MAS8404 | Statistical Learning for Data Science | Predicting Breast Cancer: Leveraging Cytological Features for Accurate Diagnosis

Mohmadzakir_Chotaliya_240572857

2024-11-22

1. Abstract

This study analyzed the BreastCancer dataset to classify tissue samples as benign or malignant using cytological features. Key predictors like Cl.thickness, Cell.size, and Bare.nuclei were identified. Among the supervised models, LASSO Logistic Regression achieved the highest accuracy (96.6%), followed by Subset Selection (96.1%). Both LDA and QDA showed strong linear classification performance at 95.6%. Unsupervised k-means clustering aligned closely with actual classifications. While the models performed well, limitations include a small dataset and difficulty with borderline cases. These findings confirm the clinical utility of cytological features for breast cancer diagnosis.

2. Exploratory data analysis: Data summary

2.1 Data Cleaning

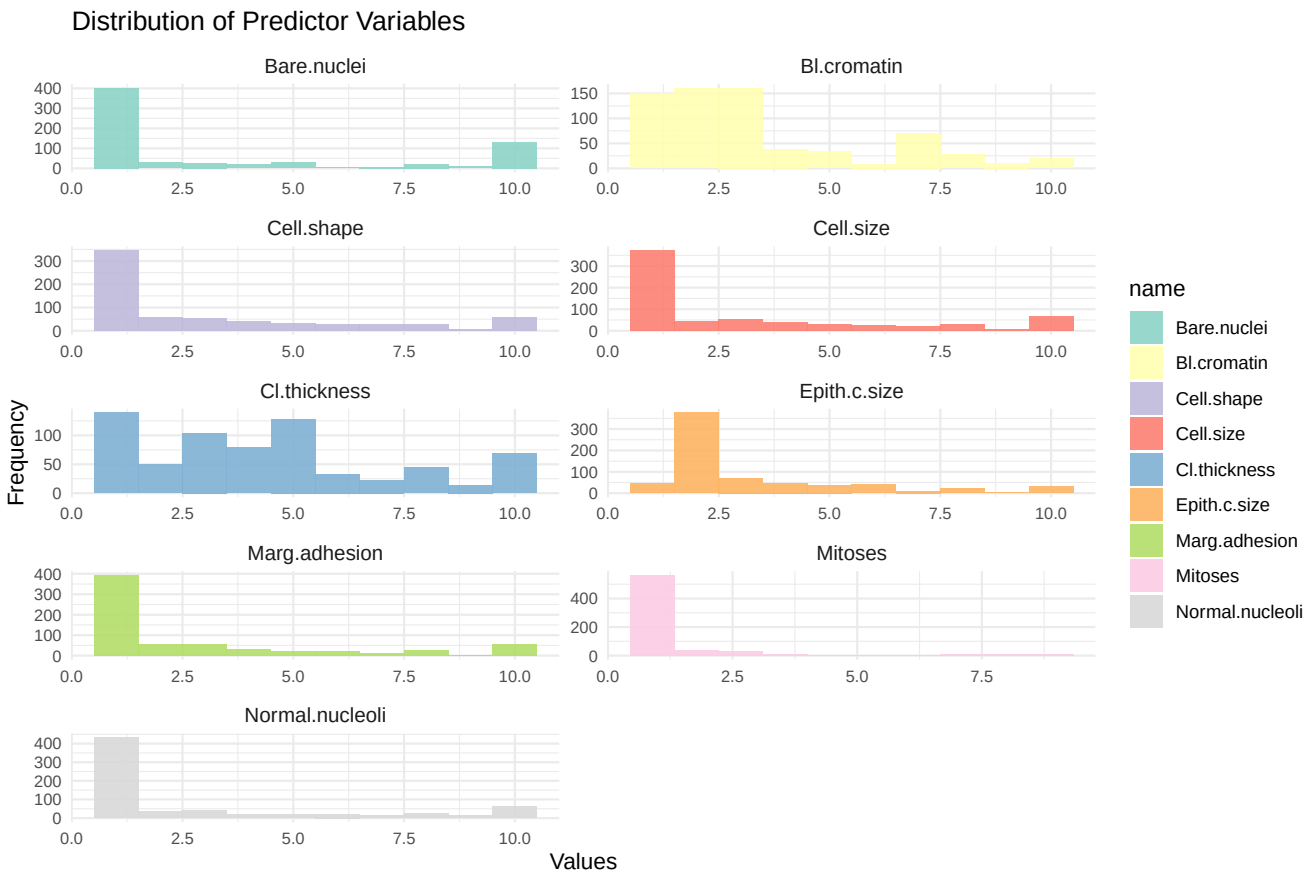
The Breast Cancer dataset was prepared for analysis by removing rows with missing values, converting categorical variables to numeric, and dropping the unnecessary Id column. This ensured the dataset is clean, consistent, and ready for further exploration.

