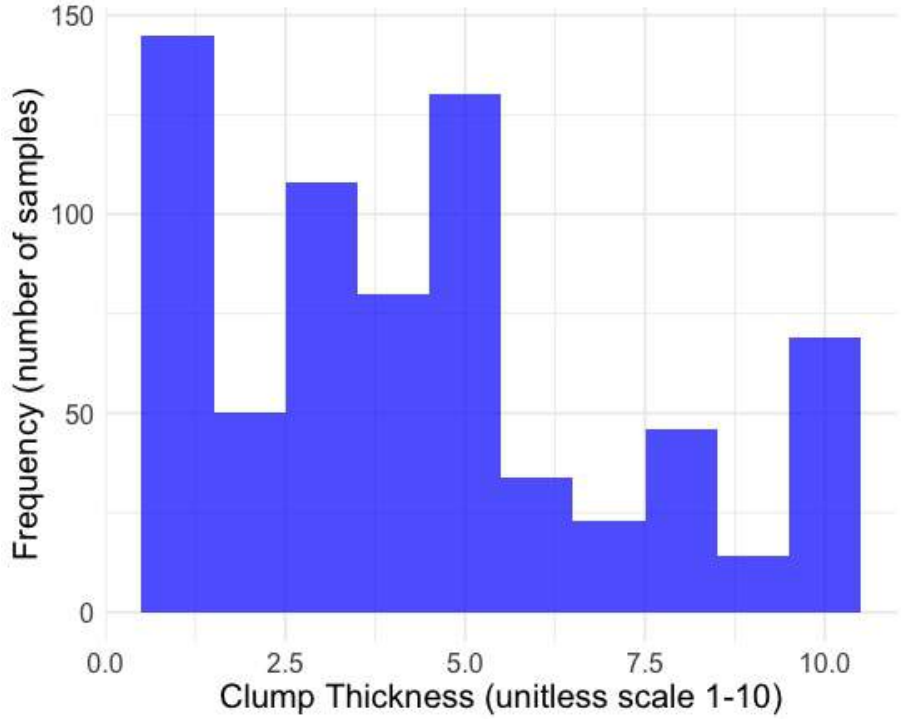

Fig. 1: Histogram of Clump Thickness

Most histograms show a right-skewed distribution. This indicates that while the majority of the samples have low values for these predictors, there are some samples with much higher values. This suggests that the characteristics measured by these variables are, in general low, but when there are higher values, it can be indicative of malignancy. For simplicity and to avoid redundancy, we describe this pattern once, and the same interpretation applies to all histograms of the variables (see Figures in Appendix A): Clump Thickness, Uniformity of Cell Size, Uniformity of Cell Shape, Marginal Adhesion, Single Epithelial Cell Size, Bare Nuclei, Bland Chromatin, Normal Nucleoli, and Mitoses. As an example, consider the histogram of `Clump-Thickness` (see Figure 1). As before, the x-axis represents clump thickness values on a scale from 1 to 10, and the y-axis represents the number of samples with each clump thickness value. The distribution is indeed right-skewed, and we observe some outliers with high values. This could suggest that while most clumps are thin, some are thicker, and it could indicate malignancy.

Bivariate Numerical Analysis

To understand the relationships between the different features of the dataset, we will compute and visualize the correlation matrix. To understand the relationships between the different features of the dataset, we visualize the correlation matrix using a heat-map shown in Figure 2. The heat-map displays the *Pearson correlation coefficients* between each pair of variables. These coefficients range from -1 to 1, where values close to 1 indicate a strong positive correlation, values close to -1 indicate a strong negative correlation, and values around 0 indicate no correlation. We then obtain the following plot on the next page.

