Breast Cancer Analysis Report

Rutuja Jagtap (230118658)

2023-12-01

INTRODUCTION

In this report, we will conduct a comprehensive analysis of the BreastCancer dataset, focusing on the distinctive characteristics of breast tissue samples collected from a cohort of 699 women. The dataset encapsulates valuable information that can shed light on various aspects of breast health and provide insights into the features associated with tissue samples. Each sample has nine easily assessed cytological characteristics that were measured. These characteristics are measured on a scale from 1 to 10, with a lower number indicating healthier cells for that specific characteristic, and a higher number indicating less healthy or unhealthy cells. There is a response column in the dataset that shows historic examination values indicating whether the sample was benign or malignant.

OBJECTIVE

The aim of this project is to first conduct data cleaning and exploratory data analysis (EDA) on the BreastCancer dataset. The next step involves constructing classifiers through different variants of classification model for the predicting classes of tissue samples based on these nine cytological features. Lastly, comparing the performance of these models using cross-validation results and recommending the most effective classifier.

DATA UNDERSTANDING

In the BreastCancer dataset, there are 699 observations on 11 variables, one being a character variable (id column), 9 being ordered or nominal (9 cytological characteristic variables), and 1 target class (class column).

Description of variables :

- Id : Sample code number
- Cl.thickness : Clump Thickness
- Cell.size : Uniformity of Cell Size
- Cell.shape : Uniformity of Cell Shape
- Marg.adhesion : Marginal Adhesion
- Epith.c.size : Single Epithelial Cell Size
- Bare.nuclei : Bare Nuclei
- Bl.cromatin : Bland Chromatin
- Normal.nucleoli : Normal Nucleoli
- Mitoses : Mitoses
- Class : Class (benign or malignant)

The 'id' column is of character type, while the rest are factors. The cytological characteristic variables are explicitly encoded as factors and will be considered as quantitative variables in this project.

The predictor variables are ‘Cl.thickness’, ‘Cell.size’, ‘Cell.shape’, ‘Marg.adhesion’, ‘Epith.c.size’, ‘Bare.nuclei’, ‘Bl.cromatin’, ‘Normal.nucleoli’ and ‘Mitoses.’ The predictor column is used to predict response variable.

The response variable is ‘Class.’ The response variable is what we are trying to predict. The class variable indicated whether the tissue sample is benign or malignant type.

DATA PREPARATION AND CLEANING

We are using ‘mlbench’ package. mlbench is an R package that provides a collection of artificial and real-world machine learning benchmark datasets.To use the mlbench package, I have installed it first, and then loaded BreastCancer dataset.

I have created a new dataframe ‘df_breastCancer’ which contains the BreastCancer data itself. Going forward, I will be making transformations on this dataframe in this analysis project. This will ensure that transformations are done on dataframe ‘df_breastCancer’ and original data will not be hampered and can be viewed in original format while making comparison study between altered data and original data.

Further, I have examined the uniqueness of the samples and discovered that out of 699 samples, there are 645 unique sample IDs. However, each of these sample IDs does not have the exact same observations for the 9 characteristic variables. Therefore, I am considering including these samples in the BreastCancer dataframe, as it also contributes to valuable information for determining the target class.

To identify missing values in the dataset, I employed is.na function on data and got the following summary and this summary clearly displays that only the ‘Bare.nuclei’ column contained missing values, encoded as ‘NA’. Approximately 16 NA values were identified in the ‘Bare.nuclei’ column of the dataset which can be seen below. I subsequently removed these rows from the dataset. As a result, we now have 683 samples remaining to conduct our analysis.

```
##      Id          Cl.thickness    Cell.size       Cell.shape
##  Mode :logical   Mode :logical   Mode :logical   Mode :logical
##  FALSE:699       FALSE:699       FALSE:699       FALSE:699
##
##  Marg.adhesion   Epith.c.size    Bare.nuclei     Bl.cromatin
##  Mode :logical   Mode :logical   Mode :logical   Mode :logical
##  FALSE:699       FALSE:699       FALSE:683       FALSE:699
##                                  TRUE :16
##  Normal.nucleoli  Mitoses          Class
##  Mode :logical   Mode :logical   Mode :logical
##  FALSE:699       FALSE:699       FALSE:699
##
```

Moving to the ‘Id’ column in data, the column does not contribute to provide meaningful information in our analysis, that is why I have removed it from the dataframe.