2.2 Numerical Summary

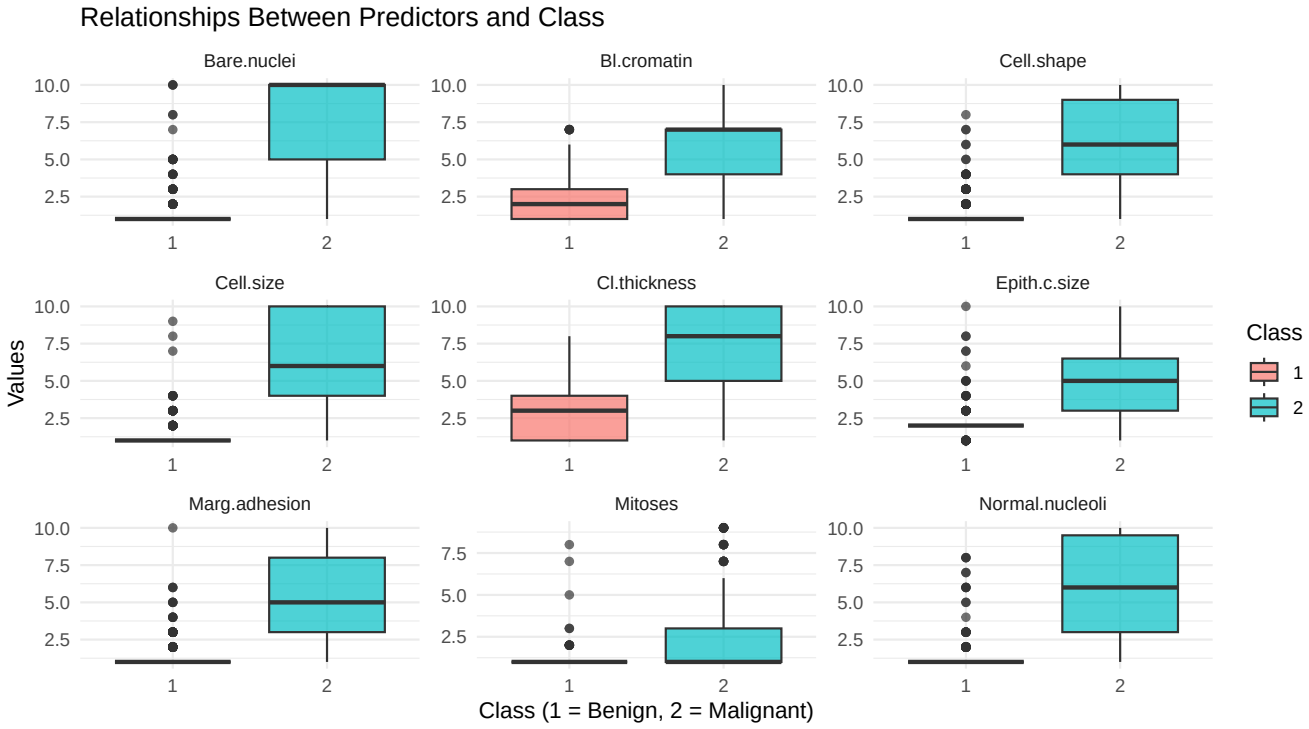
The dataset consists of 683 samples with 10 features describing cell characteristics, such as thickness, size, shape, and nuclei properties, measured on a scale of 1 to 10. Most values cluster towards the lower end, indicating healthy-looking cells for many samples. The Class variable identifies samples as benign (1) or malignant (2), with benign cases being more frequent.