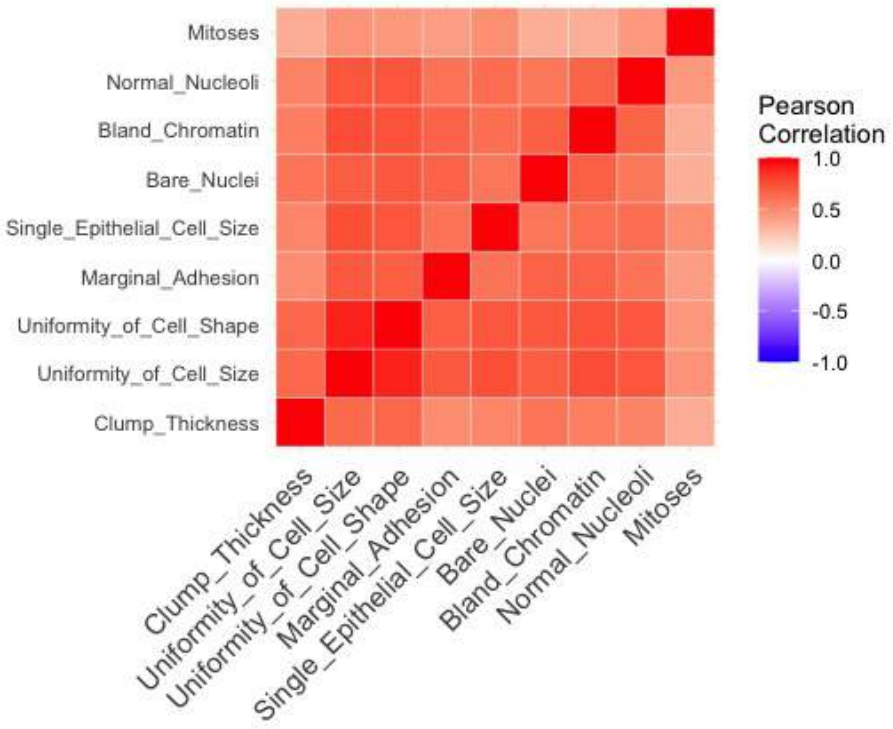

FIG. 2: Heat-map of the correlation matrix for the Breast Cancer Wisconsin dataset.

Remarks: Upon examining the heat-map, we observe the following relationships:

- *Clump Thickness and Uniformity of Cell Size*: These two variables show a strong positive correlation (coeff. of `0.82`). This suggests that as the thickness of the cell clumps increases, the uniformity of cell size also tends to increase.
- *Uniformity of Cell Shape and Uniformity of Cell Size*: There is a strong positive correlation (coeff. of `0.71`) between these two variables,

Bivariate Graphical

For the bivariate graphical analysis, we will create scatter plots for pairs of variables that show significant correlations. To choose the pairs of variables, we use the heat-map or the correlation matrix (it's equivalent). As in the univariate graphical part, due to the similarity in the conclusions drawn from these scatter plots, we provide again a generalized explanation and apply this to the scatter plots (see Figures in Appendix B). To do so, we detail one, and mention that similar patterns are observed in the others.

The scatter plots reveal positive correlations between some pairs of variables. This confirms the relationships seen in the correlation matrix. This correlations could imply that as one variable increases, the other tends to increase as well. For simplicity and to avoid redundancy, we describe this pattern once, and the same interpretation applies to all scatter plots of the variables: Clump Thickness, Uniformity of Cell Size, Uniformity of Cell Shape, Marginal Adhesion, Single Epithelial Cell Size, Bare Nuclei, Bland Chromatin, and Normal Nucleoli. As an example, we consider the scatter plot of Clump Thickness vs Bare Nuclei (Figure 3). This plot shows positive correlation, where higher clump thickness is associated with greater presence of bare nuclei. This suggests that tumors with thicker

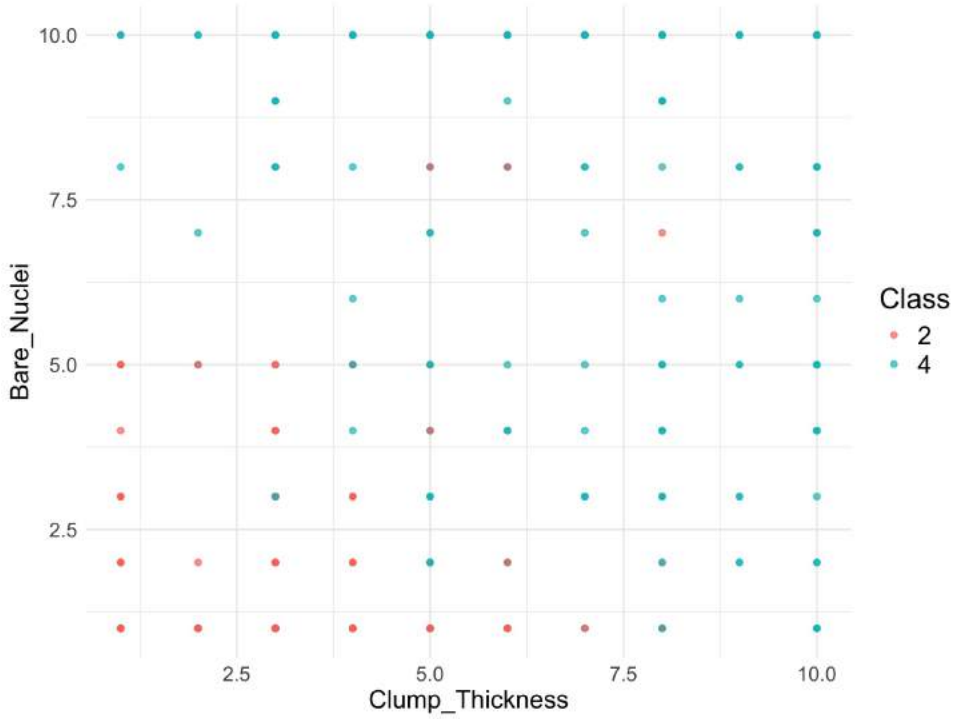

FIG. 3: Scatter plot of Clump Thickness vs Bare Nuclei.

clumps also tend to have more bare nuclei. This scatter plot shows the general pattern observed across the other pairs of variables with significant correlation.