‘Class’ column contains non-numeric data so I have converted it to numeric format by applying as.numeric function on ‘Class’ column and created a new column called ‘Class_numeric’ out of it in the dataframe. In ‘Class_numeric’ - 0 represents ‘benign’ class and 1 represents ‘malignant’ class. I’m doing this as classification model reads numeric data only and cannot understand non numeric data.

Similarly, I converted other encoded cytological characteristic variables to numeric format using the as.numeric function and stored the changes in the dataframe. The type of variables changed from factor to double. I have converted non-numeric data to numeric format for better readability during analysis and to facilitate easier interpretation by the model.

EXPLORATORY DATA ANALYSIS

Out of 683 tissue samples from women, the Table 1 indicates a higher count of the benign class, suggesting that benign breast cancer is nearly twice as prevalent as malignant cancer in this dataset.

Table 1: Class Distribution

Tissue Class	Count
benign	444
malignant	239

Swarm Plot :

Swarm plot is basically a scatter plot where x-axis is representing cytological characteristics. The blue points on the plot represent tissue classified as benign, and the red points represent tissue classified as malignant. The benign tissue is mainly concentrated in the lower half of the value range for all cytological characteristics, while the malignant tissue is spread across the upper half of the value range. In the swarm plot in Figure 1, there is a distinct and clear separation between the two classes. The features “Cl.thickness” and “Bl.cromatin” particularly show a noticeable separation compared to the other characteristics. These two characteristics are also concentrated upto higher scale than others. Very less benign tissue sample show characteristc value more than 7 and only two to three sample of benign type have “Marg.adhesion” and “Epith.c.size” values in highest scale 10. Malignant tissue exhibits a broader range of values or spread compared to benign tissue, indicating greater variability within the malignant class. Based on the plot, there are high chances that higher values of any characteristic suggest the presence of malignant tissue, while lower values suggest benign tissue.

I chose a swarm plot to clearly display the classification between benign and malignant tissue for each characteristic. This plot illustrates the relationship between predictor variables and the response variable.

Heat Map :

The heatmap in Figure 2 illustrates the correlation between predictor variables in the dataset. In the figure below, the darkest blue color represents the maximum correlation, while lighter or red colors indicate lower or negative correlation. White color denotes no correlation.

It is evident that the ‘cell.size’ and ‘cell.shape’ features exhibit the highest correlation, reaching a correlation value of approximately 0.8. On the other hand, the ‘Mitoses’ feature demonstrates the least correlation with any other feature in the dataset, with values ranging from 0 to 0.4, as indicated in the legend.

The ‘Cell size’ feature demonstrates considerable correlation with other features, except ‘Mitoses,’ with values ranging between 0.6 and 0.8. Negative correlation is not observed in this dataset, as there are no red boxes in the heatmap.

