Analysis of Cancer Data

Group 37: Gaoli Lin, Haowei Yan, Haotian Liu, Yizhou Gu

1. Introduction

Modern mass spectrometry collects thousands of molecular features, but analyzing such high-dimensional data is challenging. Identifying key patterns can aid early diagnosis and improve treatment. To explore the question, the Arcene dataset will be divided into a training set and a test set. A unique training set containing 100 samples with 5,000 randomly selected features. A fixed test set (100 samples with all 10,000 features) will be used for model evaluation. The study will employ various classification techniques to evaluate their ability to distinguish between cancerous and normal tissue samples. The primary research question is whether biochemical features can accurately differentiate between these two types of tissue. Additionally, the study will compare different classification models, including **Tree-based methods**, **Discriminant Analysis methods**, **SVM**, and **Neural Networks**, to identify the best-performing model.

2. Data Processing

2.1Initial rejection of potential probes

Based on the dataset description, we performed an initial feature selection process to reduce the dimensionality of the data and eliminate variables that may not provide meaningful insights for the model. We focused on removing features with extremely low variance, as these are typically considered to be probes or noise sequences that do not contribute valuable information for predictive modeling. By removing these low-variance features, we reduced

the data dimensionality, which facilitates more efficient analysis and model training, and helps to prevent overfitting by removing redundant or irrelevant information.

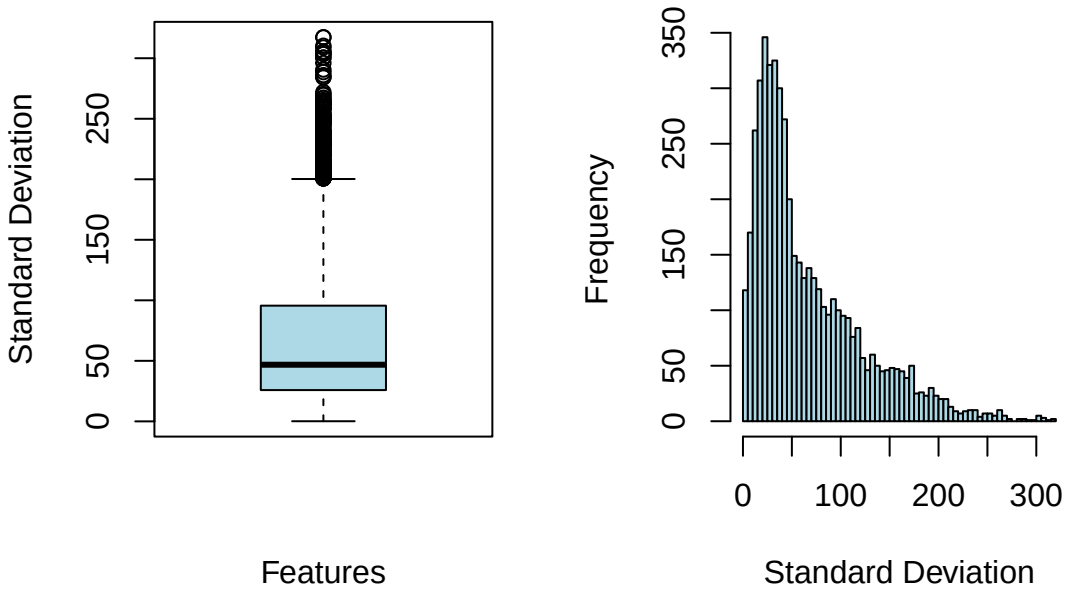


Figure 1: Variable Standard Deviation Distribution

2.2 Data Dimension Reduction Processing

PCA was further applied to reduce the dimensionality of the data. The Kaiser criterion and the cumulative variance contribution ratio were used to select the effective principal components. Based on these criteria, the original dataset was updated, leading to a significant reduction in its dimensionality. As a result, we now have three datasets for training: the original dataset, the dataset after initial variable selection, and the dataset after PCA dimensionality reduction. The subsequent analysis will be conducted using each of these datasets separately. The datasets contain features derived from different preprocessing methods: Dataset 1: Original dataset with 5000 features. Dataset 2: Feature selection reduced the number of features to 1824. Dataset 3: Principal Component Analysis (PCA) reduced the feature count to 47.