The bivariate graphical analysis, supported by the correlation matrix and by the scatter plots confirms the positive correlations of some pairs of variables. Thus this provides more insight into the characteristics that can differentiate benign and malignant tumors.

3 Model fitting

First, we define the logistic regression model with interaction terms, including both main effects and interaction effects. We first define the following: β_0 is the intercept term, $\{\beta_i\}_{i=1}^9$ are the coefficients for the main effects, and β_{ij} are the coefficients for the interaction effects. We then have the linear predictor:

$$
\begin{aligned}
z &= \beta_0 + \sum_{i=1}^{9} \beta_i X_i + \beta_{12}(X_1X_2) + \beta_{23}(X_2X_3) + \beta_{14}(X_1X_4) + \beta_{16}(X_1X_6) \\
&= \beta_0 + \sum_{i=1}^{9} \beta_i X_i + \beta_{12} \cdot (\text{Clump Thickness} \cdot \text{Uniformity of Cell Size}) \\
&+ \beta_{23} \cdot (\text{Uniformity of Cell Size} \cdot \text{Uniformity of Cell Shape}) \\
&+ \beta_{14} \cdot (\text{Clump Thickness} \cdot \text{Marginal Adhesion}) \\
&+ \beta_{16} \cdot (\text{Clump Thickness} \cdot \text{Bare Nuclei}) \\
&:= \underbrace{\mathbf{X}\boldsymbol{\beta}}_{\text{Main effects}} + \underbrace{\widetilde{\mathbf{X}}\boldsymbol{\beta}_{ij}}_{\text{Interaction terms}}
\end{aligned}
$$

The probability outcome p is given by the logistic function applied to the linear predictor:

$$p = \sigma(z) = \frac{1}{1 + e^{-z}}.$$

For the Breast Cancer Wisconsin dataset, the logistic regression model is written as:

$$\log\left(\frac{p}{1-p}\right) = \mathbf{X}\boldsymbol{\beta} + \widetilde{\mathbf{X}}\boldsymbol{\beta}_{ij} \underset{\text{after fitting the model to training data}}{\Longrightarrow} \log\left(\frac{\hat{p}}{1-\hat{p}}\right) = \mathbf{X}\widehat{\boldsymbol{\beta}} + \widetilde{\mathbf{X}}\widehat{\boldsymbol{\beta}}_{ij}$$

where p is the probability of the tumor being malignant, $\hat{p}$ is the predicted probability of the tumor being malignant, and $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_{16}$ are the estimated coefficients.

Maximum Likelihood Estimation (MLE) is used to find the parameters that maximize the likelihood given the observed data:

$$L(\beta) = \prod_{i=1}^{N} p_i^{y_i}(1-p_i)^{1-y_i}, \quad \text{and} \quad \log(L(\beta)) = \sum_{i=1}^{N} \left(y_i \log(p_i) + (1-y_i)\log(1-p_i)\right),$$

where p_i is the predicted probability for observation i, and y_i is the binary outcome (1 for malignant, 0 for benign).

The initial step in model selection was to include all predictor variables to establish a baseline model. Given the observed correlations between features, we added interaction terms, that were chosen based on the correlation analysis. To evaluate and compare models, the AIC was used. The AIC value of the model is given by the following [1]:

$$\text{AIC} = 2k - 2\ln\hat{L}$$

where k is the number of estimated parameters and $\hat{L}$ is the maximum value of the log-likelihood. We kept the model with the lower AIC value. The model selection process involved fitting an initial logistic regression model including all main effects. Interaction terms were then added based on the correlation analysis. For each model iteration, the AIC was calculated, and the model with the lowest AIC was selected as the final model. The final model included both main effects and interaction terms that provided the optimal predictive accuracy and simplicity.

4 Model assessment

This section assesses whether the assumptions of the model are met. The dependent variable, the class of the tumor (benign or malignant) is binary (Section 1.4, Hosmer, 2013 [2]). The data consists of individual observations of tumors, which are assumed to be independent of each other (Section 1.4, Hosmer, 2013 [2]). We do scatter plots of the logit vs each predictor to assess the linearity assumption (Section 4.3.1, Hosmer, 2013 [2]) and plot them in Figure 4. We observe a linear trend, indicating that the assumption is met.

