Breast Cancer Analysis

STATISTICAL LEARNING PROJECT

Ferrero Ilaria 5209648

Pagliuca Fabiana 5212402

1 INTRODUCTION

2 EXPLORATORY DATA ANALYSIS

3 MODEL ANALYSIS

4 EVALUATION AND RESULTS

5 CONCLUSIONS

🔍 **INTRODUCTION**

2 EXPLORATORY DATA ANALYSIS

3 MODEL ANALYSIS

4 EVALUATION AND RESULTS

5 CONCLUSIONS

INTRODUCTION

Breast cancer is a prevalent condition characterized by abnormal cell growth in breast tissue. Early diagnosis is critical for effective treatment and improved patient outcomes. **Mammography** serves as a primary screening tool for identifying suspicious abnormalities in breast tissue. Subsequent **biopsy** procedures, such as fine-needle aspiration (**FNA**), provide essential diagnostic information.

This project seeks to evaluate different classification models to optimize breast cancer diagnosis, focusing on **accuracy, precision, and sensitivity**.
Identifying breast cancer early is crucial for improving survival rates and choosing the right treatments for patients.

STAKEHOLDERS

Researchers and Developers of Medical Technologies: This group is responsible for the research and development of medical technologies, including machine learning algorithms used for breast cancer diagnosis. The study provides them with a clear understanding of the algorithms' performance and effectiveness in breast cancer diagnosis.

Clinicians and Healthcare Professionals: This group includes physicians, oncologists, and other healthcare professionals, who utilize these classification models to devise personalized treatment plans, optimizing care for each patient.

AIM OF THE PROJECT

The purpose of the project is to predict whether a breast cancer cell is malignant or benign. Initially, the data was transformed to make it suitable for analysis.

Subsequently, we applied the Principal Component Analysis (PCA) technique to reduce the number of variables in the original dataset while retaining most of the significant information.

After applying PCA, an analysis was conducted on both the original and reduced datasets to compare their performance. This allowed us to assess which of the two datasets performed better, using metrics such as precision, accuracy, or sensitivity.

1 INTRODUCTION

🔍 **EXPLORATORY DATA ANALYSIS**

3 MODEL ANALYSIS

4 EVALUATION AND RESULTS

5 CONCLUSIONS

THE DATASET

The dataset contains **539 observations** and **32 parameters.** All parameters could be useful to classify cancer: if these parameters have relatively large values, it can be a sign of malignant tissue.

HOW IS IT COMPOSE?

- **Response variable:** Class ‘0’ = **Malign**, Class ‘1’ = **Benign**
- **10 feature (for each cell nucleus) -> all numerical variables**

- **mean,**
- **standard error**
- **worst** or **largest** (mean of the three largest values)

The dataset contains the average, standard error, and "worst" values for each of the 10 features.

- **Bit imbalanced:** From this diagnosis, 357 of the cases were classified as benign tumors and 212 were considered malignant tumors.

In the *figure 1* we can see that our response classes are imbalanced. In order to deal with imbalanced classes we undersample the majority class. After removing 100 random observations from the majority class, we check the class distribution again and we see that the classes are now more balanced. *(figure 2)*