2.3 Distribution of Predictor Variables

Most predictor variables, like Bare.nuclei, Mitoses, and Normal.nucleoli, have the majority of their values concentrated at the lower end (1–3), indicating minimal abnormalities in many samples. Variables like Cl.thickness and Cell.size show more variation, with values spread between 3 and 7 for about half the samples. This highlights that while most samples appear normal, a subset shows significant irregularities critical for diagnosis.



2.4 Relationships Between Predictors and Class (Response Variable)

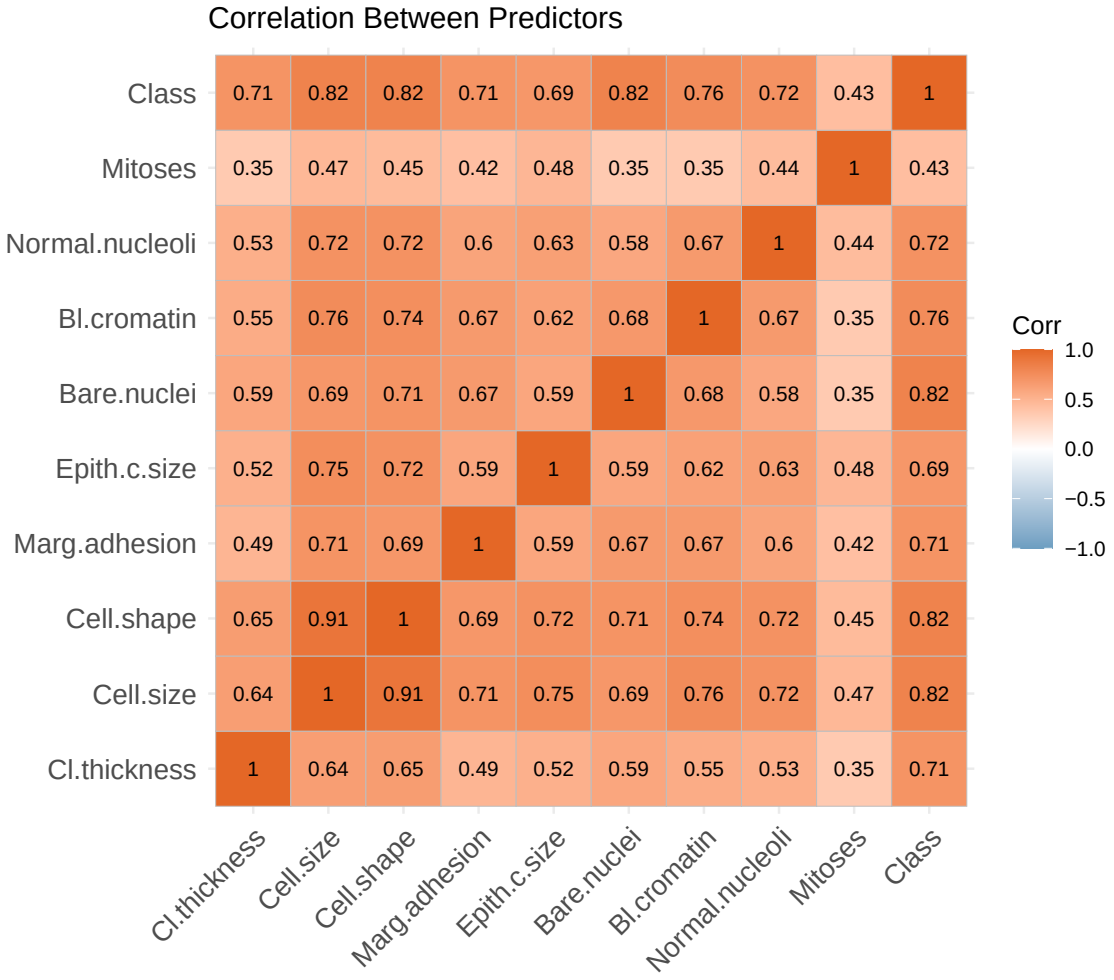


The relationships between predictor variables show that some are strongly connected. For example, Cell.size and Cell.shape are closely related, with a correlation of 0.91, meaning they often increase to-

gether. Similarly, Normal.nucleoli and Bl.cromatin are strongly linked (0.76). Variables like Epith.c.size and Bare.nuclei also show moderate connections