3. Formal data analysis

3.1 Discriminant-based methods

3.1.1 Linear Discriminant Analysis

LDA is a classification method based on probability distributions. It assumes that all classes share the same covariance structure, resulting in a linear decision boundary. This assumption makes LDA effective when the class distributions are similar but may limit its performance when dealing with more complex data.

However, LDA relies on the assumption of homogeneous covariance matrices across classes.The Box's M-test for homogeneity of covariance matrices produced a very small p-value, suggesting that the assumption of equal covariance matrices does not hold. The violation of this assumption suggests that LDA may not be the most suitable method for this dataset, as it could lead to suboptimal classification performance.

3.1.2 Quadratic Discriminant Analysis

QDA extends LDA by allowing each class to have its own covariance structure, leading to more flexible, curved decision boundaries. This flexibility makes QDA suitable for non-linearly separable data, but it also increases the risk of overfitting, especially when the dataset is small.

Two classification models, LDA and QDA, were applied to the data. The accuracy of LDA was 73%, with a confusion matrix showing 45 true negatives, 28 true positives, 11 false positives, and 16 false negatives. QDA achieved a higher accuracy of 79%, with 48 true negatives, 31 true positives, 8 false positives, and 13 false negatives. The ROC curves for both models were plotted, and the AUC values were displayed. QDA had a higher AUC than LDA, indicating better performance in distinguishing between the two classes.

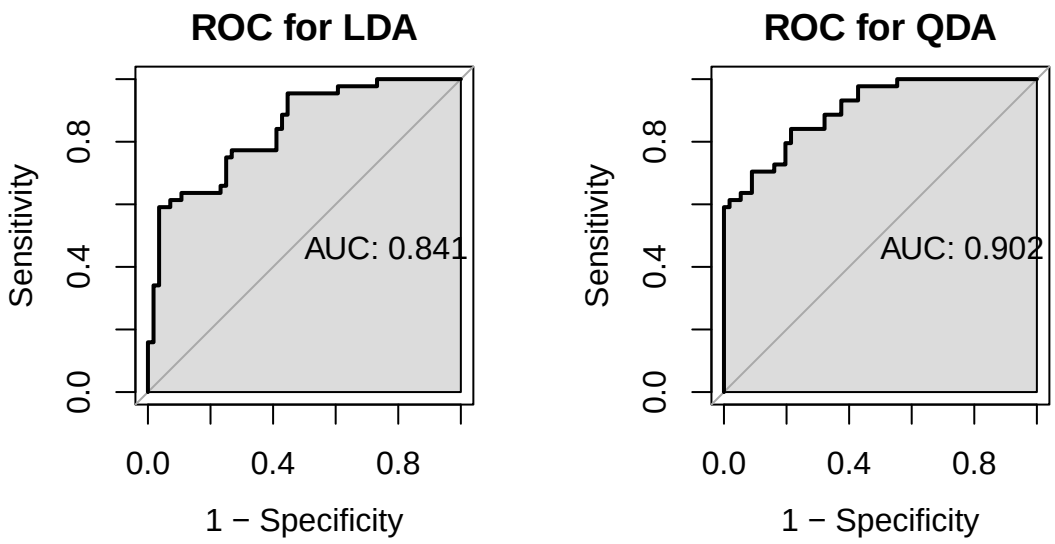


Figure 2: ROC

3.2 Tree-based methods

3.2.1 Classification Tree

Classification Tree partition the feature space into a number of disjoint and non-overlapping regions.And predict the class of a given observation as the most commonly occurring class of training observations is the region to which it belongs. A classification tree is typically suitable for smaller datasets because it is easy to interpret during the training process and can quickly generate predictions. The analysis was then performed using the dataset processed with PCA

The pruning process is based on the complexity parameter (cp) selection. The optimal cp is chosen as the largest value within the range where the cross-validation error remains within one standard deviation of the minimum error.Based on this criterion, a cp of 0.032 is selected to prune the new tree.

