Breast Cancer Analysis

Davies Luo (Zheng Luo)

2023-11-24

Introduction

Breast cancer diagnosis is a critical and challenging task in medical science, where accurate classification of tumors can significantly impact patient outcomes. The complexity of cancer pathology necessitates the use of advanced statistical techniques to discern patterns and relationships within clinical data. This report delves into an analytical exploration of the BreastCancer dataset, comprising cytological characteristics from fine needle aspiration biopsies. I employ a multifaceted approach, utilizing exploratory data analysis, hierarchical clustering, principal component analysis (PCA), and logistic regression models to unravel the intrinsic data structure and develop robust classifiers. My objective is to identify the most informative cytological features that can accurately distinguish between benign and malignant breast tissue samples, ultimately contributing to the enhancement of diagnostic procedures.

Data Exploratory

Prior to the in-depth analysis of the BreastCancer dataset, I performed essential data cleaning, which involved converting several key cytological characteristics to numerical values and removing records with missing data to enhance data integrity. Following this preparation, I employed a combination of graphical and numerical approaches to unravel and understand the complex relationships inherent in the dataset. This exploratory analysis is pivotal in providing a comprehensive overview of the dataset's structure, highlighting key patterns, distributions, and correlations amongst the variables. By integrating both visual and quantitative methods, I aimed to gain a deeper insight into the dataset, setting a solid foundation for more advanced statistical modeling and interpretation in the later stages of this research.

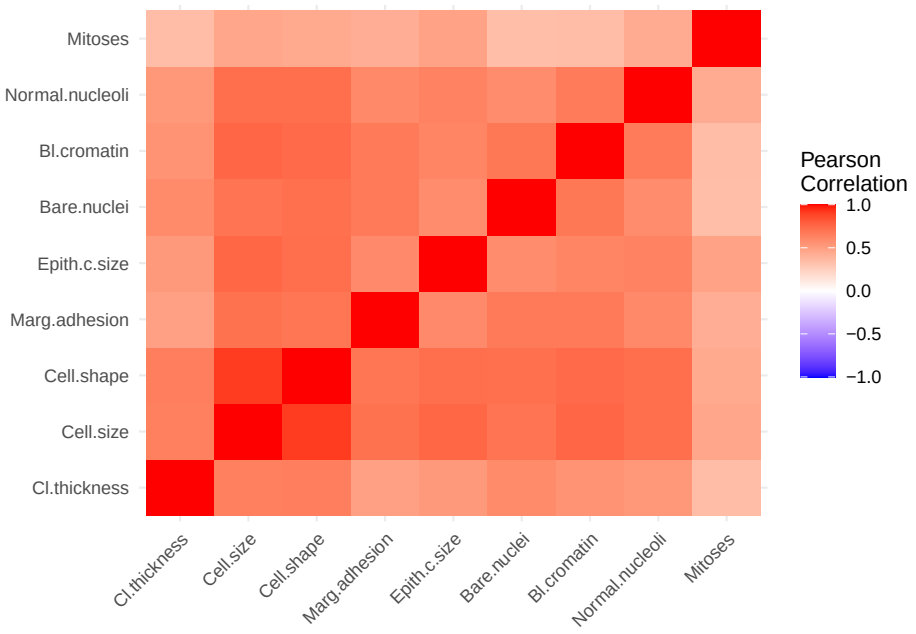

Figure 1: The graphical summary of the data

Table 1: The numerical summary of the data

Cl.thickness	Min. : 1.000	1st Qu.: 2.000	Median : 4.000	Mean : 4.442	3rd Qu.: 6.000	Max. :10.000
Cell.size	Min. : 1.000	1st Qu.: 1.000	Median : 1.000	Mean : 3.151	3rd Qu.: 5.000	Max. :10.000
Cell.shape	Min. : 1.000	1st Qu.: 1.000	Median : 1.000	Mean : 3.215	3rd Qu.: 5.000	Max. :10.000
Marg.adhesion	Min. : 1.00	1st Qu.: 1.00	Median : 1.00	Mean : 2.83	3rd Qu.: 4.00	Max. :10.00
Epith.c.size	Min. : 1.000	1st Qu.: 2.000	Median : 2.000	Mean : 3.234	3rd Qu.: 4.000	Max. :10.000
Bare.nuclei	Min. : 1.000	1st Qu.: 1.000	Median : 1.000	Mean : 3.545	3rd Qu.: 6.000	Max. :10.000
Bl.cromatin	Min. : 1.000	1st Qu.: 2.000	Median : 3.000	Mean : 3.445	3rd Qu.: 5.000	Max. :10.000
Normal.nucleoli	Min. : 1.00	1st Qu.: 1.00	Median : 1.00	Mean : 2.87	3rd Qu.: 4.00	Max. :10.00
Mitoses	Min. : 1.000	1st Qu.: 1.000	Median : 1.000	Mean : 1.603	3rd Qu.: 1.000	Max. :10.000
Class	benign :444	malignant:239	NA	NA	NA	NA