with others, while Mitoses stands out as more independent, with weaker relationships (below 0.5). These patterns highlight overlaps between some features and the unique role of others in distinguishing between benign and malignant samples.

2.5 Correlation Between Predictors

The heatmap highlights strong relationships between some predictors, like Cell.size and Cell.shape (correlation: 0.91), showing they tend to increase together. Normal.nucleoli also closely aligns with Bl.cromatin (0.76). However, Mitoses has much weaker links with other variables, mostly below 0.5, suggesting it behaves more independently. These patterns show that some features are closely related, which could affect analysis.

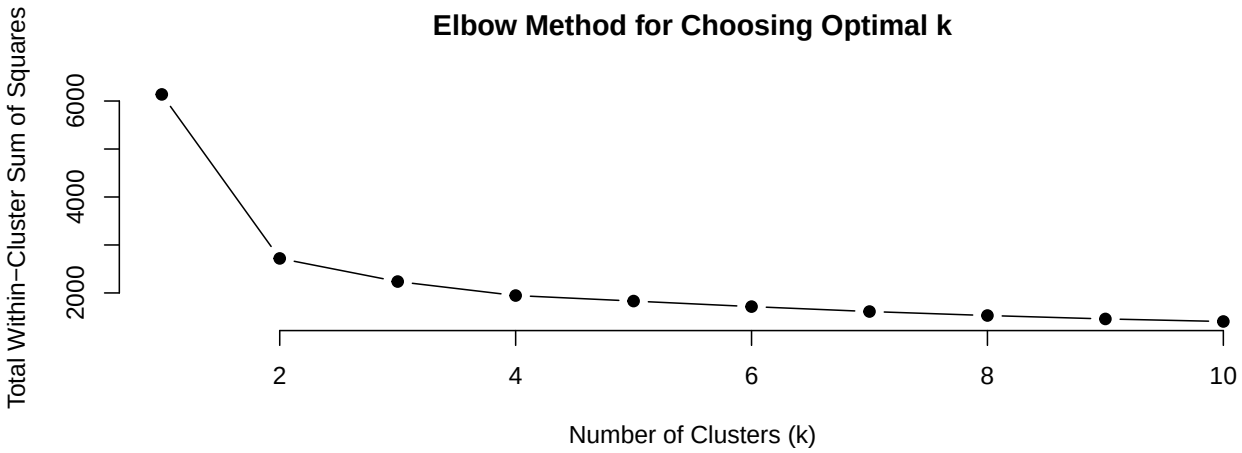


3. Exploratory data analysis: Unsupervised leaning

This section uses k-means clustering to analyze the BreastCancer dataset and understand whether unusual samples are likely to be benign or malignant.