For the VIF table, the predictor names have been abbreviated for simplicity. Multicollinearity was assessed using the *Variance Inflation Factor* (VIF) (Section 4.3.3, Hosmer 2013 [2]). VIF measures the impact of collinearity among the predictor variables in our regression model. For a given predictor, X_i, the VIF is computed using Kutner's formula [3]:

$$\text{VIF}(X_i) = \frac{1}{1 - R_i^2}$$

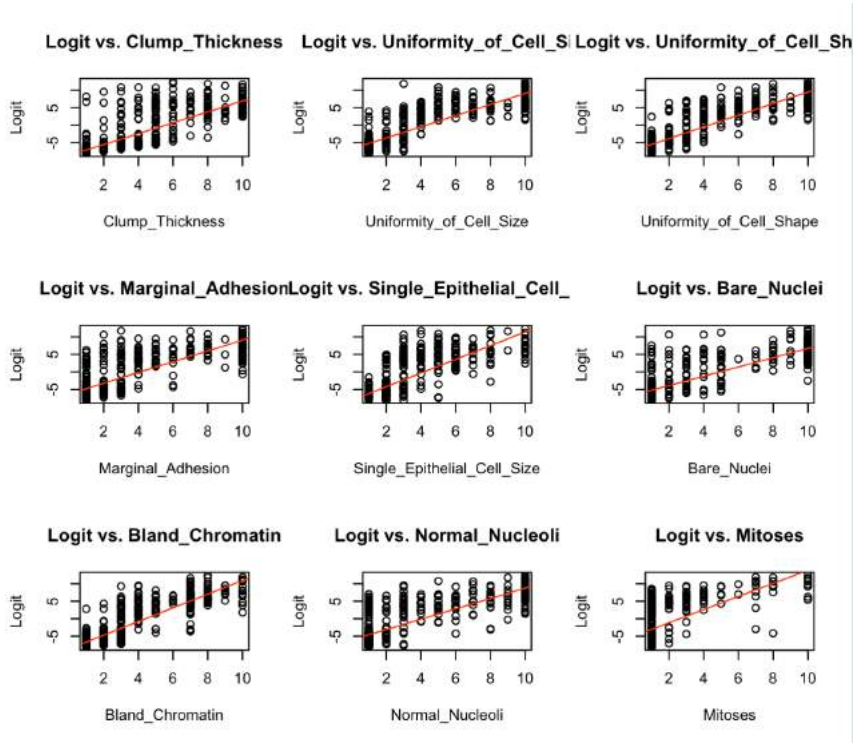

FIG. 4: Linearity of Logit

C.T	U.C.S	U.C.Sh	M.A	S.E.C.S	B.N	B.C
3.5	4.2	4.1	2.9	2.8	3.6	3.2
N.N	**M**	**C.T * U.C.S**	**U.C.S * U.C.Sh**	**C.T * M.A**	**C.T * B.N**	
3.4	2.5	5.0	4.7	3.8	3.9	

TABLE 2: Variance Inflation Factor (VIF) values for each predictor variable in the logistic regression model.

where R_i^2 is the R^2 value from a regression of predictor X_i on all other predictors in the model. By the *rule of thumb*, we use the fact that if the VIF > 10, it's a value considered to be large. Therefore, in our case, by looking at Table 2 we have no significant multicollinearity among the predictors.

Recall that the *Cook's distance*, from Montgomery's book [4], is defined as follows:

$$C_j = \frac{r_j^2 h_{jj}}{p(1 - h_{jj})},$$

where r_j is the outlier, and $h_{jj} = ((\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X})_{jj}$ is the leverage point. We use Cook's distance and standardized residuals to identify outliers and influential points (Section 4.3.2, Hosmer 2013 [2]). The results are shown in Figures 5 and 6. No significant points where found as indicated by the plots.

The logistic regression model's assumptions were assessed according to the guidelines provided by Hosmer (2013) [2] and found to be reasonably met. The binary nature of the outcome and the independence of observations were confirmed. The scatterplots of the logit versus predictors indicated a linear relationship. The VIF values showed no evidence of multicollinearity. Cook's distance and standardized residuals identified no significant outliers or influential points. These diagnostics confirm that the logistic regression model

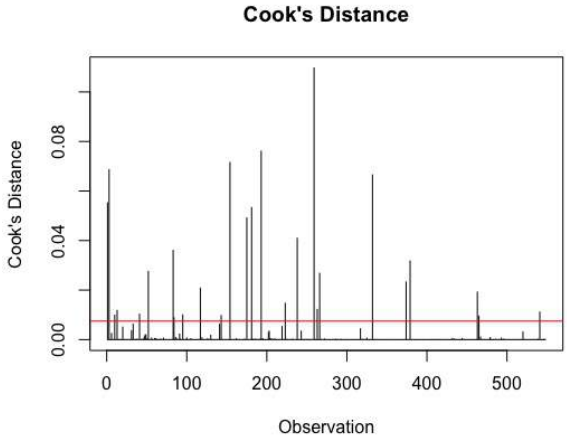

FIG. 5: Cook's distance

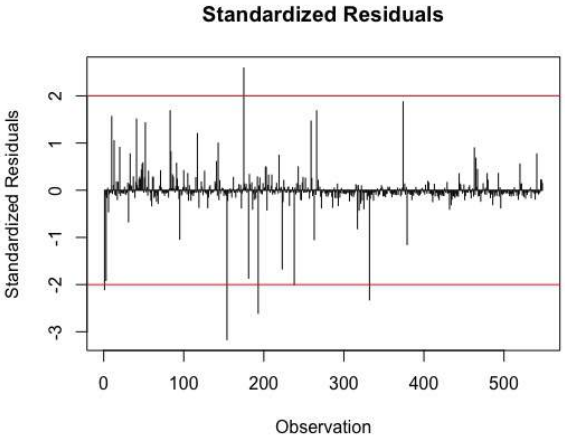

FIG. 6: Standardized residuals

is appropriate for the data.