The heatmap stands out as an effective graphical tool, immediately visualizing the correlations between cytological characteristics. Shades ranging from light to dark red indicate the strength of positive correlations, with the most intense reds denoting the strongest relationships. This is particularly evident with variables such as Cl.thickness, Cell.size, and Cell.shape, which consistently exhibit higher correlations with each other, suggesting a potential combined role in the progression of malignancy.

The numerical summaries complement these findings by quantitatively describing the distribution of each characteristic. Notably, variables such as Bl.cromatin and Normal.nucleoli show a pronounced skew towards

higher values, suggesting that a subset of the dataset contains samples with notably severe cytological abnormalities. The distributional data further aid in discerning the nuances in the dataset, highlighting right-skewed patterns that are indicative of variations in the progression or severity of disease states.

Relationships Between Variables

Looking at the heatmap in conjunction with the response variable, Class, I notice distinct patterns where certain characteristics like Cl.thickness, Cell.size, and Cell.shape have a pronounced correlation with the cancer classification. This is suggested by the deeper red hues along the corresponding rows and columns in the heatmap, indicating their potential as significant predictors in distinguishing malignancy. In contrast, variables such as Mitoses exhibit lighter shades, suggesting a weaker direct relationship with the response variable.

The numerical summary sheds light on the range and median values of these characteristics, further differentiating between benign and malignant samples. For instance, the higher mean and maximum values for Cell.size and Cell.shape within malignant samples reinforce the idea that these variables are significant indicators of malignancy. The contrast in values between the two classes highlights the potential of these variables to contribute to an accurate cancer classification, supporting the efficacy of a multivariate approach in predictive modeling.

Predictor Variables Relationships

The heatmap also offers insights into the relationships among predictor variables, revealing red blocks where strong positive correlations exist. The deep red square between Cell.size and Cell.shape, for example, suggests redundancy between these variables, which could have implications for models that assume predictor independence. In contrast, the numerical summary provides a different perspective, revealing the spread and distribution of values that underpin these correlations. This statistical context is invaluable for interpreting the heatmap and understanding the multifaceted relationships within the data.

By integrating the graphical patterns observed in the heatmap with the statistical context provided by the numerical summary, a more comprehensive picture of the data emerges. This combined analysis highlights the intricate interplay between predictor variables, allowing for a richer understanding of the data structure. It informs subsequent modeling decisions and preprocessing steps, ensuring that the predictors used in any subsequent models reflect the true nature of the underlying relationships within the dataset.

Hierarchical Clustering

In this section of the analysis, I delve into Hierarchical Clustering, a crucial technique for uncovering the inherent structure within our breast cancer dataset. Hierarchical Clustering is particularly adept at revealing natural groupings and relationships among cytological characteristics, which are pivotal in understanding the complex nature of breast cancer. By employing this method, I aim to explore how different variables cluster together and whether these clusters can provide insights into distinguishing between benign and malignant cases. This approach is not only fundamental in identifying patterns in high-dimensional data but also serves as a basis for more informed and nuanced data-driven decisions. Through the application of different linkage methods, such as single, complete, and average linkage, I will assess the robustness and suitability of each in capturing the underlying biological processes, thereby providing a comprehensive view of the relationships within the data.

Single-linkage Clustering

Figure 2: Single Linkage Dendrogram

The dendrogram produced from the single-linkage hierarchical clustering, which relies on correlation-based distances, fails to delineate two distinct groups of variables that could correspond to benign and malignant classes.

