MAS8404 Statistical Learning for Data Science | Predicting Breast Cancer Malignancy: A Comparative Study of Supervised and Unsupervised Learning Approaches

Monisha Dabbara (Student ID: 240503817)

2024-11-16

Abstract

This study uses cytological features to predict breast tissue malignancy with unsupervised and supervised learning techniques. After cleaning the data, Exploratory Data Analysis (EDA) identified key variables, including Bare.nuclei, Cl.thickness, Cell.shape, Marg.adhesion, and Mitoses, linked to malignancy. Unsupervised learning methods like K-means and hierarchical clustering differentiated benign and malignant samples, with K-means performing best. In the supervised phase, models such as logistic regression with subset selection, Lasso-regularized logistic regression, and discriminant analysis (LDA and QDA) were evaluated. Model performance was assessed using training and testing errors, accuracy, and AUC. Logistic regression with subset selection had a training error of 0.0311 and a testing error of 0.0256. Lasso-regularized logistic regression performed similarly with a training error of 0.0275 and a testing error of 0.0291. LDA and QDA had slightly higher errors. These results suggest Lasso-regularized logistic regression is the most effective model for distinguishing between benign and malignant samples. Further validation is needed. Further validation and adjustments are recommended to improve robustness and generalizability.

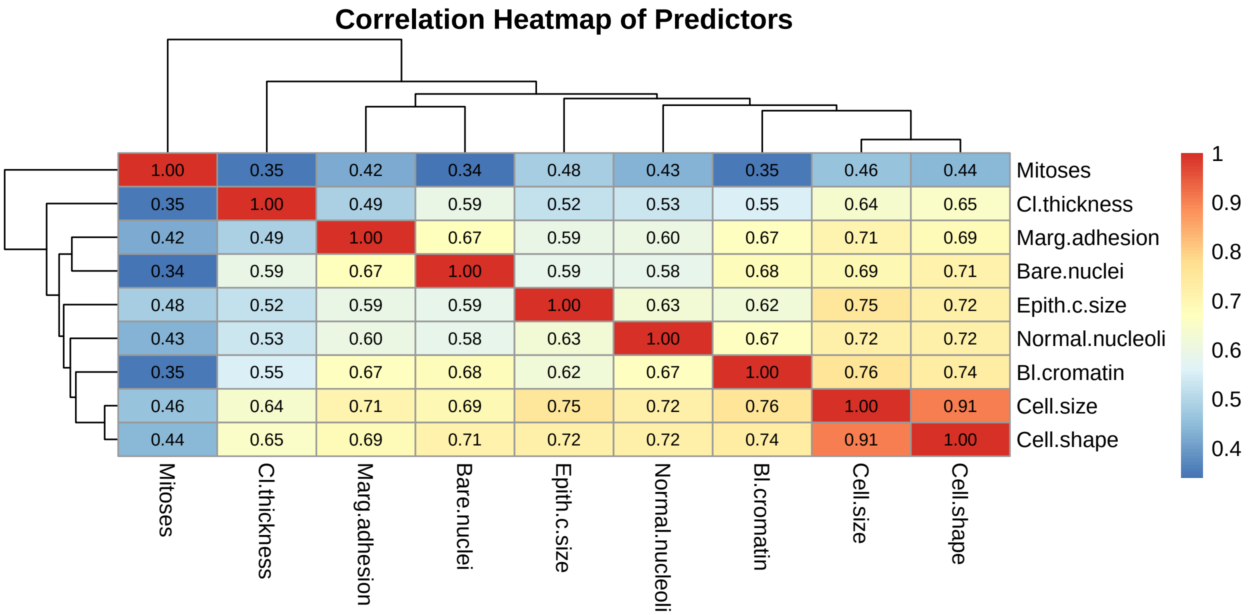
Exploratory data analysis: Data summary

Before starting with EDA, data cleaning was performed, which included converting factors to quantitative variables and removing rows with missing observations.