3.1 K-Means Clustering

i) Choose Optimal Number of Clusters



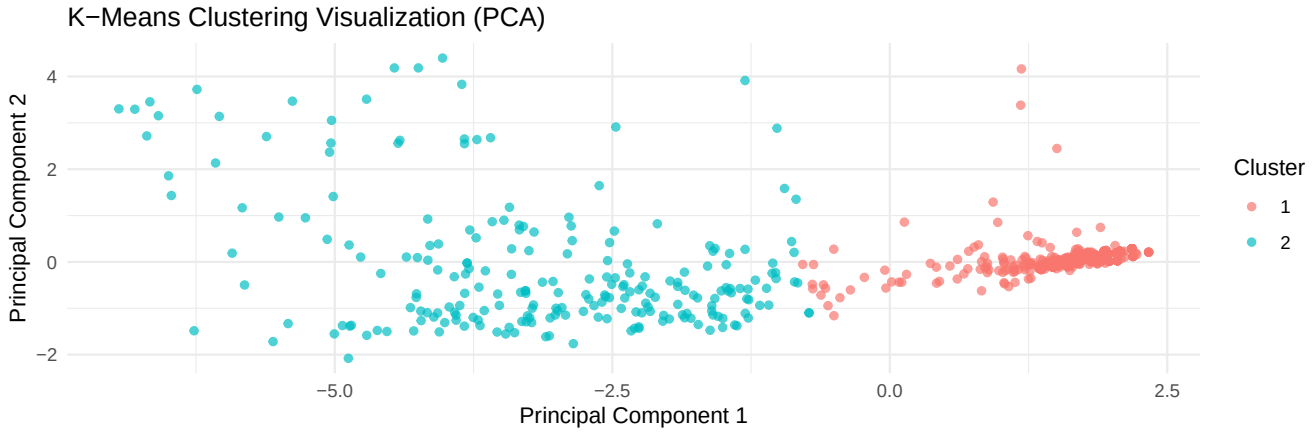
The Elbow Method plot shows the within-cluster sum of squares (WSS) for different numbers of clusters (k). The steep drop in WSS from k=1 to k=2 indicates that dividing the data into two clusters significantly improves clustering efficiency. After k=2, the reduction in WSS becomes marginal, forming an “elbow” at k=2. This supports the hypothesis that two primary clusters exist in the data, likely corresponding to benign and malignant tissue. Beyond k=2, additional clusters offer little improvement, suggesting two clusters are optimal.

ii) Perform K-Means with Optimal Clusters

```
##
##        1   2
##    1 434  18
##    2  10 221
```

The clustering closely matches the actual classifications, with most malignant samples (434) in one cluster and benign samples (221) in the other. However, 28 samples were misclassified, possibly due to overlapping features or unusual cases. Overall, the features do a good job of separating benign and malignant tissue.

iii) Visualize Clusters



The PCA visualization shows clear separation, with Cluster 1 containing 434 malignant samples and Cluster 2 aligning with 221 benign samples. Overlap exists, with 10 benign and 18 malignant samples

misclassified, and a few isolated outliers suggesting unusual cases. Overall, clustering effectively distinguishes benign from malignant samples.

Moving forward, supervised learning methods like logistic regression or LDA, along with advanced techniques like hierarchical clustering, could help refine predictions and uncover deeper insights.

4. Supervised learning

This section explores supervised learning methods, including logistic regression, LASSO, LDA, and QDA, to classify breast tissue as benign or malignant, comparing their performance to find the most effective approach.