The classification tree also highlights the importance of variables, showing that PC1 is the most significant, followed by PC12 and PC2. Variables like PC18, PC13, PC21, and PC24 are considered less important.

The test accuracy is 0.68, indicating poor performance. While the classification tree has some predictive ability, its overall effectiveness is limited.

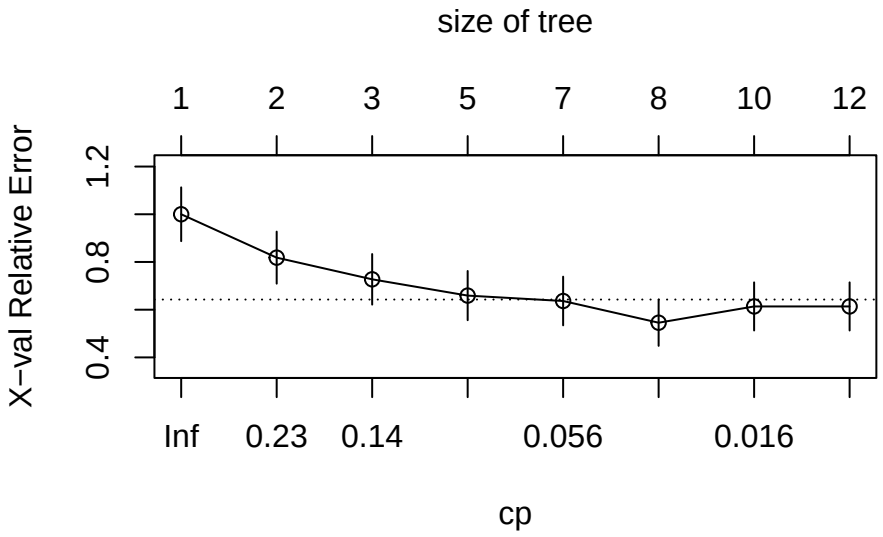


Figure 3: Model Performence with CP

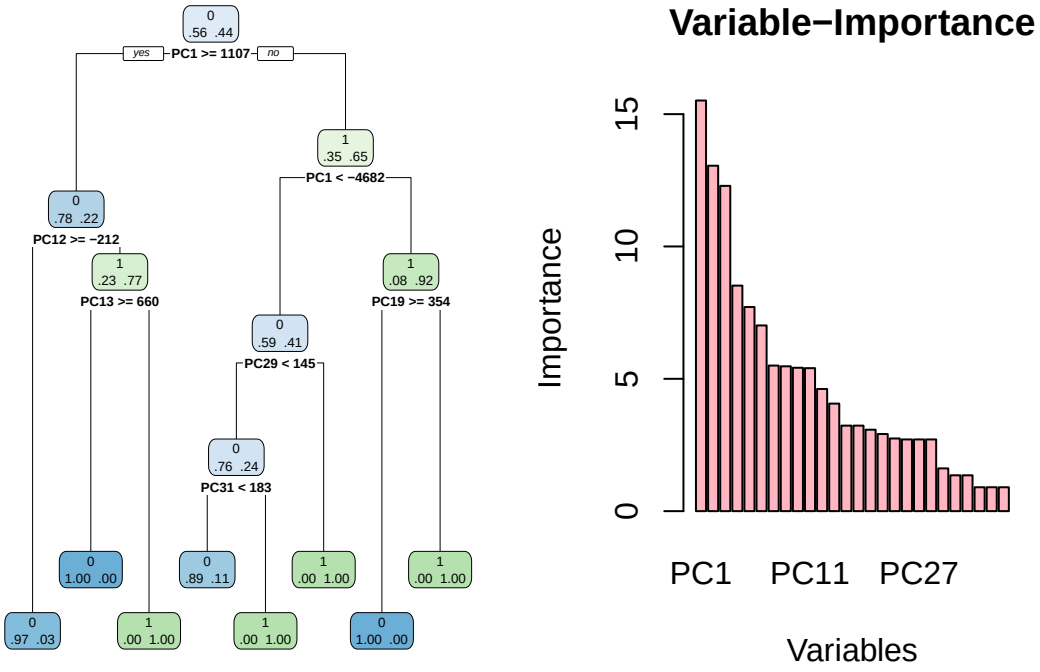


Figure 4: Pruned Classification Tree (cp=0.032)