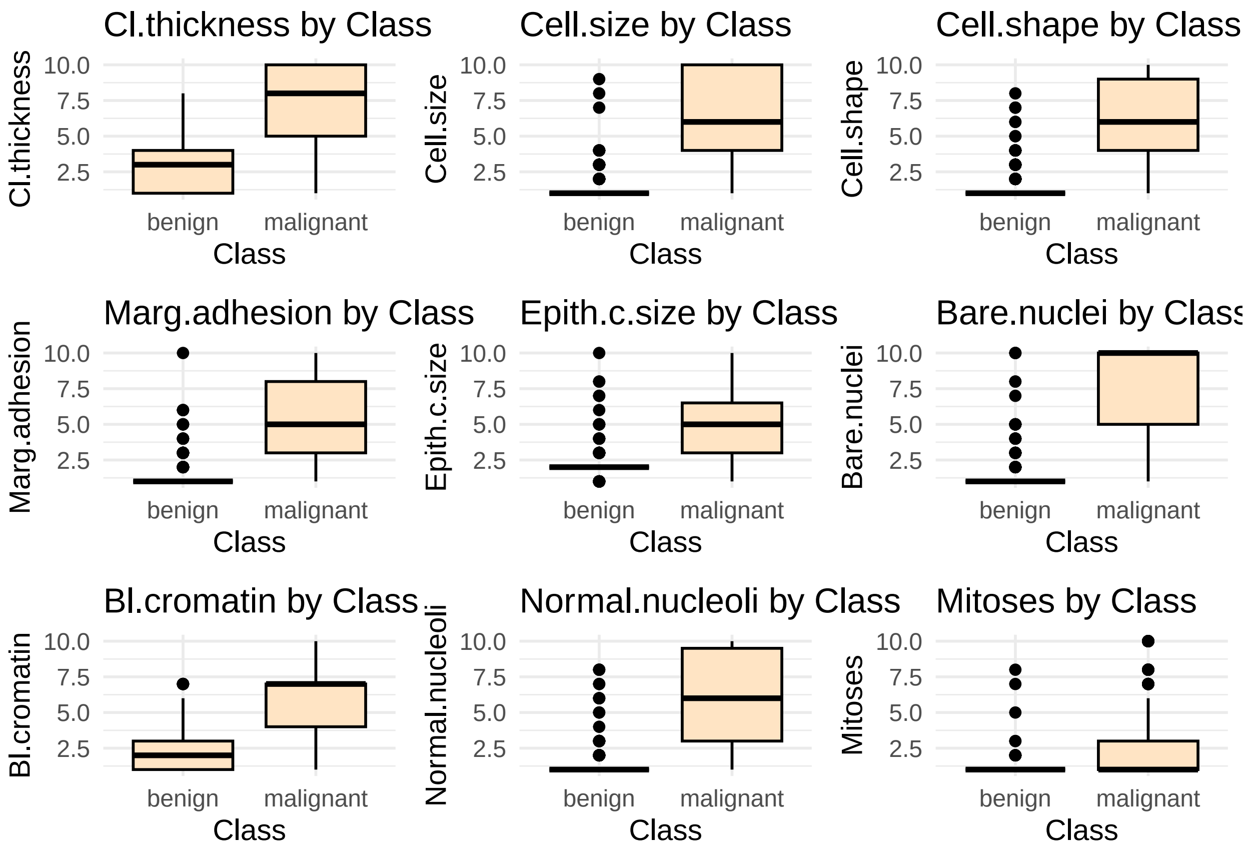
The `MyBreastCancer` dataset contains **444 samples** belonging to the **Benign** category and **239 samples** belonging to the **Malignant** category. This dataset is imbalanced, which could affect model training particularly for classifiers like logistic regression as they may become biased toward the majority class.

Predictors in benign samples (Class = 0) generally have lower means, medians, and variability, while malignant samples (Class = 1) show higher values and greater heterogeneity.

The correlation matrix below, indicates that **Cell.size**, **Cell.shape**, **Bl.cromatin**, and **Bare.nuclei** exhibit a strong correlation with each other. In contrast, **Mitoses** shows weaker correlations with most of the other predictors, suggesting it may offer distinct information not reflected by the other features.