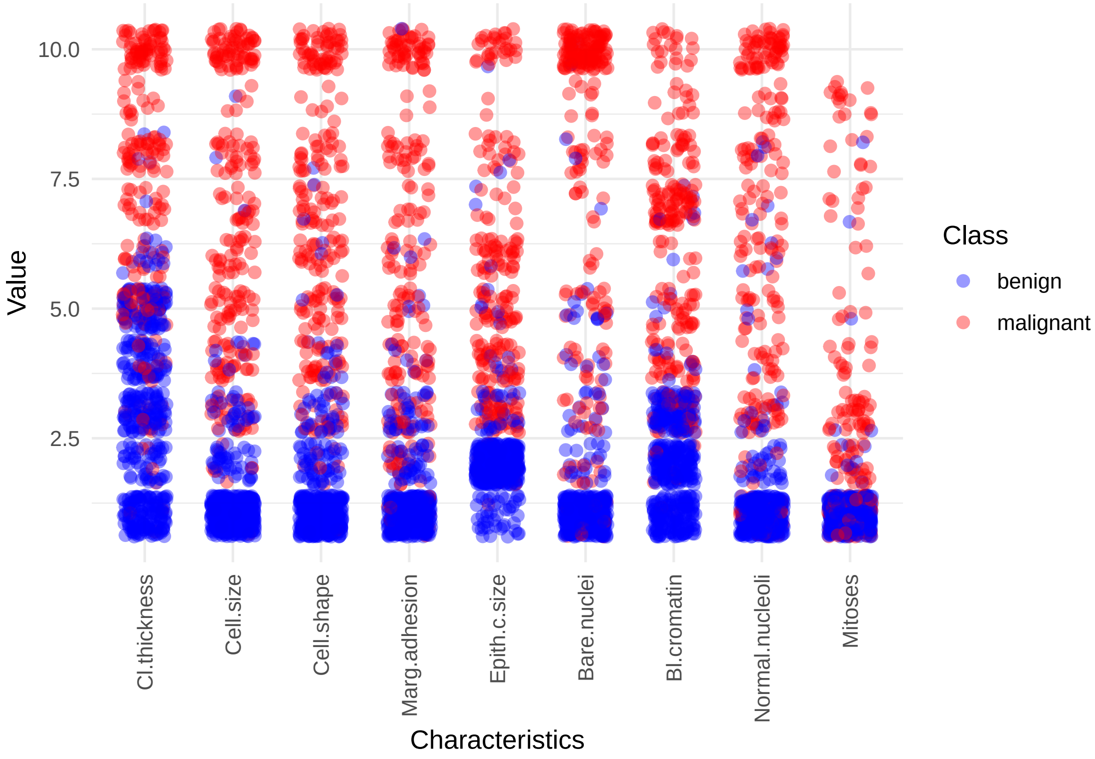

Figure 1: Swarm Plot of characteristics by class

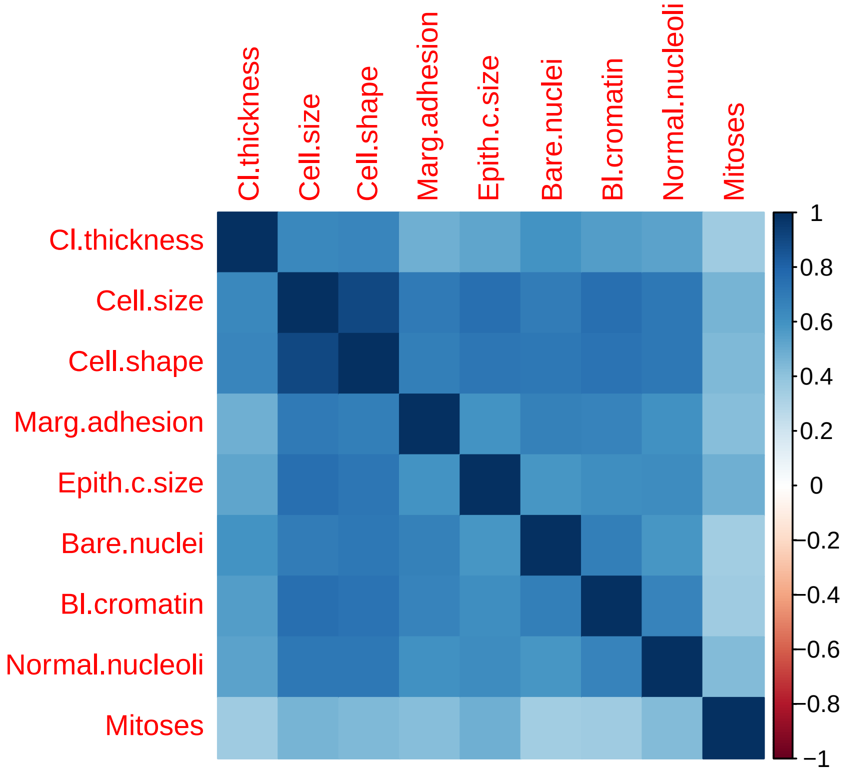

Figure 2: Correlation Heatmap of Cytological Features

BUILDING DIFFERENT VARIANTS OF CLASSIFICATION MODELS

To classify response variable based on predictor variables, it is essential to perform feature selection as focusing on only insightful feature will increase the accuracy and performance of the model. So in this report I have performed classification on Logistic Regression Model using three techniques :