Statistically, single-linkage clustering is sensitive to outliers, it forms clusters by successively linking individual observations with the smallest distance. This can lead to a chaining effect, where variables are clustered together based on the strength of the nearest neighbor without considering the overall data structure. Such chaining can artificially elongate clusters, pulling in loosely related variables and obscuring meaningful groupings. The method's inability to capture the broader inter-variable correlations means that it cannot reliably differentiate clusters that may represent the critical underlying biological processes distinguishing tumor classes. Consequently, the single-linkage approach does not provide a clear or robust separation in this biomedical context, suggesting the need for alternative clustering strategies that consider the global data structure for a more meaningful biological interpretation.

Complete-linkage and Average-linkage Clustering

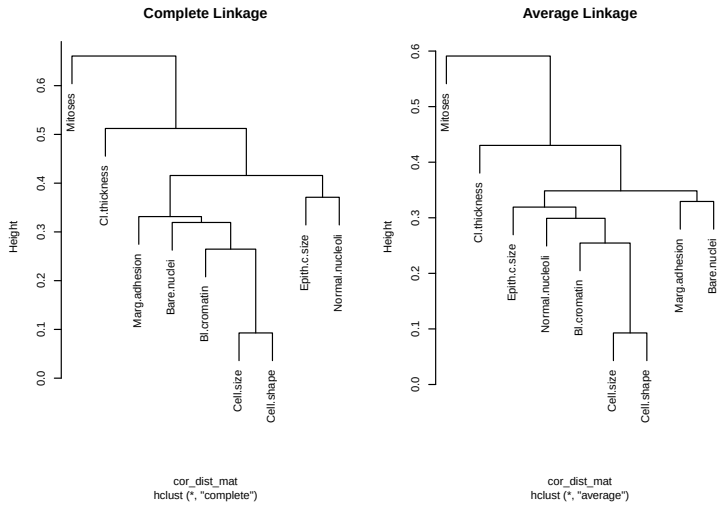

Figure 3: Complete and Average Linkage Dendrograms

Based on the hierarchical clustering results, it becomes evident that the type of linkage employed significantly impacts the structure of the resulting dendrograms. The complete-linkage dendrogram, which clusters variables based on their maximum pairwise distance, suggests a pronounced distinction between clusters. This method's tendency to form compact clusters is clearly observed, with variables joining at various heights, indicating different levels of similarity and a conservative clustering approach. Statistically, this method is robust against outliers, as it does not allow a single close relationship to unduly influence the clustering process. As a result, the dendrogram exhibits a balanced branching pattern, with no individual variable disproportionately distanced from the rest, reflecting a more uniform intra-cluster similarity.

In contrast, the average-linkage dendrogram presents a more nuanced picture. Clusters form by calculating the average distance between all observation pairs, offering a middle ground between the sensitivity of single-linkage to outliers and the potential for complete-linkage to overemphasize dissimilarities. The dendrogram produced by this method reveals a gradual separation of clusters, indicating a spectrum of similarity levels rather than a strict binary division. This method provides a comprehensive view of the data's structure, with the dendrogram's graduated branch heights suggesting a more refined clustering that incorporates a broader range of inter-variable relationships.

The observations from these dendrograms lead to the conclusion that the clustering results are indeed dependent on the chosen linkage method. Complete-linkage clustering is characterized by its creation of well-defined, tightly-knit clusters, which could be preferable in studies where the strongest correlations are of interest. Average-linkage clustering, however, with its more graded approach, may be more suitable for datasets where the relationships are complex and a single outlier does not fully represent the data's structure. These differences highlight the importance of selecting a linkage method that aligns with the objectives of the analysis and the nature of the dataset at hand. The choice of linkage method is not merely a procedural detail but a fundamental decision that can alter the interpretation and conclusions drawn from the data.