3.2.2 Bagging Tree

Bagging involves repeatedly drawing samples from the original dataset and building a classification tree on each bootstrapped sample. For each test observation, the class predicted by each tree is recorded. The final prediction is determined by a majority vote, where the most frequent class across all predictions is chosen. To find the minimum number of trees that stabilize the OOB error, the model was trained with varying tree numbers, and the OOB error was monitored. Once the error stabilized, the smallest number of trees achieving this was selected to build the final Bagging Tree model.

Since bagging is an ensemble method that uses multiple classification trees, it typically requires more data to effectively train the trees and improve prediction accuracy through ensemble learning. Therefore, using a larger dataset (the dataset after initial variable selection) allows better utilization of the data's diversity, enhancing the model's stability and generalization ability.

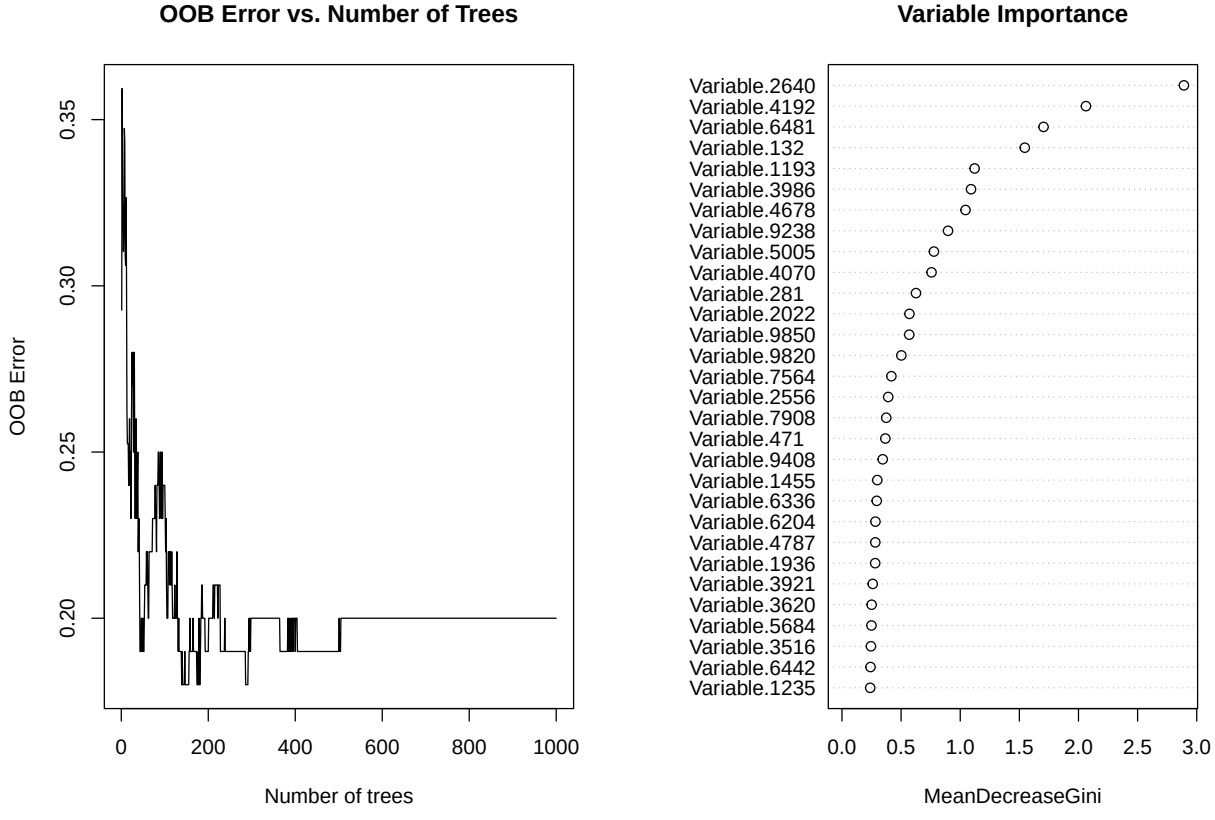


Figure 5: Model Performance

According to the Bagging Tree, Variable.2640 is the most important factor, followed by

Variable.4192 and Variable.6481, Variable.132 and Variable.1193, Variable.6442 and Variable.1235 are relatively unimportant.The Accuracy of test is 0.79, which is good. Showing the Bagging Tree has good classification ability.