5 Final Estimated Model

In this section, we present the final estimated logistic regression model, including the interaction terms. The coefficients are presented with a maximum of two significant digits for clarity. The estimated coefficients are presented in the table below:

$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\beta}_5$	$\hat{\beta}_6$
-3.27	0.54	0.65	0.74	0.44	0.34	0.56
$\hat{\beta}_7$	$\hat{\beta}_8$	$\hat{\beta}_9$	$\hat{\beta}_{12}$	$\hat{\beta}_{23}$	$\hat{\beta}_{14}$	$\hat{\beta}_{16}$
0.41	0.43	0.29	0.15	0.21	0.17	0.11

TABLE 3: Estimated coefficients for the final logistic regression model.

To provide a clearer understanding of the model coefficients, we calculated the odds ratios and their 95% confidence intervals and plotted them in Figure 7. The odds ratio for a coefficient indicates the change in odds of the outcome when the predictor variable increases for one unit. Figure 7 displays the odds ratios for each predictor and also their 95% confidence intervals. Most predictor variables have odds ratios greater than 1, indicating that increases in these variables are associated with higher odds of the tumor being malignant. As an example, for Clump Thickness, the odds ratio is 1.72 and the 95%-C.I is $(1.41, 2.11)$. For each one-unit increase in Clump Thickness, the odds of the tumor being malignant increase by 72%. This indicates that thicker cell clumps are associated with a higher likelihood of malignancy. We can apply the same reasoning for the rest of predictor variables. By interpreting these coefficients and odds ratios, we can better understand the influence of each predictor on the probability of a tumor being malignant. We observe that variables with higher odds ratios have more impact on increasing the likelihood of malignancy (for example, Uniformity of Cell Shape with an odds ratio of 2.10 and so an increase by 110%, suggesting a strong association with malignancy).

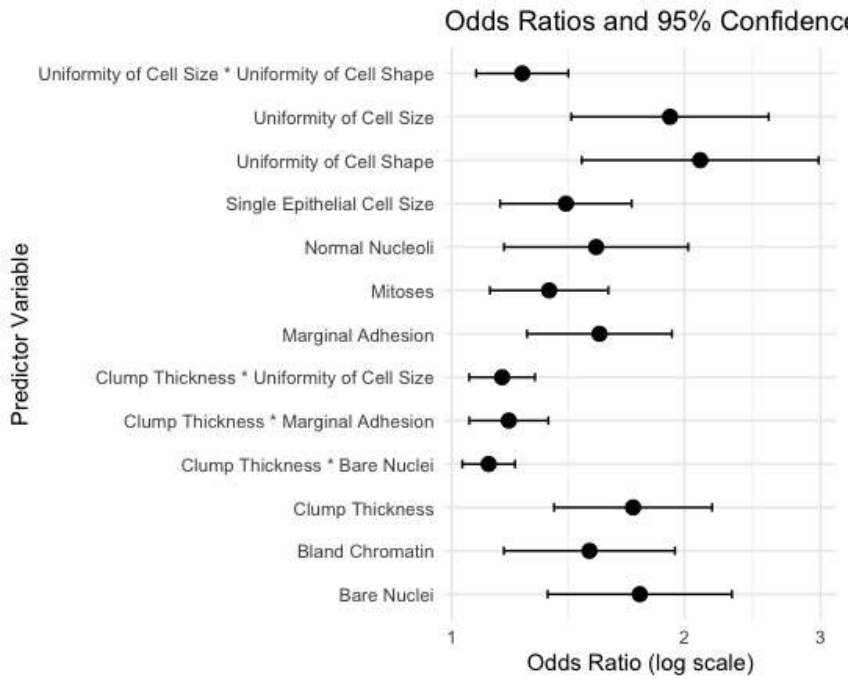

Fig. 7: Odds Ratios and 95% Confidence Intervals

6 Conclusion

In this report, we developed and assessed a logistic regression model to classify tumors as benign or malignant based on the dataset from the Wisconsin Breast Cancer Dataset (Original). Our analysis began with an EDA where we summarized the central tendency and spread of each variable, identifying skewness in the distributions. Histograms revealed right-skewed distributions for most variables, suggesting that while most observations have low values, higher values could indicate malignancy. A correlation matrix and heatmap were generated to visualize relationships between variables, revealing strong positive correlations between pairs. Scatter plots of these highly correlated pairs confirmed the positive relationships and suggested that higher values in these pairs might be associated with malignancy.