Boxplots are plotted to compare the distribution of each predictor variable across the classes

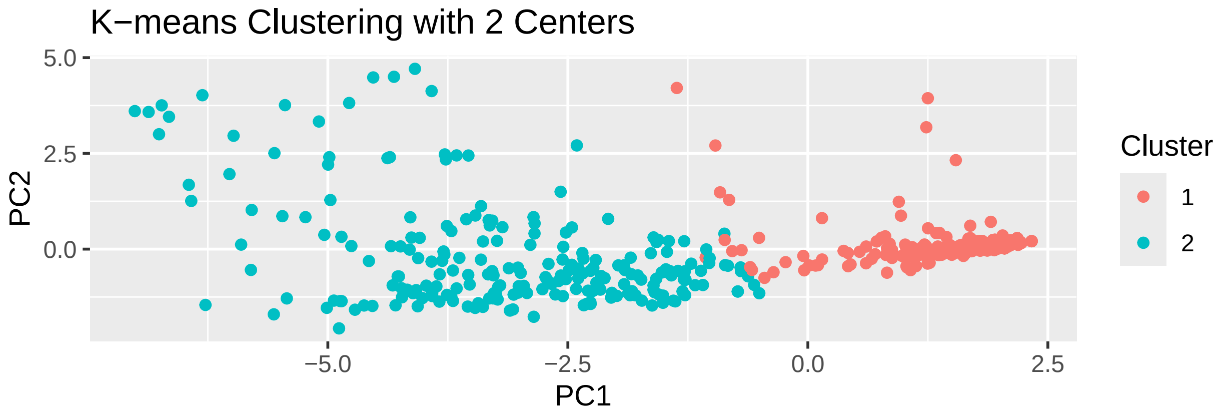


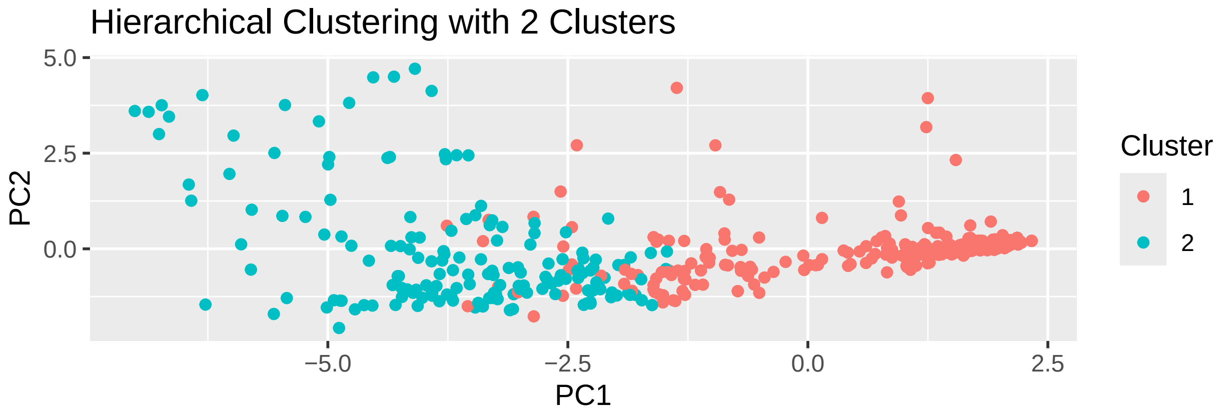
Based on the graph, **Cl.Thickness, Cell.Size, Cell.Shape, Epith.C.Size, Bare.Nuclei**, and **Normal.Nucleoli** show strong relationships with the predictors, effectively separating benign and malignant classes. These variables should be prioritized for classification.

To confirm the significance of each predictor, it would be useful to apply supervised learning methods, such as logistic regression or LDA.

Exploratory Data Analysis (EDA): Unsupervised learning

To better understand the data, unsupervised machine learning methods were applied, with k-means and hierarchical clustering chosen as the primary models. Both are effective for forming well-defined clusters, which is essential for our goal of distinguishing between benign and malignant tissue samples.





Cluster	K-Means (benign)	K-Means (malignant)	Hierarchical (benign)	Hierarchical (malignant)
1	435	18	441	75
2	9	221	3	164

By examining the confusion matrix above,

K-means Clustering: Cluster 1 mostly contains benign samples (435 benign vs. 18 malignant), while Cluster 2 predominantly consists of malignant samples (221 malignant vs. 9 benign).