PCA Analysis

In the pursuit of understanding which variables most significantly differentiate between the two groups within the dataset, Principal Component Analysis (PCA) is employed:

```
##                    PC1    PC2    PC3    PC4    PC5    PC6    PC7    PC8    PC9
## Cl.thickness    -0.302 -0.141  0.866  0.108 -0.080 -0.243 -0.009 -0.248  0.003
## Cell.size       -0.381 -0.047 -0.020 -0.204  0.146 -0.139 -0.205  0.436  0.733
## Cell.shape      -0.378 -0.082  0.034 -0.176  0.108 -0.075 -0.127  0.583 -0.667
## Marg.adhesion   -0.333 -0.052 -0.413  0.493  0.020 -0.655  0.124 -0.163 -0.046
## Epith.c.size    -0.336  0.164 -0.088 -0.427  0.637  0.069  0.211 -0.459 -0.067
## Bare.nuclei     -0.335 -0.261  0.001  0.499  0.125  0.609  0.403  0.127  0.077
## Bl.cromatin     -0.346 -0.228 -0.213  0.013 -0.228  0.299 -0.700 -0.384 -0.062
## Normal.nucleoli -0.336  0.034 -0.134 -0.417 -0.690  0.022  0.460 -0.074  0.022
## Mitoses         -0.230  0.906  0.080  0.259 -0.105  0.148 -0.132  0.054 -0.007
```

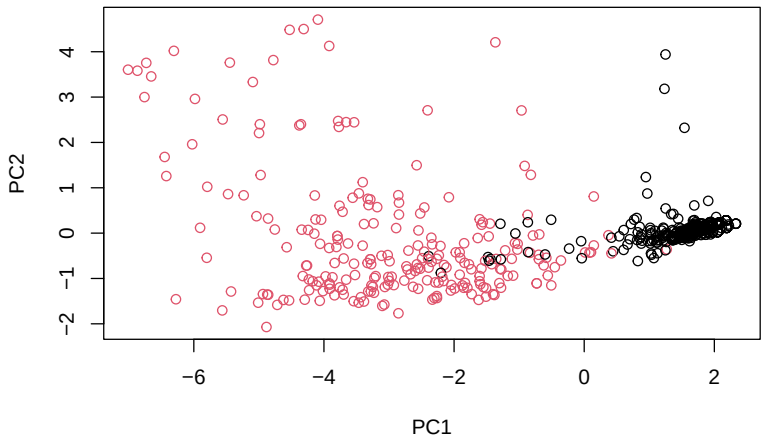

Figure 4: PCA Loadings and Scatter Plot

```
PC1 = (-0.302 * Cl.thickness) + (-0.381 * Cell.size) + (-0.378 * Cell.shape) + (-0.333 *
Marg.adhesion) + (-0.336 * Epith.c.size) + (-0.335 * Bare.nuclei) + (-0.346 * Bl.cromatin)
+ (-0.336 * Normal.nucleoli) + (-0.230 * Mitoses)

PC2 =  (-0.141 * Cl.thickness) + (-0.047 * Cell.size) + (-0.082 * Cell.shape) + (-0.052 *
Marg.adhesion) + (0.164 * Epith.c.size) + (-0.261 * Bare.nuclei) + (-0.228 * Bl.cromatin)
+ (0.034 * Normal.nucleoli) + (0.906 * Mitoses})
```

PCA reduces the dimensionality of the data by transforming the original variables into a new set of uncorrelated features, known as principal components (PCs). These components are constructed in such a way that the first few retain the majority of the variation present in all of the original variables. By examining the loadings of the original variables on these components and their graphical representation, I can identify which features contribute most to the variance and potentially distinguish between benign and malignant classes.

The first principal component (PC1) is more influenced by Cell.size, Cell.shape, and Bl.cromatin, which indicates their importance in explaining the variance between samples. This is further validated by the clear visual separation between benign and malignant classes along PC1 in the scatter plot. Suggesting that PC1 represents a measure of overall cell abnormality and chromatin pattern. Negative coefficients indicate that higher values on these variables (which might indicate more abnormal cells) contribute to a higher value on PC1. Thus, PC1 can be seen as representing general cell abnormality and chromatin patterns, capturing the majority of variance in these features.

The second principal component (PC2) prominently features Mitoses. This variable's distinct loading on PC2 suggests its role in distinguishing between the classes, which is also evident in the scatter plot. Potentially representing the rate of cell division. Variables like Bare.nuclei and Bl.cromatin have negative loadings, indicating that higher values on these decrease the PC2 score, offering a contrast to Mitoses.Therefore, PC2 appears to differentiate samples based on the rate of cell division, contrasting it with nuclear and chromatin features.

Both these components underscore the multifaceted nature of the data, paving the way for dimensionality reduction that retains critical diagnostic information.