For model fitting, we defined a logistic regression model that included both main effects and interaction terms, identified based on correlation analysis. The model was fitted using Maximum Likelihood Estimation (MLE), and the Akaike Information Criterion (AIC) was employed to select the final model. The final model included significant main effects and interaction terms, providing the best predictive accuracy.

In the model assessment, we evaluated the assumptions of the logistic regression model. We followed the guidelines of Hosmer (2013) [2], confirming that the model assumptions were reasonably met. Our final estimated model provided a comprehensive representation of the relationship between predictor variables and the probability of a tumor being malignant. The model's coefficients provided an understanding of how each variable influences the outcome. Furthermore, the calculation of odds ratios and their 95% confidence intervals offered additional insights into the impact of each predictor. For instance, higher odds ratios indicated that increases in these predictors significantly raised the likelihood of malignancy, enhancing the interpretability of the model.

Overall, this report shows the effectiveness of logistic regression to predict if the tumor

is malignant using the Wisconsin Breast Cancer Dataset.

A Appendix : Detailed histograms

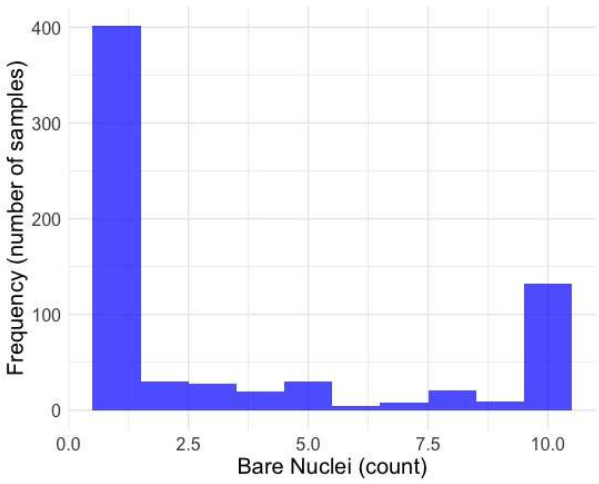

FIG. 8: Histogram of Bare Nuclei

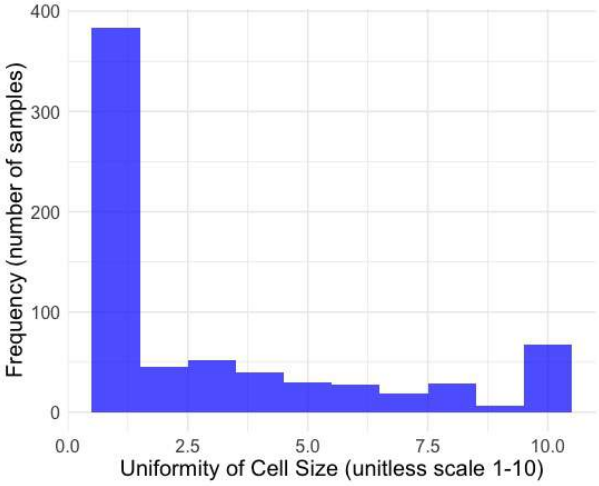

FIG. 9: Histogram of Uniformity of Cell Size.

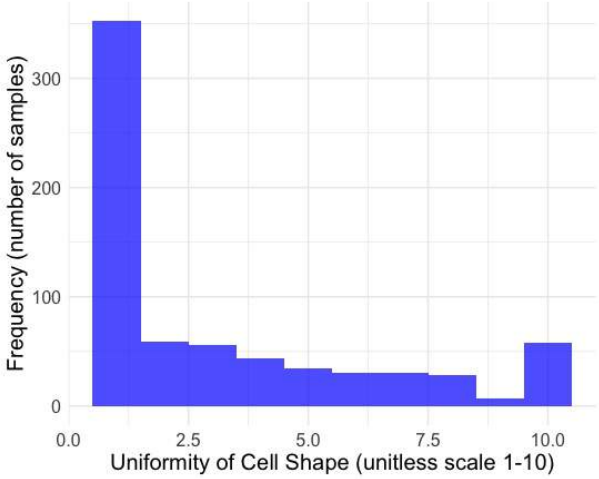

FIG. 10: Histogram of Uniformity of Cell Shape.

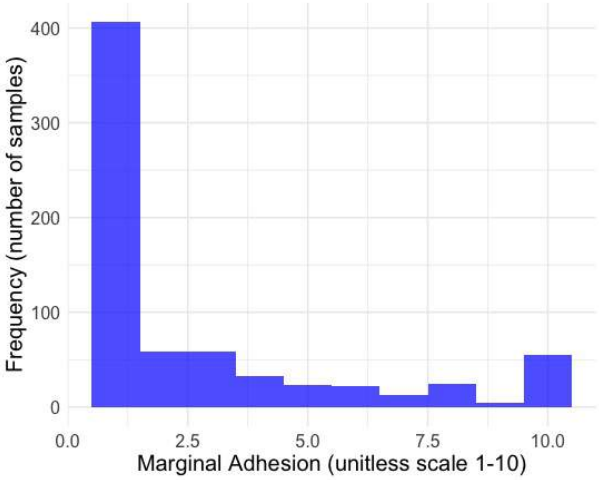

FIG. 11: Histogram of Marginal Adhesion.