Hierarchical Clustering: Cluster 1 contains more benign samples (441 benign vs. 75 malignant), but the misclassification rate is higher than in K-means. Cluster 2 includes 164 malignant samples and 3 benign samples incorrectly classified as malignant.

K-means is the better choice for separating benign and malignant tissue, as it achieved a clearer class separation, aligning with the goal of distinguishing between the two. This makes k-means a more reliable method for identifying whether unusual tissue is benign or malignant.

Results: Supervised Learning

Split the Data into Training and Testing Sets

A typical split is to use 70-80% of the data for training and 20-30% for testing.

I chose 80% for training and 20% for testing as it is suitable to experiment with multiple models and a robust test set for comparison.

Exploring the following classification models:

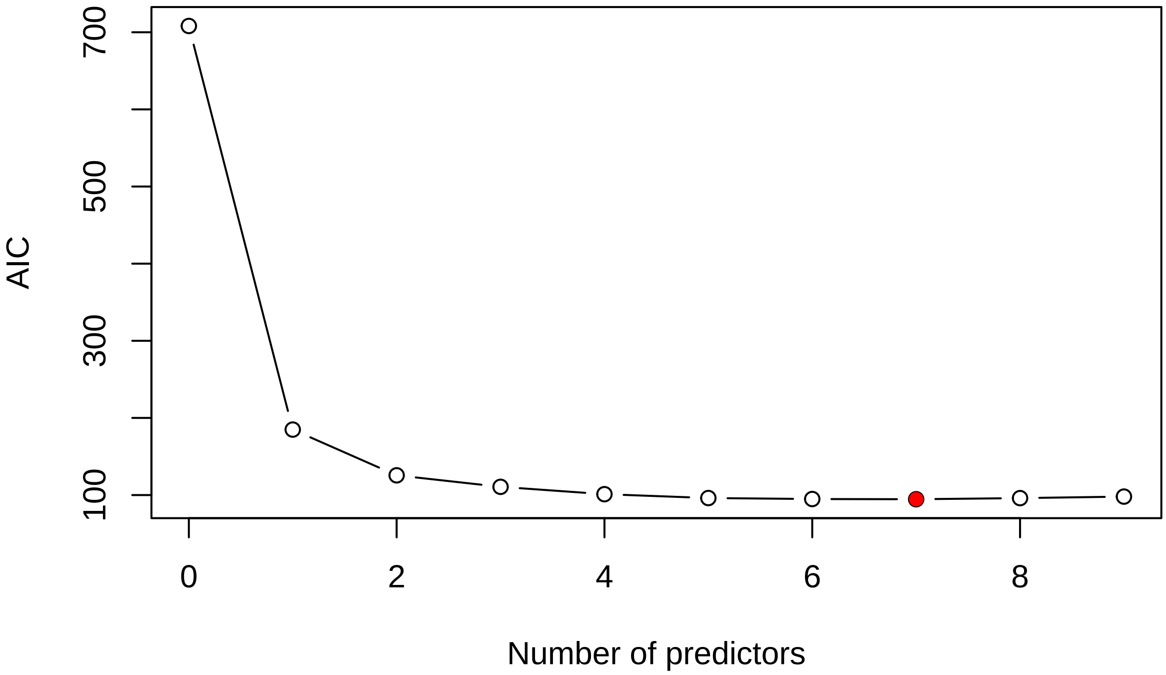
1. **Logistic Regression with Subset Selection**
2. **Regularized Logistic Regression (Lasso)**
3. **Discriminant Analysis Methods (LDA and QDA)**