- Backward Stepwise Subset Selection
- Least Absolute Shrinkage and Selection Operator (LASSO) and Ridge Regularization
- Linear Discriminant Analysis (LDA)

1. Backward Stepwise Subset Selection Method

I have chosen to use backward stepwise selection method for feature selection because in this dataset as we have many exploratory variables. Moreover, this method is computationally efficient as compared to best subset selection method and reduces risk of overfitting. It also helps to avoid multicollinearity and generates more simpler model by reducing complexity as it eliminates variables step by step. Backward stepwise subset selection model also performs better than forward stepwise subset selection model as it initially considers all predictors and successively reduces least useful predictors which makes model more generalized and gives better accuracy on unseen data as well.

Initially, I divided dataset into x(all predictor variables data) and y(response variable data). Further, I scaled the x set because all the variables had the same range of values from 1 to 10; however, the 'Mitoses' column had a range of values from 1 to 9. Since the data was not completely standardized, I preferred standardizing it for better model performance.

The backward stepwise selection process during analysis began by constructing a model that incorporated all nine predictor variables, resulting in an initial AIC value of 122.9. During the first step, the AIC improved to 120.9; however, the variable 'Cell.size' was removed from the model as its inclusion did not contribute to significance in AIC, showing the lowest AIC value among all variables.

In the second step, the AIC further improved to 119.28, leading to the removal of the variable 'Epith.c.size' from the model. Hence, the final model consists of the variables: 'Cl.thickness', 'Cell.shape', 'Marg.adhesion', 'Bare.nuclei', 'Bl.cromatin', 'Normal.nucleoli', and 'Mitoses.'

Mathematically, the AIC value is calculated using the following formula:

AIC = -2 x Log-Likelihood + 2 x Number of Parameters

Log-Likelihood: This is a measure of how well the model explains the observed data. The higher the log-likelihood, the better the fit of the model to the data.

Number of Parameters: This represents the complexity of the model, typically the number of predictors or variables used in the model. AIC penalizes models with more parameters, helping to prevent overfitting.

The AIC value is used for model selection in the context of comparing different models. Lower AIC values indicate a better trade-off between model fit and complexity. Therefore, we finally got model with lowest AIC. Apart from AIC , another important parameter to consider while choosing predictor is the coefficients of the predictors. According to Table 2 below it is evident that coefficient of these two variables were considerably low in original logistic model which got automatically removed with backward stepwise selection process as shown in Table 3. Therefore, the final best model of backward subset selection is generated with 7 predictors.

Table 2: Logistic Regression Coefficients

Predictors	Coefficient
Cl.thickness	1.50983
Cell.size	-0.01822
Cell.shape	0.96273
Marg.adhesion	0.94729
Epith.c.size	0.21519
Bare.nuclei	1.39565
Bl.cromatin	1.09600
Normal.nucleoli	0.65044
Mitoses	0.88124

Table 3: Backward Stepwise selection Coefficients

Predictors	Coefficient
Cl.thickness	1.50698
Cell.shape	1.03115
Marg.adhesion	0.98142
Bare.nuclei	1.41490
Bl.cromatin	1.13229
Normal.nucleoli	0.69045
Mitoses	0.87601

2. Regularization with LASSO and RIDGE penalty

Least Absolute Shrinkage and Selection Operator (LASSO) is a helpful technique in logistic regression that automatically picks the most important variables from data, it prevents overfitting and simplifies models by shrinking some coefficients to zero. It handles multicollinearity, promotes model interpretability and offers a bias-variance trade-off through parameter tuning. This techniques is applied when number of predictors are more.

Ridge penalty tends to shrink coefficients towards zero but retains all predictors in the model. LASSO combines shrinkage with variable selection, setting some coefficients to exactly zero and, therefore, performing automatic feature selection. Ridge penalty is beneficial when all predictors are expected to contribute, while LASSO is preferred when there is need to identify a subset of important predictors. In our report we are focusing on reducing number of features in dataset to understand only important features that help in better classification. Therefore I had initially chosen to use LASSO technique over ridge so that unwanted variable coefficient drops to zero and we can build enhanced classification model. However, I found that final lasso model did not converge to exact zero coefficient value for any predictor variable so I also built another model with ridge regularization to conduct comparison study between these two techniques.