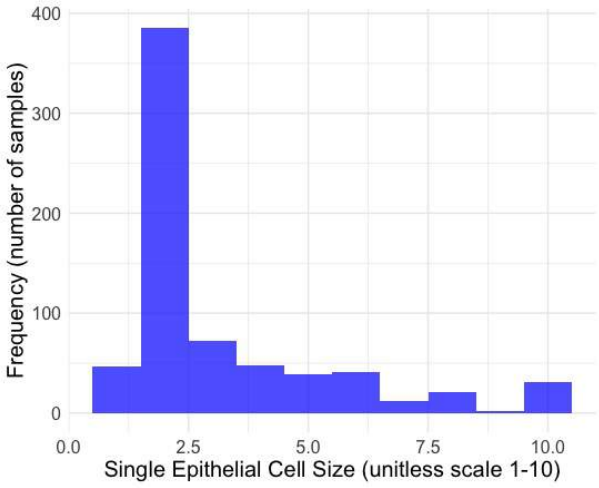

FIG. 12: Histogram of Single Epithelial Cell Size.

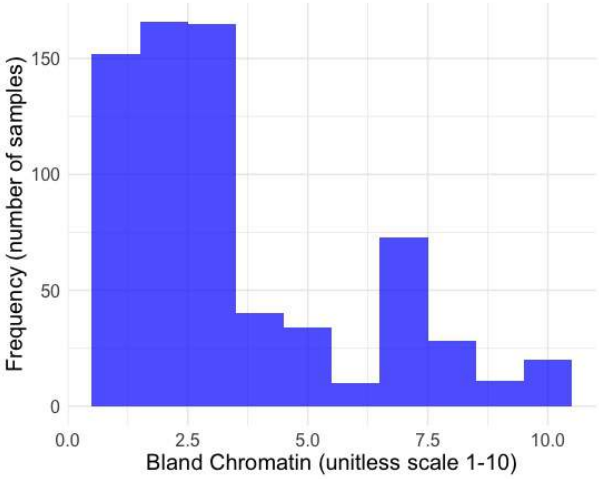

FIG. 13: Histogram of Bland Chromatin.

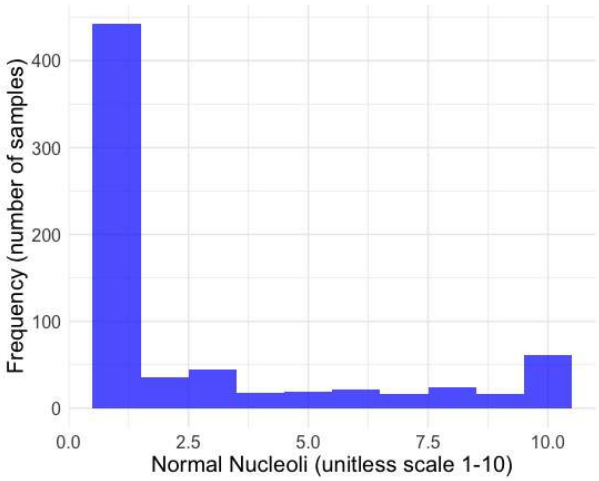

FIG. 14: Histogram of Normal Nucleoli.

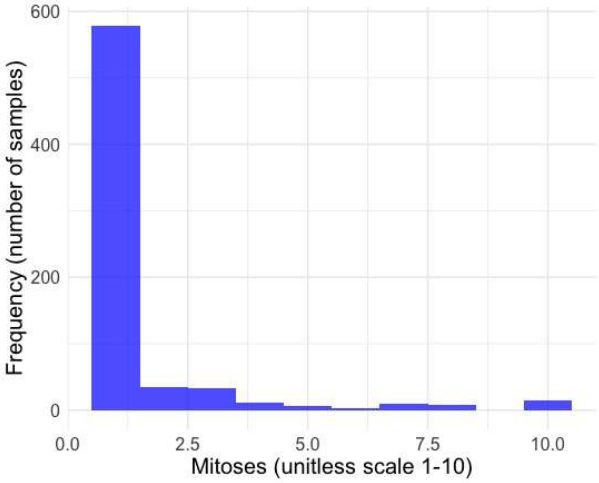

FIG. 15: Histogram of Mitoses.