4.1 Data Preparation

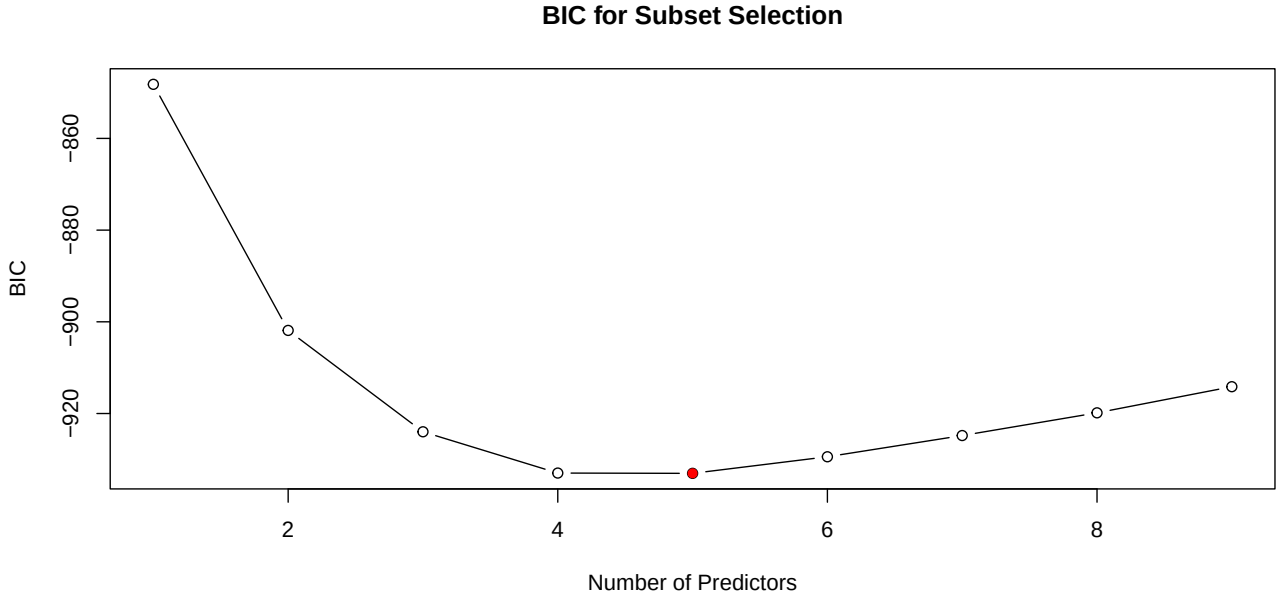
```
## Training samples: 478
## Testing samples: 205
```

The data is split into 70% training (478 samples) and 30% testing (205 samples) to train models effectively and evaluate them on unseen data, with Class as the response variable.

4.2 Logistic Regression with Subset Selection

```
## Number of predictors in the best model: 5
```

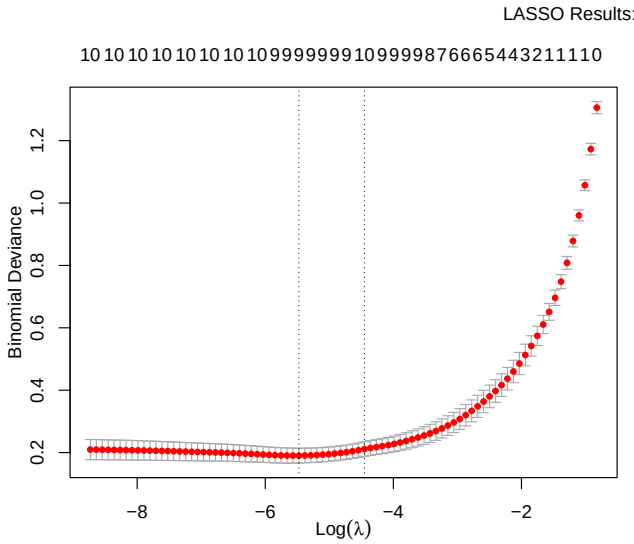
```
## Predictors in the best model:
##  Cl.thickness, Cell.size, Bare.nuclei, Bl.cromatin, Cluster
```




```
## Subset Selection Logistic Regression Accuracy: 0.9609756
```

Subset selection achieved an impressive accuracy of 96.1%, identifying Cl.thickness, Cell.size, Bare.nuclei, Bl.cromatin, and Cluster as the most important predictors. These features provide a clear and simple way to reliably distinguish between benign and malignant samples.