3.2.3 Random Forests

Random forests improve upon Bagging Trees by reducing the correlation between individual trees. By randomly excluding a subset of variables at each split, the trees become more diverse, leading to more stable and reliable predictions. Similarly, the model was trained using the dataset after initial variable selection.

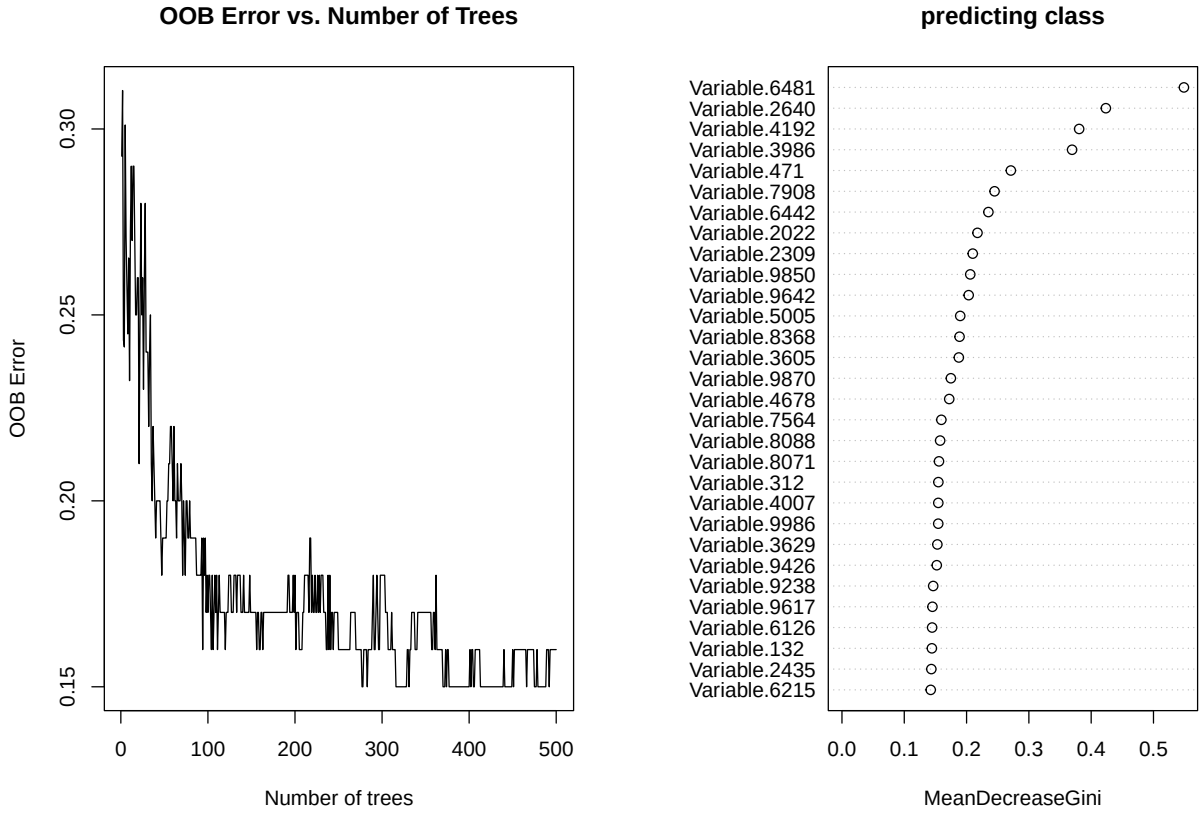


Figure 6: Model Performance

According to the random forests, Variable.2640 and Variable.4192, which were also important in the Bagging model, remain influential.