K-means Clustering

The K-means clustering algorithm is a pivotal method in unsupervised machine learning, commonly used for identifying patterns and groupings within datasets without pre-labeled outcomes. My goal in implementing this algorithm from scratch is to not only understand the inner workings of the clustering process but also to tailor the algorithm to suit specific needs that may arise in data analysis. By building a custom version, I can explore various modifications and optimizations, and potentially improve upon the algorithm's performance for the particular characteristics of the breast cancer dataset I am examining.

```
my_kmeans <- function(X, K, max.iter = 100) {
  # Randomly assign initial cluster_center
  cluster_center <- X[sample(nrow(X), K), ]
  clusters <- rep(0, nrow(X))
  for (i in 1:max.iter) {
    # Assign clusters based on closest cluster_center
    clusters <- apply(X, 1, function(x) {
      which.min(colSums(t(cluster_center - x)^2))})

    # Recalculate cluster_center
    new_cluster_center <- matrix(ncol = ncol(X), nrow = K)
    for (k in 1:K) {
      new_cluster_center[k, ] <- colMeans(X[clusters == k, , drop=FALSE], na.rm = TRUE)}

    # Check for convergence
    if (all(cluster_center == new_cluster_center)) {
      message("Convergence reached after ", i, " iterations.")
      break}

    cluster_center <- new_cluster_center}

  # Calculate within-cluster sum-of-squares
  SSW <- sum(sapply(1:K, function(k) {
    sum(rowSums((X[clusters == k, , drop=FALSE] - cluster_center[k, ])^2))}))

  return(list(cluster = clusters, centers = cluster_center, ssw = SSW))}
```

The custom my_kmeans function created here is designed to partition the dataset into K distinct clusters based on the similarity of data points. It begins by randomly selecting K points from the dataset to serve as initial cluster-centers. The algorithm then iterates over two main steps: assignment and update. In the assignment step, each data point is assigned to the nearest centroid, forming K clusters. The update step recalculates the cluster-centers as the mean of all points in each cluster. This process repeats until the cluster-centers no longer change significantly, indicating convergence. The function outputs the final cluster assignments, the locations of the cluster-centers, and the within-cluster sum-of-squares (SSW), a measure of clustering quality. The SSW is a critical outcome, as it quantifies the compactness of the clusters; the lower the SSW, the tighter and more coherent the clusters are, which is often desirable in clustering scenarios.

Applying Custom K-means and Validation with R's K-means

My objective is to apply the custom K-means function to the breast cancer dataset to identify natural groupings that could correspond to benign and malignant tumor classifications. I aim to run this clustering multiple times to minimize within-cluster variation and compare the results with R's built-in kmeans function to validate the robustness and accuracy of my custom implementation.

Initially, I preprocessed the breast cancer dataset to ensure numeric values and removed any incomplete cases to maintain the quality of the analysis. A consistent random seed was set before multiple runs to ensure the reproducibility of the results.

In the clustering phase, I invoked the custom my_kmeans function multiple times, each time with a different seed to avoid local minima and to find the clustering that results in the smallest within-cluster sum of squares (SSW). The SSW is crucial as it quantifies how compact the clusters are, with a lower value indicating that data points within a cluster are close to each other, suggesting a good quality of clustering.

Upon finding the best clustering run after 6 attemps, I observed that the SSW value was 30166.39, which served as a benchmark for clustering quality. However, when comparing the partitioning from my custom function with that from R's kmeans, which also minimizes the SSW, the comparison returned FALSE. This discrepancy implies that while both methods aim to optimize the same criterion, they may converge to different solutions, likely due to differences in initial cluster-centers or the inherent randomness in the K-means algorithm.

Then, I created a contingency table to compare the cluster assignments and see if there is a consistent one-to-one mapping between the clusters of both methods:

```
##
##        1   2
##   1 452   5
##   2   1 225
```

Based on the contingency table output from comparing the cluster assignments of my custom K-means function and R's built-in kmeans function, I observed a strong agreement between the two clustering results. The table indicates that 452 samples in cluster 1 from my function correspond to cluster 1 of the kmeans function, and 225 samples in cluster 2 from my function match cluster 2 of the kmeans function, with only a small number of samples (6 in total) being assigned to different clusters.

This high degree of correspondence suggests that despite the inherent randomness in the K-means algorithm and possible differences in the implementation details, both methods have converged to a similar partitioning of the data. The minor discrepancies can be attributed to the stochastic nature of the algorithm, where different initializations can lead to slightly varied outcomes.