4.3 Regularized Logistic Regression (LASSO)



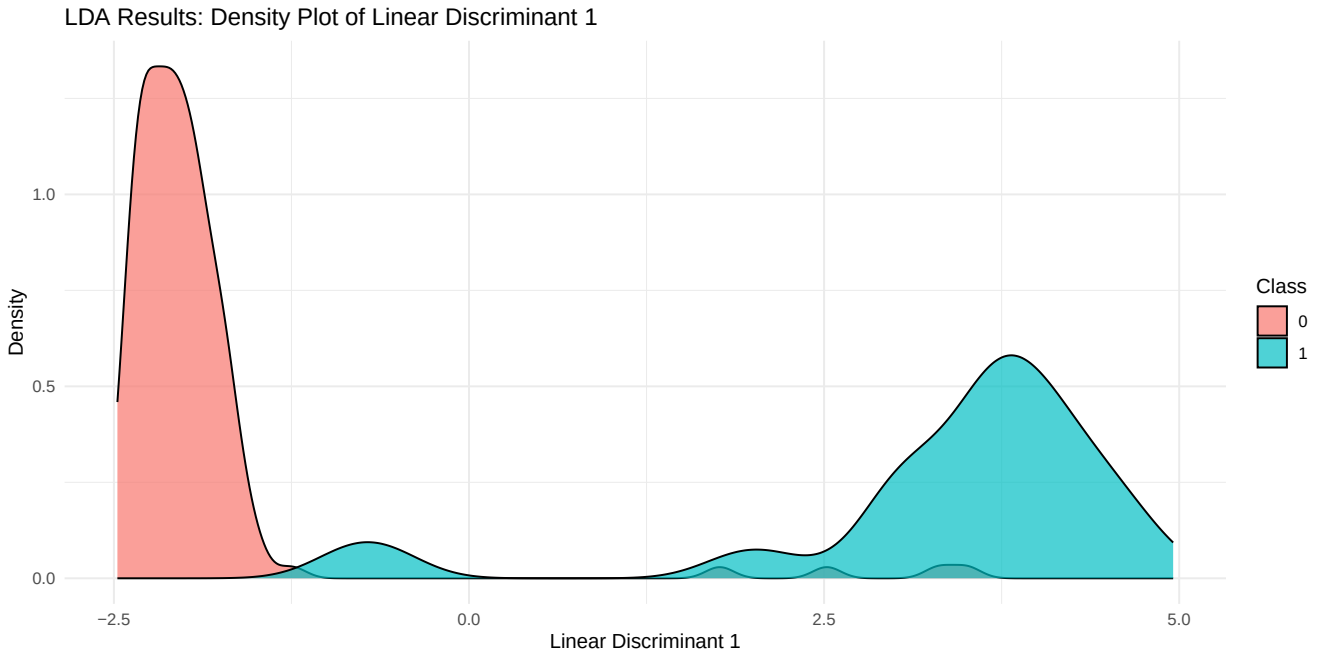
Predictor	Coefficient
(Intercept)	−7.8943273
Cl.thickness	0.3190993
Cell.size	0.2053212
Cell.shape	0.2009035
Marg.adhesion	0.2333599
Epith.c.size	0.0921439
Bare.nuclei	0.4041764
Bl.cromatin	0.3199654
Normal.nucleoli	0.1131226
Mitoses	0.2602894
Cluster	0.0000000

```
## LASSO Logistic Regression Accuracy: 0.9658537
```

LASSO Logistic Regression achieved an impressive 96.6% accuracy, highlighting important predictors of malignancy like Cl.thickness, Cell.size, Bare.nuclei, Bl.cromatin, and Mitoses, while leaving out less relevant features. Using cross-validation to fine-tune the penalty (lambda), the model stays simple yet effective, reducing multicollinearity and focusing on the most impactful predictors for accurate insights.