B Appendix: Detailed Scatter plots

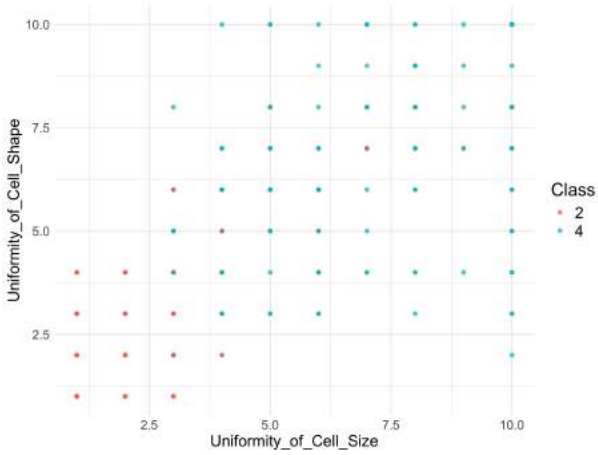

FIG. 16: Scatter plot of Uniformity of Cell Size vs Uniformity of Cell Shape.

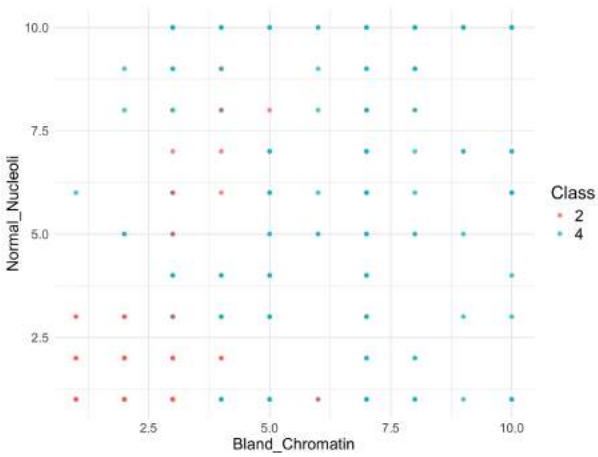

FIG. 17: Scatter plot of Bland Chromatin vs Normal Nucleoli.

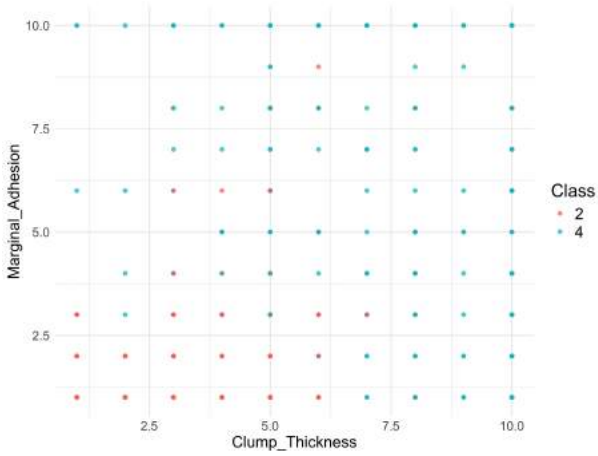

FIG. 18: Scatter plot of Clump Thickness vs Marginal Adhesion.

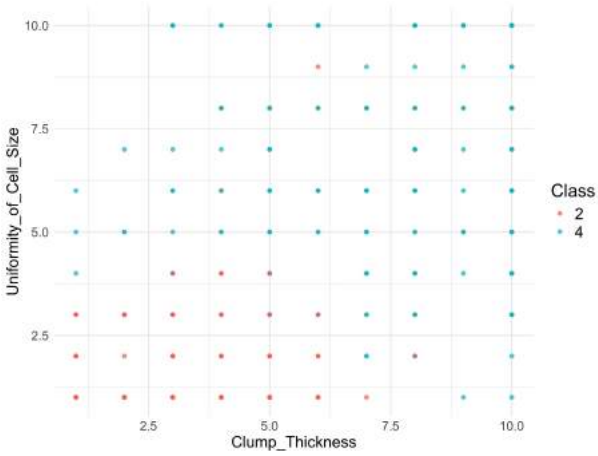

FIG. 19: Scatter plot of Clump Thickness vs Uniformity of Cell Size.

Remark: for the scatter plots, *Class 2* indicates that the tumor is benign, and *Class 4* indicates that the tumor is malignant.

References

[1] K.P. Burnham. *Model Selection and Multimodel Inference*. Springer, 2014. ISBN: 9781475777116.

[2] David W Hosmer, Stanley Lemeshow, and Rodney X Sturdivant. *Applied Logistic Regression*. John Wiley & Sons, 2013.

[3] Michael H. Kutner et al. *Applied Linear Statistical Models*. 5th. Boston, MA: McGraw-Hill/Irwin, 2004. ISBN: 978-0073108742.

[4] Douglas C. Montgomery, Elizabeth A. Peck, and Geoffrey G. Vining. *Introduction to Linear Regression Analysis*. 5th. Wiley, 2012. ISBN: 978-0470542811.

[5] Akash Nag and Soumya Sarkar. “Identifying Patients at Risk of Breast Cancer through Decision Trees”. In: *International Journal of Advanced Computer Research* 8 (Oct. 2017).