In conclusion, my implementation of the K-means algorithm is consistent with the built-in R function in identifying clusters within the breast cancer dataset, as evidenced by the large majority of data points being grouped into corresponding clusters by both methods. This gives me confidence in the robustness of my clustering approach and its suitability for analyzing this dataset.

Exploring Cluster Partitions with R's Kmeans

If there is a given dataset and choice of K, even with the same stopping conditions for R's kmeans function and my own K-means implementation, Due to the random initialization of cluster_center in the K-means algorithm, which can lead to different solutions. Especially in datasets that have clusters that are not well-separated or have multiple potential cluster_center that could be considered as "optimal" depending on the starting points.

Another factor is the potential presence of local minima in the dataset. K-means is susceptible to converging to local minima, which means that different runs might settle into different local minima, resulting in different cluster partitions. This is particularly true if the dataset does not have clear, distinct clusters, or if there is a lot of overlap between data points belonging to different clusters.

Moreover, the order of operations and computational precision can also lead to differences. Although two algorithms may have the same theoretical approach, practical implementations may differ slightly due to computational aspects, which can lead to diverging results.

Therefore, unless the initialization of cluster_center is controlled and is the same across both procedures, complete equivalence of the cluster partitions is not guaranteed.

Classification

In the classification phase of this study, my primary objective is to develop predictive models that can accurately distinguish between benign and malignant breast cancer cases based on cytological characteristics. To achieve this, I have partitioned the BreastCancer dataset into training and testing sets, utilizing 80% for training to ensure comprehensive model learning and 20% for testing to validate the models' effectiveness:

Table 2: Summary Statistics for Training and Test Data

	Mean	Median	Min	Max	Variance	SD
Training.Id	1086817.980	1173511.5	76389	13454352	458360949439.561	677023.596
Training.Cl.thickness	4.394	4.0	1	10	7.758	2.785
Training.Cell.size	3.201	1.0	1	10	9.644	3.105
Training.Cell.shape	3.267	1.5	1	10	9.220	3.036
Training.Marg.adhesion	2.897	1.0	1	10	8.456	2.908
Training.Epith.c.size	3.231	2.0	1	10	4.956	2.226
Training.Bare.nuclei	3.599	1.0	1	10	13.507	3.675
Training.Bl.cromatin	3.469	3.0	1	10	6.224	2.495
Training.Normal.nucleoli	2.868	1.0	1	10	9.293	3.048
Test.Id	1036476.628	1145420.0	63375	1368882	92804948617.015	304639.046
Test.Cl.thickness	4.635	4.0	1	10	8.763	2.960
Test.Cell.size	2.949	1.0	1	10	8.416	2.901
Test.Cell.shape	3.007	1.0	1	10	7.787	2.790
Test.Marg.adhesion	2.562	1.0	1	10	7.174	2.679
Test.Epith.c.size	3.248	2.0	1	10	4.923	2.219
Test.Bare.nuclei	3.328	1.0	1	10	12.399	3.521
Test.Bl.cromatin	3.350	3.0	1	9	5.141	2.267
Test.Normal.nucleoli	2.876	1.0	1	10	9.492	3.081

The summary statistics for both the training and test datasets reveal striking similarities in their distributions of key variables. Across various variables, including cell thickness, size, shape, and others, the mean values are remarkably close between the two datasets, differing only by small decimal places. Similarly, the median values, which are robust to outliers, exhibit minimal differences. This consistency suggests that the random split of approximately 80% for training and 20% for testing has resulted in representative subsets that capture the underlying data distribution effectively. The tight alignment of minimum and maximum values further supports this notion, indicating that extreme data points are well-distributed in both sets. Overall, these findings provide confidence in the quality of the data split, suggesting that the training and test datasets accurately reflect the broader population of patients with breast cancer. This balanced representation is crucial for building a robust and generalizable predictive model.