- LASSO Model : - glmnet function is used to fit predictors, response variables, as it has to be logistic regression model family is set to binomial, alpha is set to 1 indicating use of lasso penalty and grid value is added for parameter tuning.
- RIDGE Model : - It is same as lasso except the apha value is set 0 indicating use of ridge penalty.

Note - Standardization during model fitting is set to FALSE is both techniques as standardization was already done.

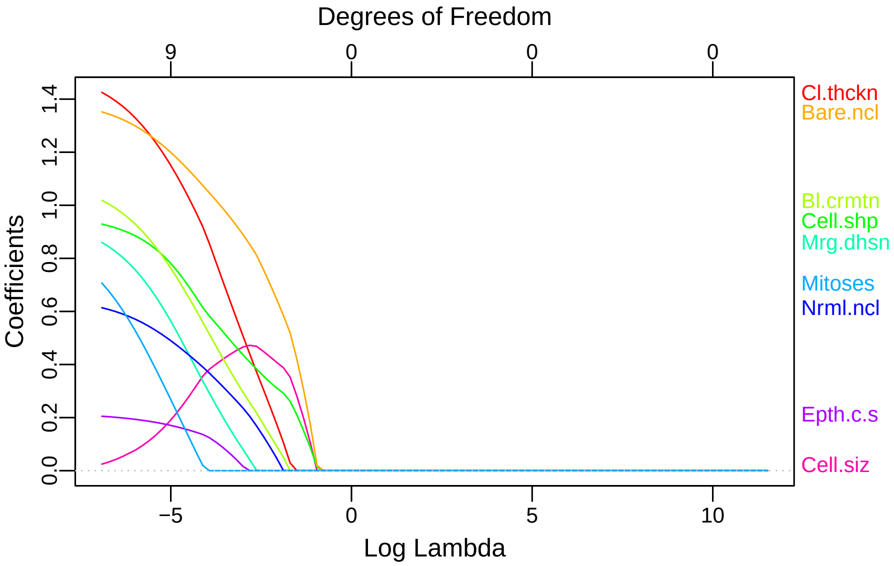

Figure 3: LASSO Regularization Model Plot

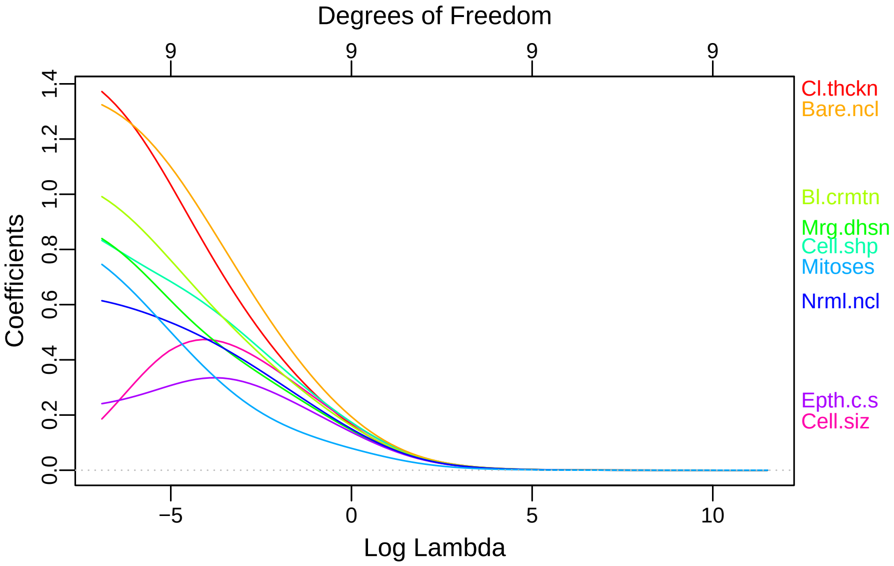

Figure 4: RIDGE Regularization Model Plot

The resulted plot in Figure 3 and Figure 4 is showing the effect of the tuning parameter (lambda) on the regularization path for each predictor in the logistic regression model using LASSO and RIDGE method respectively.