The ROC comparison shows that the AUC values for the classification tree, bagging tree, and random forest are 0.6936, 0.8953, and 0.9111, respectively.Since the random forest achieves

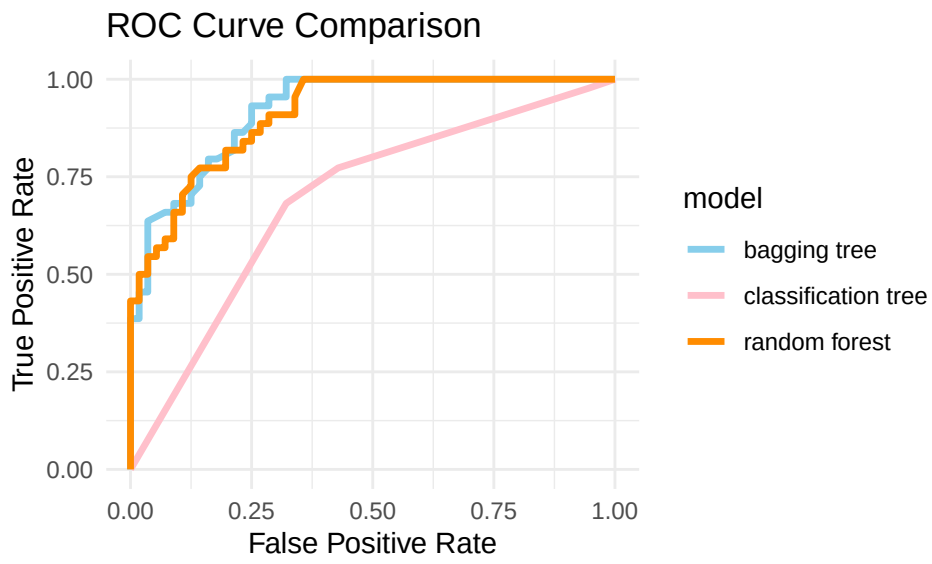


Figure 7: ROC Plot Compare

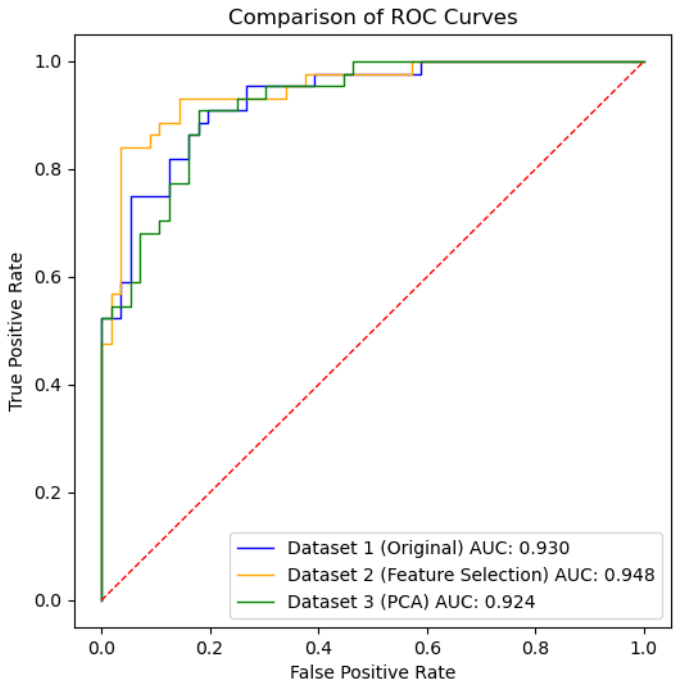
the highest AUC, it demonstrates the best predictive performance among the three models.

3.3 SVM

Support Vector Machine (SVM) is a classification algorithm that finds the optimal boundary between different classes by maximizing the margin between data points. It identifies key data points, known as support vectors, that are closest to the decision boundary and uses them to define the classification rule. SVM is effective for high-dimensional datasets and works well with small to medium-sized data. So for each dataset, SVM models were trained, tested, and optimized identify the best hyperparameters.