4.4 Linear Discriminant Analysis (LDA)

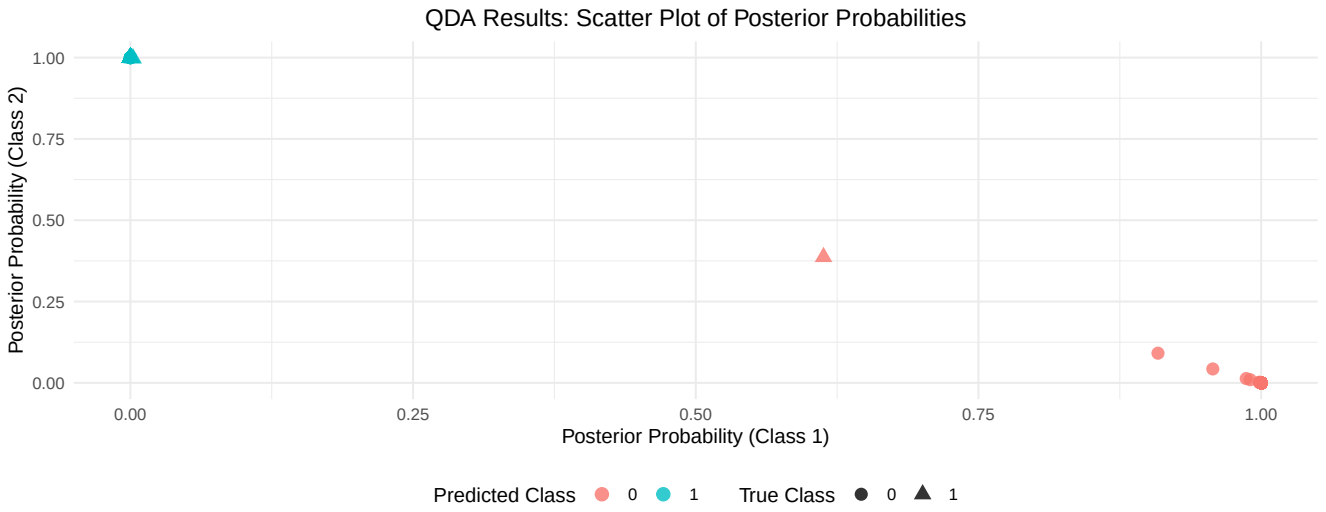
```
## LDA accuracy: 0.9560976
```



LDA delivered an impressive accuracy of 95.6% (0.9561) on the test data, showing it's very effective at distinguishing between benign and malignant cases. This suggests the dataset works well with linear classification, and further analysis could help identify which features have the biggest impact on separating the classes.

4.5 Quadratic Discriminant Analysis (QDA)

```
## QDA accuracy: 0.9560976
```



QDA also achieved 95.6% accuracy, just like LDA, suggesting that the data doesn't need more complex decision boundaries and that the class separation is likely linear. The key predictors are probably similar for both methods, and looking at misclassified cases could help uncover any subtle overlaps or patterns in the data.

4.6 Model Comparison

Table 1: Model Comparison: Accuracy Across Methods

Model	Accuracy
Subset Selection Logistic Regression	0.9609756
LASSO Logistic Regression	0.9658537
Linear Discriminant Analysis (LDA)	0.9560976
Quadratic Discriminant Analysis (QDA)	0.9560976

LASSO Logistic Regression performed the best, achieving 96.6% accuracy by focusing on the most important features like Bare.nuclei and keeping the model straightforward. Subset Selection was close behind with 96.1% accuracy, offering a simple and clear approach. Both LDA and QDA reached 95.6%, showing the data works well with linear classification, but QDA's complexity didn't add much value. Overall, LASSO is the top choice for its balance of accuracy, simplicity, and relevance to the data.

5. Conclusions and Discussion

The analysis found LASSO Logistic Regression to be the top performer, with an impressive accuracy of 96.6%. It pinpointed key predictors like Cl.thickness, Cell.size, and Bare.nuclei while leaving out less important ones like Marg.adhesion. Subset Selection Logistic Regression also performed well, with 96.1%

accuracy, offering strong results with added interpretability. Both models confirmed the importance of these predictors in distinguishing benign from malignant cases, aligning with earlier data insights. Misclassification mostly occurred in borderline cases with overlapping features, suggesting the need for more features or refined techniques. To improve accuracy and robustness, expanding the dataset or exploring advanced methods like ensemble models or neural networks could be beneficial. Overall, the analysis highlights the clinical value of these nine cytological features in accurately diagnosing breast cancer.