Figure 3 and Figure 4 show tuning parameter (lambda) on x-axis, different color lines are regression coefficients of different predictors and y-axis shows the coefficient value.

I utilized the additional 'plotmo' R library to label lines with exact variable names on the right hand side of the plot, avoiding the default numbering pattern that caused overlapping digits for variables with nearly similar coefficient values. Labeling with the exact variable names reduces visual confusion and maintains clarity in conveying the correct information to the audience.

From both the plots above, we can see that all coefficients are positive. Only 'cell.size' had lowest initial coefficient (In Figure 3 value is around 0 and in Figure 4 value is around 0.2). Also, 'Epith.c.size' variable is second lowest of all in both the plots. Overall plot depicts that with increase in lambda of predictor , there is decrease in coefficient value and eventually, coefficient of all predictors are pushed to zero. We can see in LASSO plot, that by the time lambda is slightly more than zero, all the variables converge to coefficient zero. There is drastic drop to zero observed in LASSO. On the other hand, in RIDGE plot, coefficients of all predictors together have gradually and smoothly shrunk to zero at the same point which is slightly before Log of Lambda 5.

The coefficient for 'Mitoses' dropped to zero first, while 'Bare.nuclei' dropped last. Depending on lambda we get some predictor subset selection between -5 and 0 log of lambda value in both plots.

The numbers at the top of the plot indicate the number of coefficients which are not equal to zero. In LASSO plot its seen zero coefficients which are not at zero that means all variables have touched to zero coefficient. While in RIDGE plot, all nine predictors have not reached zero coefficient.

The perfect lambda or parameter tuning value which gives best predictive performance lies somewhere in between absolute loss function and shrinkage.

Further, I have performed cross-validation tuning parameter to select the optimal lambda value where we can achieve best predictive performance of model.