For Dataset 1, a linear kernel was optimal, while for Datasets 2 and 3, an RBF kernel performed better. Feature selection dataset(1800 features) provided the best performance in accuracy and AUC, while PCA dataset(40 features) caused a slight accuracy drop but remains a viable dimensionality reduction method.

Dataset	Features	Accuracy	AUC
Original	5000	0.84	0.930
Feature selection	1824	0.88	0.948
PCA Reduction	47	0.81	0.924



3.4 Neural networks Neural networks enhance predictive performance by capturing complex nonlinear relationships between features and the target variable. With multiple hidden layers and structured activation functions, they learn intricate patterns in the data and reduce dependence on any single feature, improving generalization. To save computational resources and retain as much information as possible, the model was trained using the dataset after initial variable selection and the dataset after PCA process.

3.4.1 Simple Neural networks

A simple neural network was designed with one hidden layer based on the dataset after the PCA process. This choice was made because the dimensionality of the data after PCA is smaller, and a simpler network structure is sufficient to capture the necessary information.

The model's performance was satisfactory, with an accuracy of 0.73 and an AUC of 0.76. However, there is still room for improvement in the neural network to enhance its performance further.

3.4.2 Multilayer Neural Network

To enhance the model's learning capacity, we increased the number of neurons in the first hidden layer to 533, improving its ability to capture important features. In the subsequent hidden layers, we progressively reduced the number of neurons to 256, 60, and finally 20, in order to refine the model's representation and reduce the risk of overfitting. This layered architecture allows the network to extract high-dimensional features in the initial layers and gradually distill the most relevant information in the deeper layers. The model was trained using the dataset after initial variable selection, which provided more informative input for learning.

The training performance improved significantly, with the AUC increasing from 0.64 to 0.95. This dramatic improvement suggests that the adjusted network architecture effectively enhanced the model's ability to learn complex patterns, leading to a much better classification performance. The increased network capacity in the initial layers allowed for better feature extraction, while the gradual reduction in neurons helped refine the representations, ultimately resulting in a more robust and well-generalized model.

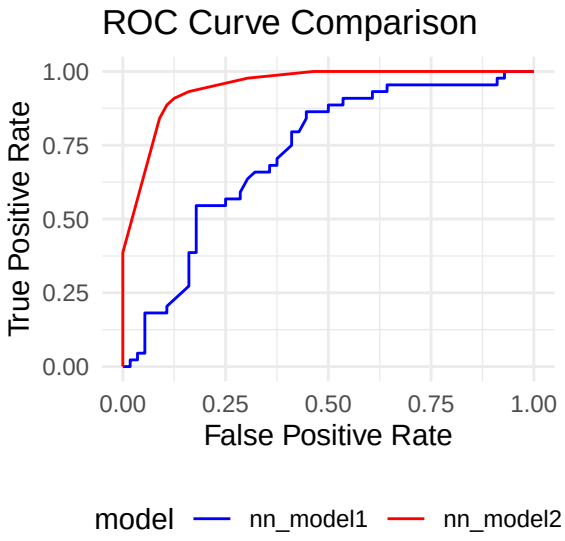


Figure 8: ROC of nn1 and nn2

4. Conclusion:

Based on the results, the neural network model performed the best, with an accuracy of 0.89 and an AUC of 0.950, closely followed by the support vector machine (SVM) with an accuracy of 0.88 and AUC of 0.948. Among the tree-based models, random forests achieved the highest AUC (0.911) and accuracy (0.81), outperforming both bagging trees and classification trees. The linear discriminant analysis (LQA) and quadratic discriminant analysis (QDA) models showed comparable performance with accuracies of 0.73 and 0.79, and AUC values of 0.908 and 0.902, respectively. Overall, the results indicate that more complex models, particularly neural networks and SVM, offer the best predictive performance.

Classification Model	Accuracy	AUC
Classification tree	0.68	0.693
Bagging tree	0.79	0.895
Random forests	0.81	0.911
LQA	0.73	0.908
QDA	0.79	0.902
SVM	0.88	0.948
Neural networks	0.89	0.950