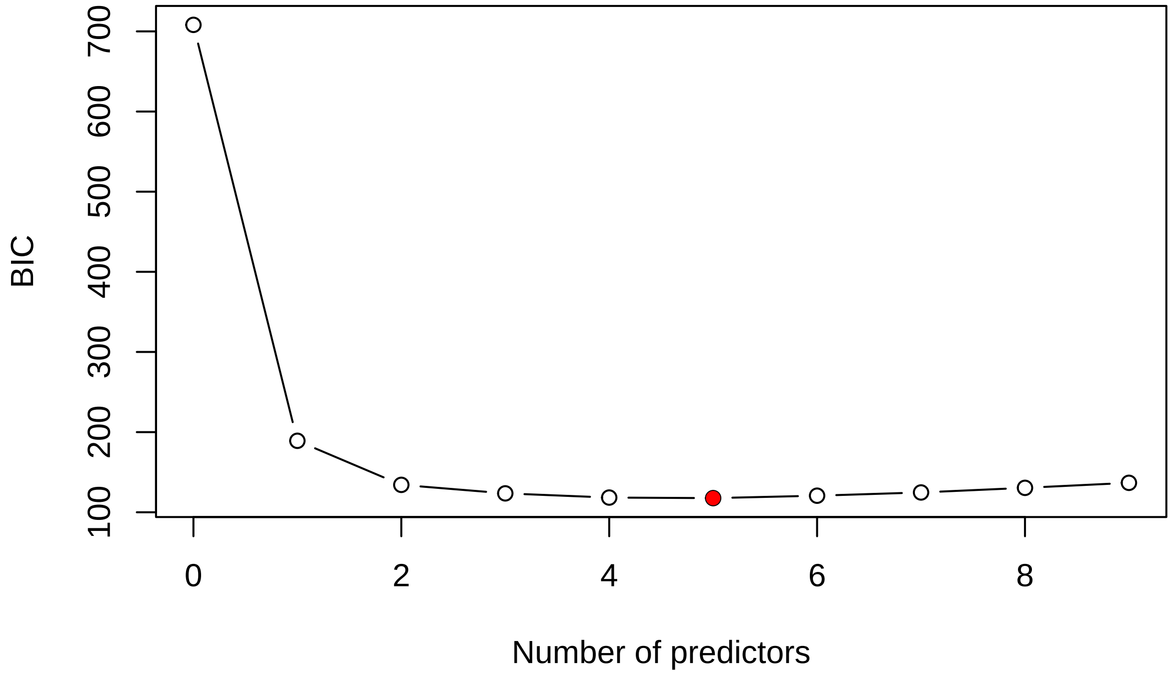
Logistic Regression with Subset Selection

Best subset selection is conducted on the training set to identify the optimal set of predictors.

Best subset selection can be applied using criteria such as AIC and BIC.

The models minimising the AIC and BIC are highlighted in below figure,





As depicted in the above figures, AIC agrees on 7 while BIC agrees on 5, so 6 subset of predictors is likely the best choice for the analysis. This subset is now used to build the final logistic regression model.

Out of the 9 predictors, **Cl.thickness, Cell.shape, Marg.adhesion, Bare.nuclei, Bl.cromatin,** and **Mitoses** predictors are selected using the Best Subset Selection for logistic regression.

Variable	**Estimate**	**Std. Error**	**z value**	**Pr(>
(Intercept)	-0.9271	0.3555	-2.608	0.00911 **
Cl.thickness	1.2161	0.4272	2.846	0.00442 **
Cell.shape	1.4861	0.5230	2.842	0.00449 **
Marg.adhesion	0.9818	0.3869	2.538	0.01115 *
Bare.nuclei	1.5436	0.3791	4.071	4.67e-05 ***
Bl.cromatin	1.5776	0.4864	3.244	0.00118 **
Mitoses	0.9482	0.6314	1.502	0.13315

Logistic regression model identifies five significant predictors for malignancy with p-values less than 0.05 with **Bare.nuclei** showing the strongest association, with a reasonable AIC value(96.867), which is used to compare this model against other potential models.

Training error

To evaluate the performance of logistic regression model with Best Subset Selection,computed the training confusion matrix and training error.