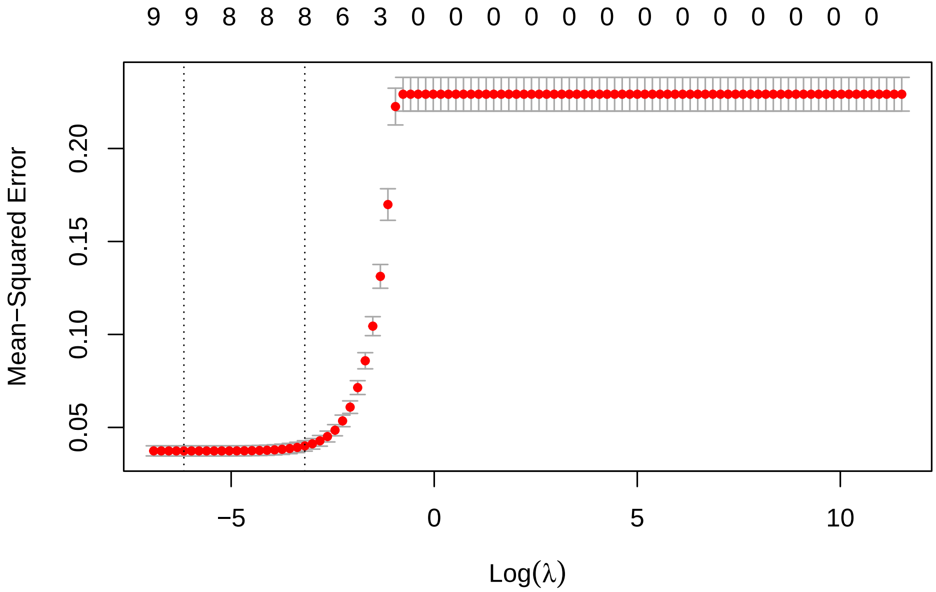

Figure 5: LASSO Regularization Mean Square Error Plot

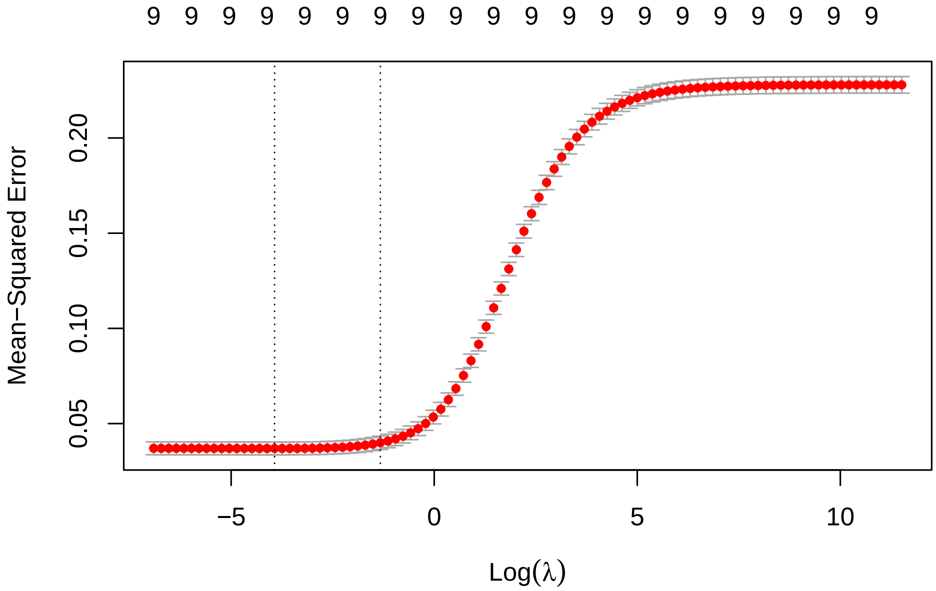

Figure 6: RIDGE Regularization Mean Square Error Plot

The plot of the cross-validated in Figure 5 and Figure 6 shows mean squared error rates against the log(lambda) values. We are looking for model that minimizes cross validation error.

Mean-squared error(MSE) is least which is below 0.05 and this is the region at which model should give best prediction performance. So LASSO solution lies somewhere before -5 log of lambda where there are two vertical dotted lines shown in plot , at this region the degree of freedom lies between 9 and 8. On the other hand, RIDGE MSE plot shows solution quite after -5 log of lambda with constant degree of freedom 9.

Mean Squared Error of Ridge is less than LASSO penalty regularization. While minimum Lambda of LASSO is greater than Ridge. This suggests that Ridge penalty regularization is better model than LASSO model.

```
## Minimum lambda value for lasso logistic regression model :  0.002104904


##
## Minimum lambda value for ridge logistic regression model :  0.01963041


##
## Corresponsing Mean Square Error value (LASSO):  0.03735276


##
## Corresponsing Mean Square Error value (RIDGE):  0.03692725
```

Now lets compare coefficients of LASSO and RIDGE with help of Table 4 and Table 5.

Table 4: Lasso Regularization Model Coefficients

Predictors	Coefficient
Cl.thickness	1.35310
Cell.size	0.06496
Cell.shape	0.89646
Marg.adhesion	0.78208
Epith.c.size	0.19625
Bare.nuclei	1.31202
Bl.cromatin	0.95083
Normal.nucleoli	0.58126
Mitoses	0.56838

Table 5: Ridge Regularization Model Coefficients

Predictors	Coefficient
Cl.thickness	0.78903
Cell.size	0.47293
Cell.shape	0.59167
Marg.adhesion	0.48351
Epith.c.size	0.33469
Bare.nuclei	0.89014
Bl.cromatin	0.60552
Normal.nucleoli	0.46996
Mitoses	0.35587

In both tables, the magnitudes of each coefficient suggest their relationship with the response variable. All predictors exhibit a positive coefficient relation with the response variable. The coefficients have shrunk close to zero values compared to the simple logistic regression model coefficients seen at the very beginning in Table 2. Specifically, 'Cell.size' and 'Epith.c.size' again show the lowest coefficient values among all variables in the LASSO model. Meanwhile, 'Epith.c.size' and 'Mitoses' exhibit the least coefficient values among all predictors in the RIDGE model. Some coefficients in LASSO are still above 1, which is not the case in the RIDGE model.

From this, we can observe that RIDGE regularization converges the coefficients to zero more effectively compared to LASSO.