Build Subset Selection and LASSO Model

In the pursuit of advancing breast cancer diagnostics, statistical modeling stands as a pivotal element in the interpretation and prediction of clinical outcomes. The complexity of cancer pathology necessitates the employment of sophisticated modeling techniques that can decipher intricate patterns and relationships within the data. This section of the analysis is dedicated to the development of two distinct predictive models: the Subset Selection Logistic Regression model and the LASSO Logistic Regression model.Here are the key summary of both subset selection in logistic regression and regularized logistic regression (LASSO):

Table 3: Coefficients from Subset Selection Logistic Regression

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-9.914	1.257	-7.885	0.000
Cl.thickness	0.415	0.155	2.687	0.007
Cell.shape	0.388	0.183	2.116	0.034
Marg.adhesion	0.326	0.134	2.427	0.015
Bare.nuclei	0.419	0.103	4.075	0.000
Bl.cromatin	0.576	0.200	2.876	0.004
Normal.nucleoli	0.178	0.122	1.463	0.144
Mitoses	0.525	0.348	1.509	0.131

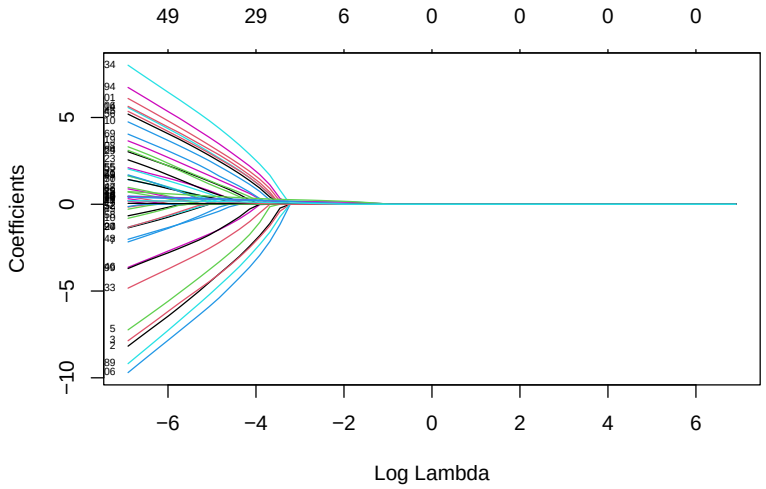

Figure 5: LASSO Coefficient Paths

The Subset Selection Logistic Regression model, a traditional approach, includes multiple predictor variables such as Cl.thickness, Cell.shape, Marg.adhesion, Bare.nuclei, Bl.cromatin, Normal.nucleoli, and Mitoses to gauge the likelihood of breast tissue being malignant. The positive coefficients for Cl.thickness, Cell.shape, Bare.nuclei, Bl.cromatin, and Marg.adhesion suggest an increased risk of malignancy with their heightened values, indicating their critical role in the model's predictive power.

In contrast, the LASSO path plot for the Regularized Logistic Regression model demonstrates the effect of the L1 penalty in variable selection and coefficient shrinkage. The plot reveals how coefficients for predictors approach and reach zero as the strength of regularization increases, signifying the diminishing contribution of some variables to the model as penalization intensifies. This technique highlights the most robust predictors while diminishing others, streamlining the model's complexity.

Particularly in the LASSO model, while Mitoses initially appears to have a non-zero coefficient at lower lambda values, it is eventually shrunk to zero as lambda increases, reflecting its lesser predictive importance compared to other variables. Meanwhile, the coefficients for Cl.thickness, Cell.shape, Bare.nuclei, and Bl.cromatin remain non-zero across a broad range of lambda values, underlining their consistent relevance in distinguishing between benign and malignant samples.

The observation from the LASSO path plot, complemented by the coefficient estimates from the Subset Selection model, provides a nuanced understanding of feature importance. While the Subset Selection model offers a detailed coefficient estimate for each variable, the LASSO model elucidates the variable importance

across various regularization levels, allowing for a dynamic assessment of feature significance. This dynamic view, afforded by the LASSO path plot, confirms the robustness of certain predictors and offers a pragmatic approach to model simplification by phasing out less critical predictors like Mitoses when appropriate.

Variables Drop-out

Table 4: Comparison of Coefficients from Subset Selection and LASSO Logistic Regression

Variable	Subset_Coefficient	LASSO_Coefficient
(Intercept)	-9.914	-8.255
Bare.nuclei	0.419	0.464
Bl.cromatin	0.576	0.302
Cell.shape	0.388	0.304
Cell.size	NA	0.264
Cl.thickness	0.415	0.330
Epith.c.size	NA	0.141
Marg.adhesion	0.326	0.033
Mitoses	0.525	0.000
Normal.nucleoli	0.178	0.169

The comparison of coefficient values between the Subset Selection and LASSO Logistic Regression models reveals interesting insights into variable significance. In the Subset Selection model, variables such as Cell.size and Epith.c.size were not included, suggesting their exclusion during the model simplification process due to their lower significance in predicting malignancy within the context of other variables.