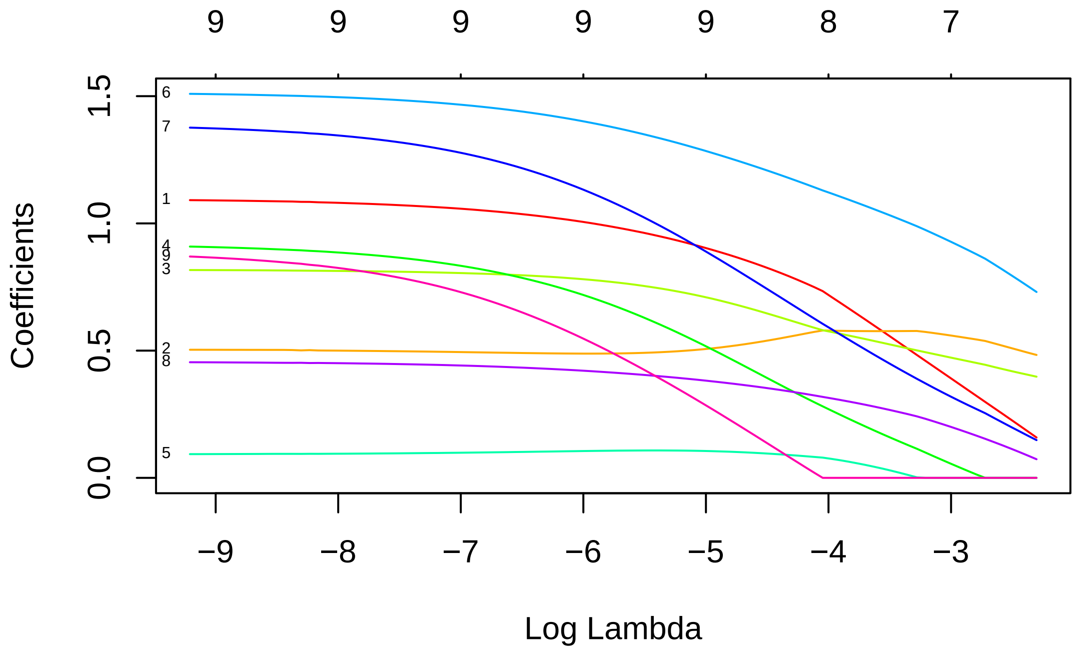
The model misclassified about **3.11%** of the training data, correctly predicting the class for **96.89%** of instances. This low error rate suggests good performance on the training data.

Testing error

Using the fitted model to predict whether the class is benign or malignant in the test dataset.

The model misclassifies approximately 65.69% of the test data, this suggests that the logistic regression model using best subset selection is not performing well on the test dataset.

2. Regularized Logistic Regression Using Lasso (Least Absolute Shrinkage and Selection Operator) penalty



A LASSO logistic regression (alpha = 1) used 10-fold cross-validation to determine the optimal lambda value of **0.007054802** for regularization to enhance model performance.

Variable	**s1**
(Intercept)	-1.0012529
Cl.thickness	0.8967974
Cell.size	0.5086950
Cell.shape	0.7046495
Marg.adhesion	0.5060159
Epith.c.size	0.1049850
Bare.nuclei	1.2778685
Bl.cromatin	0.8762577
Normal.nucleoli	0.3796286
Mitoses	0.2715968

The LASSO model identifies **Bare.nuclei** as the most important feature for predicting malignancy.

Both Best Subset Selection and LASSO suggest that **Bare.nuclei**, **Bl.cromatin**, and **Cl.thickness** are significant predictors of malignancy, with **Mitoses** being less relevant. LASSO is more aggressive in feature selection, tending to zero out less important features like **Mitoses**, whereas Best Subset Selection retains all features and evaluates their significance.

Training error

The LASSO model showed a **training error of 2.56%**, meaning the model correctly classified **97.44%** of training instances.

Testing error

The LASSO model achieved a **test error of 1.87%**, demonstrating strong generalization to unseen data and effective prediction of malignancy.

3. Discriminant Analysis (LDA and QDA)

Linear discriminant analysis

Linear Discriminant Analysis (LDA) classifies samples as benign or malignant by finding a linear combination of predictors that maximizes class separation. This approach aids in both classifying new samples and interpreting the importance of variables in distinguishing between the two classes.

The discriminant functions highlight the contribution of each predictor, such as cl.thickness and Bare.nuclei, in distinguishing between the classes. The model demonstrates high accuracy and effectively separates the two classes based on the provided features.

For consistency with the logistic regression approach, we will apply the validation set method using the same training and validation data split.

The Linear Discriminant Analysis (LDA) model achieved strong performance, correctly classifying **88 benign** and **40 malignant** samples. However, it misclassified **7 malignant** as benign and **2 benign** as malignant. The overall test error rate was **6.57%**, with **93.43%** of test samples correctly classified, indicating reliable performance.