3. Linear Discriminant Analysis (LDA)

In exploratory data analysis part of this report we saw swarm plot which clearly showed that benign and malignant classes are well separated across all the predictor variables. Therefore, I decided to use LDA over Quadratic Discriminant Analysis(QDA), as LDA method of classification is most preferred when classes are well-separated.

Table 6: LDA Model Coefficients

Predictors	Coefficient
Cell.size	0.38525
Cell.shape	0.26913
Marg.adhesion	0.13525
Epith.c.size	0.12798
Bare.nuclei	0.95268
Bl.cromatin	0.27100
Normal.nucleoli	0.32514
Mitoses	0.01406

From the given Table 6, it's evident that 'Bare.nuclei' has the highest coefficient value, indicating its significant contribution in determining the class type. On the other hand, 'Mitoses' exhibit lowest coefficient, suggesting a comparatively lesser impact on the class type determination.

Table 7: Group Mean of LDA model

Predictors	Benign(0)	Malignant(1)
Cl.thickness	-0.52404	0.97354
Cell.size	-0.60177	1.11792
Cell.shape	-0.60256	1.11941
Marg.adhesion	-0.51782	0.96197
Epith.c.size	-0.50657	0.94108
Bare.nuclei	-0.60315	1.12050
Bl.cromatin	-0.55589	1.03270
Normal.nucleoli	-0.52689	0.97883
Mitoses	-0.31620	0.58742

The group means presented in Table 7 depict the average values of each predictor variable for different classes (benign and malignant). Larger differences in the mean values between these classes suggest that those specific variables play a more informative role in distinguishing between benign and malignant cases. There shows least difference in Mitoses characteristic that means it contributes less in classification whereas largest difference is observed in Bare.nuclei characteristic that means it contributes more in classification i.e to some extent Bare.nuclei variables value can decided whether the tissue is benign or malignant.

COMPARING PERFORMANCE OF ALL MODELS USING CROSS VALIDATION TECHNIQUE

Cross validation is popular technique to evaluate error and performance of the model. Here we split data into k-folds. The first iteration goes on where 1st fold is used as test data and k-1 folds as train data. In second iteration, 2nd fold will be the test set and previous 1st fold and remaining k-1 folds will be train set, this goes on on until all k folds finish there turn of test data.

K-fold cross validation ensures that data is not wasted and testing is carried on all folds of data which is not possible in validation set approach or in-sample approach, therefore I decided to use out-of-sample or cross validation techniques to compute test errors and assess accuracy in above three variants of logistic models.

Comparison of all variants of logistic regression model is fair using k-fold technique as same number of folds are applied to models for cross validation steps. I chose to use 10 fold cross validation as it is a standard practice. It reduce bias in validation results and is also computationally less expensive.

I obtained means squared error values for each model as seen in Table 8 below. Based on the error value we can conclude that model having least error shall perform well and will give good accuracy. Therefore, RIDGE model with 0.0244 mean squared error is best model of all. While backward stepwise selection has performed poorly, resulting in the highest test error.

Table 8: K-fold Cross Validation Test Result

Classification_Model	Mean_Square_Error
Backward Stepwise Selection	35.59718
LASSO Regularization	0.02531
RIDGE Regularization	0.02438
LDA Classifier	1.00878

CONCLUSION

The best classifier is the regularized form of logistic regression. Both ridge and lasso have demonstrated the least test error compared to other classification models displayed in Table 8. Furthermore, they provide a clear understanding of important features in the data through coefficient and mean square error plots, illustrating how a model behaves for breast cancer data. To select one, I would say ridge regularization has performed better than LASSO. Initially, it was presumed that LASSO should have performed well, helping in automatic variable selection and giving exact zero values to coefficients. However, when comparing coefficients of both, ridge regularization showed a considerable reduction in coefficients towards zero. Whereas not a single variable in LASSO model was equal to zero coefficient.

Therefore, we can conclude that the Ridge regression model is the better model of the two, as it exhibits a lower test error value than LASSO also. It tends to shrink coefficients towards zero but retains all predictors in the model. There is no direct drop in variables. Ridge penalty logistic regression includes all predictor variables. It is the property of ridge penalty to retain all predictors, proving that all predictors are meaningful for the classification of data which has been already proved above in brief.