Conversely, the LASSO model presents a different aspect of variable selection, as evidenced by the coefficient for Mitoses being reduced to zero, thereby indicating its exclusion. This is in stark contrast to the Subset Selection model, where Mitoses holds a substantial coefficient (0.5250), suggesting its potential influence on the outcome. This delineation showcases the LASSO model's capacity for regularization and variable selection, where it has discerned Mitoses to be less pivotal amidst the presence of other more influential predictors.

Relationships between the response and predictor variables

Analyzing the coefficients from the Subset Selection logistic regression model, I observe that predictors like Cl.thickness, Cell.shape, and Bare.nuclei positively influence the log odds of a sample being malignant. Specifically, Bare.nuclei exhibits a notable coefficient (0.4186), reflecting a strong positive relationship with the probability of malignancy.

The LASSO model retains key predictors such as Cl.thickness, Cell.shape, and Bare.nuclei with substantial non-zero coefficients, thereby affirming their predictive importance as seen in the Subset Selection model. The consistent retention of these variables across both models accentuates their significant roles in classifying the tissue samples. Although Mitoses is dismissed in the LASSO model, its inclusion in the Subset Selection model may imply a degree of influence that the LASSO penalization overshadowed.

Collectively, the variables retained in both models are instrumental to the classification process. The consistent coefficients across models enlighten us on the attributes most strongly correlated with malignant breast cancer tissue. These findings illustrate the nuanced interplay of variables and reinforce their collective utility in differentiating between benign and malignant samples, with an acknowledgment of Mitoses' nuanced contribution in the broader model context.

Comparing the Accuracy of Two Models

In order to compare the performance of both models, I used cross-validation based on the test error:

Table 5: Test Error Camparison Between Two Models

Model	Test_Error
Subset Selection	0.0583942
LASSO	0.0656934

The test dataset comparison between the subset selection logistic regression and the LASSO logistic regression models demonstrates a discernible difference in performance. The subset selection model registers a test error of approximately 5.84%, which is slightly lower than the 6.57% error rate of the LASSO model. This indicates a marginally higher predictive accuracy for the subset selection model in classifying breast cancer samples.

Despite the LASSO model's regularization leading to a sparser model with fewer variables, it does not surpass the subset selection model in predictive accuracy. Notably, the LASSO model includes variables such as Cell.size and Epith.c.size, which the subset selection model omits. This inclusion might suggest that the LASSO model, while slightly less accurate overall, could be capturing additional nuances of the data through these variables.

This finding suggests a nuanced trade-off between model parsimony and predictive power. In situations where maximum accuracy is essential, such as medical diagnostics, the subset selection model might be favored for its inclusivity of significant variables, despite its higher complexity.

The comparison underscores the necessity of a balanced approach to model selection, considering both the simplicity of the model and its capacity for accurate predictions. The LASSO model's slight reduction in accuracy points to the potential cost of its parsimony, particularly when excluding variables that may hold some predictive importance.

Final Choice on Classifier

In determining the most suitable classifier, the subset selection model is chosen over the LASSO model, primarily due to its slightly superior test accuracy. This model selectively incorporates variables that significantly predict the classification of breast cancer tissue, potentially leading to its marginally better performance.

Interestingly, the LASSO model, while typically known for its ability to simplify models by excluding less critical predictors, has actually retained a broader set of variables including Cell.size and Epith.c.size. This suggests a complexity to the LASSO model that is not present in the subset selection model, which has opted to exclude these variables.

When considering misclassification errors, both models are susceptible to false positives and false negatives. Given the high stakes of breast cancer diagnosis, the nature of these errors is crucial. While the precise error tendencies of the models are not elaborated upon in this analysis, it is imperative to conduct a detailed assessment of the models' performance to understand their propensities for different error types.

Ultimately, the selection of the subset selection model as the classifier of choice strikes a balance between achieving the highest accuracy and maintaining an acceptable level of complexity. It is chosen based on its marginally better performance on the test dataset, while also considering the potential implications of misclassification errors in a sensitive medical setting.