Quadratic discriminant analysis

The Quadratic Discriminant Analysis (QDA) model classifies samples as benign (Class 0) or malignant (Class 1) based on discriminant scores, assigning each sample to the class with the higher score. The model showed effective performance with minimal misclassification.

The Quadratic Discriminant Analysis (QDA) model correctly classified **82 benign** and **46 malignant** samples, with **8 benign** misclassified as malignant and **1 malignant** misclassified as benign. The test error rate was **6.57%**, indicating **93.43%** accuracy and strong classification performance.

The results from the supervised learning models align with the EDA, confirming the importance of **Bare.nuclei**, **Cl.thickness**, and **Bl.cromatin** as key predictors of malignancy. These features were consistently selected in logistic regression and clustering, while **Mitoses**, showing weaker correlation in EDA, was less significant in the models, supporting the initial findings.

Conclusions and Discussion

This analysis evaluated several models for predicting breast cancer malignancy, including Logistic Regression with Subset Selection, Regularized Logistic Regression (Lasso), and Discriminant Analysis methods (LDA and QDA). The goal was to identify a model that provides the most accurate predictions, particularly by evaluating performance on both training and test datasets. Based on the evaluation of various models, the LASSO model emerged as the best classifier for this task.

The best model was selected based on its ability to generalize well to the test data, indicated by a low **test error rate** of approximately **1.87%**. This model was fitted using the **LASSO penalty**, with an optimal tuning parameter of **0.0087**. LASSO is effective here as it regularizes the model, shrinking less important coefficients to zero, which helps reduce complexity, mitigate overfitting, and improve generalization. The optimal lambda was chosen through cross-validation to minimize misclassification.

When considering the predictors, the LASSO model highlighted a subset of important variables, specifically **Bare.nuclei**, **Bl.cromatin**, and **Cl.thickness**, which were found to have the greatest impact on predicting malignancy. These variables were consistently selected by both Best Subset Selection and LASSO, making them key features for our model. However, Mitoses was not selected as a significant predictor by the LASSO model, suggesting that its contribution to the model is minimal. The feature selection provided by LASSO ensures that only the most relevant predictors are included, which not only improves model performance but also reduces the risk of overfitting.

The model's misclassification errors primarily involve false negatives, where malignant cases are incorrectly predicted as benign, or false positives, where benign cases are incorrectly predicted as malignant. These errors are typical in medical diagnoses, and further evaluation is needed to balance the costs (e.g., the risk of missing a malignant case versus misdiagnosing a benign one). While the LASSO model provides low test error, it's important to consider these misclassifications and their potential implications in clinical settings.

Regarding contradictions between methods, both Best Subset Selection and LASSO agreed on key predictors (**Bare.nuclei**, **Bl.cromatin**, and **Cl.thickness**), but LASSO was more aggressive in excluding irrelevant features like **Mitoses**. This contrast highlights the difference in feature selection: LASSO uses regularization to be more stringent, while Best Subset Selection evaluates all features without exclusion. Despite these differences, both methods identified the same core features, reinforcing the reliability of the results.

The main scientific goal of this analysis was to develop a model capable of accurately predicting the malignancy of breast cancer. Based on the results, the LASSO model appears to fulfill this goal effectively, with a high predictive accuracy as evidenced by the low test error rate. The model provides valuable insights into which predictors are most important for distinguishing between benign and malignant cases.

However, there are some limitations to this analysis. First, the dataset is imbalanced, with a higher number of benign cases compared to malignant ones, which could influence the performance of the classifier. While the low misclassification rate suggests good model performance, the imbalance could lead to biased predictions. Techniques like oversampling, undersampling, or using metrics like the F1 score or ROC curve to assess model performance in the context of imbalanced data. Additionally, the models used in this analysis assume that the predictors are linearly related to the outcome, which may not always hold in real-world scenarios. Exploring more advanced techniques such as ensemble methods or non-linear classifiers could potentially improve the predictive power further.

In conclusion, while the LASSO model provides strong predictive performance and successfully addresses the scientific goal, further improvements could be made by handling data imbalance and considering non-linear modeling techniques. These steps would help refine the model for real-world applications and increase its robustness.