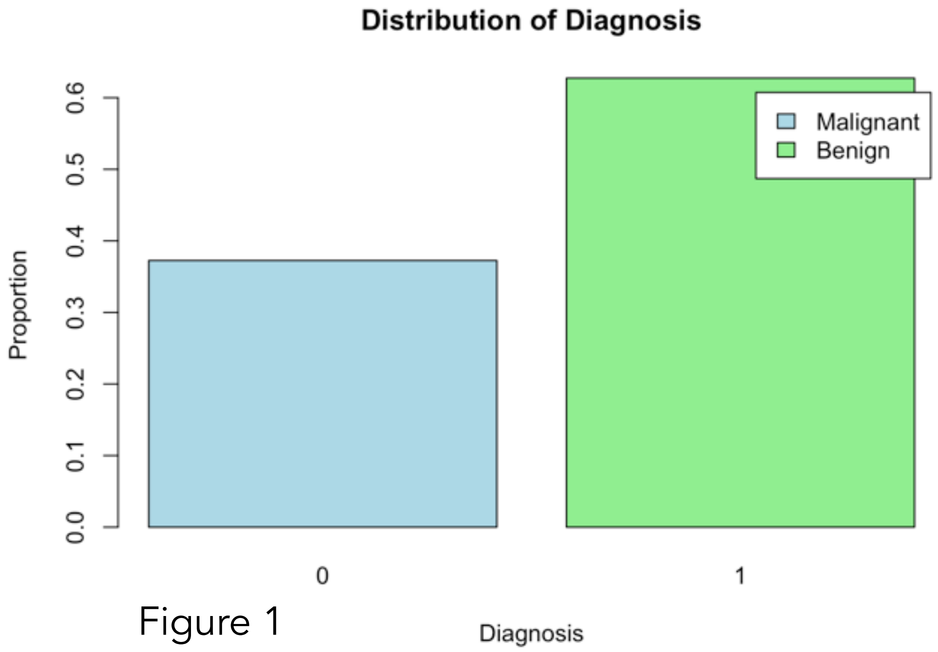


Figure 1

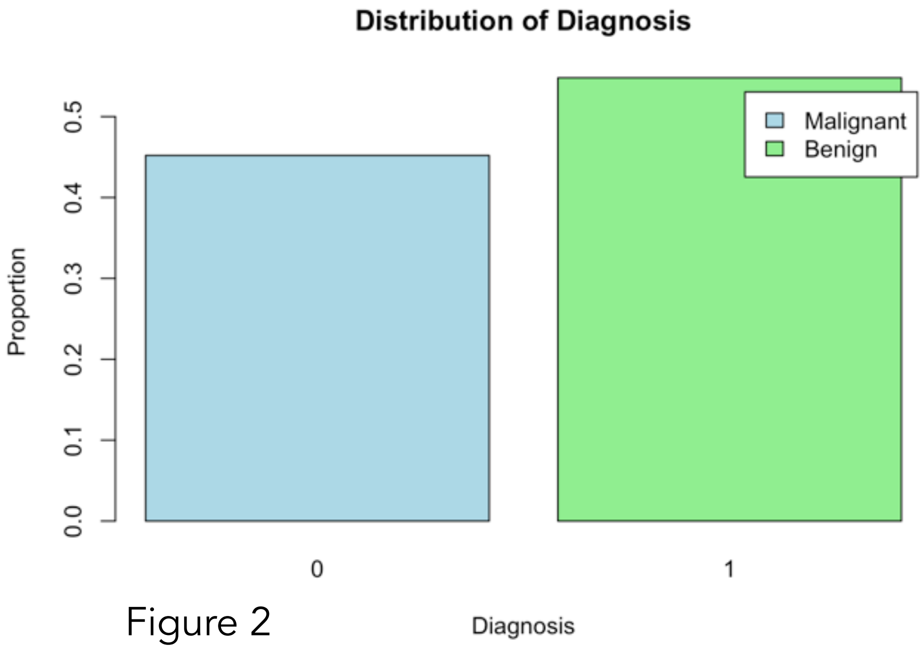


Figure 2

0	1
0.3725	0.6274

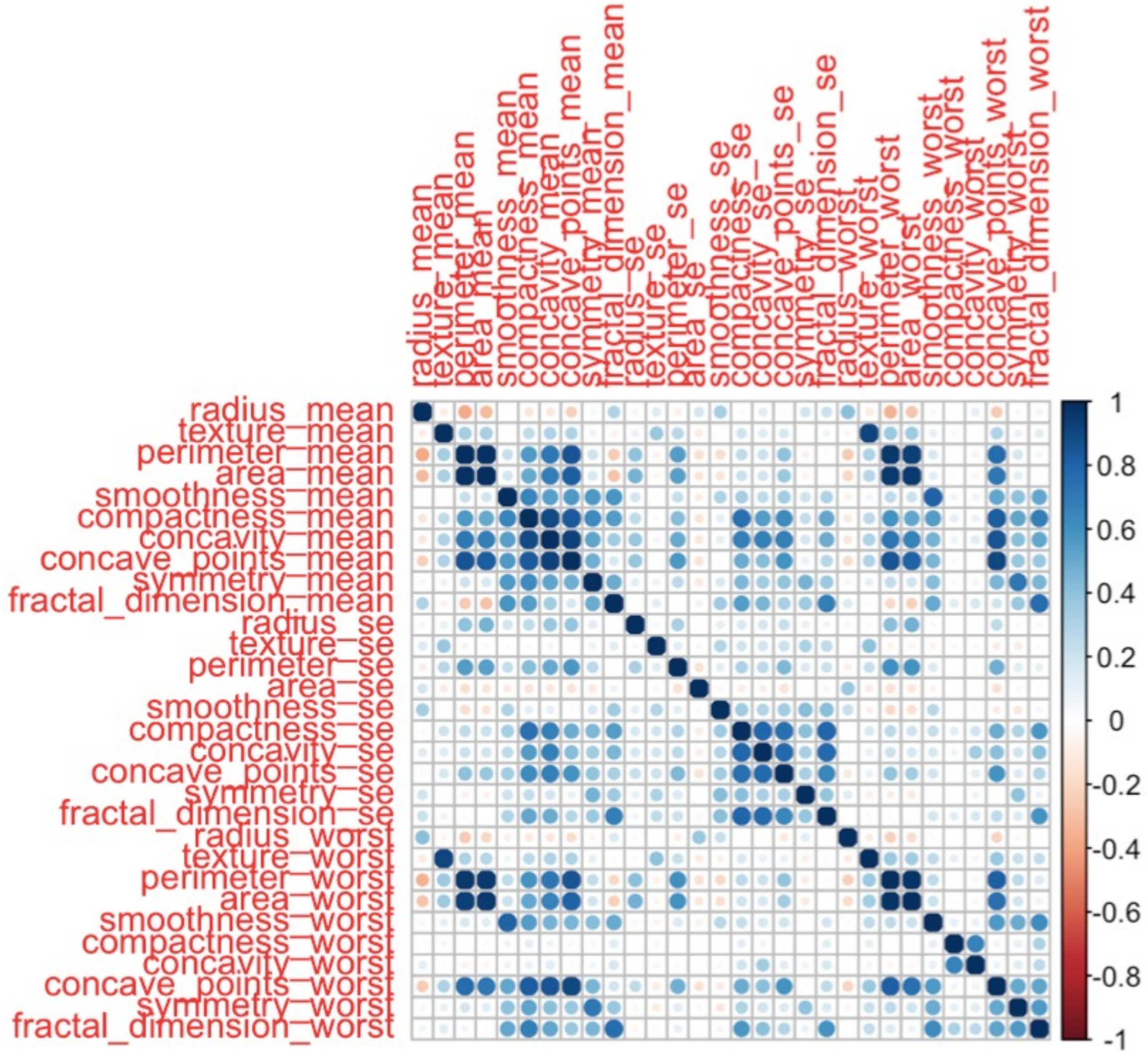
0	1
0.4520	0.5479

Correlation Matrix

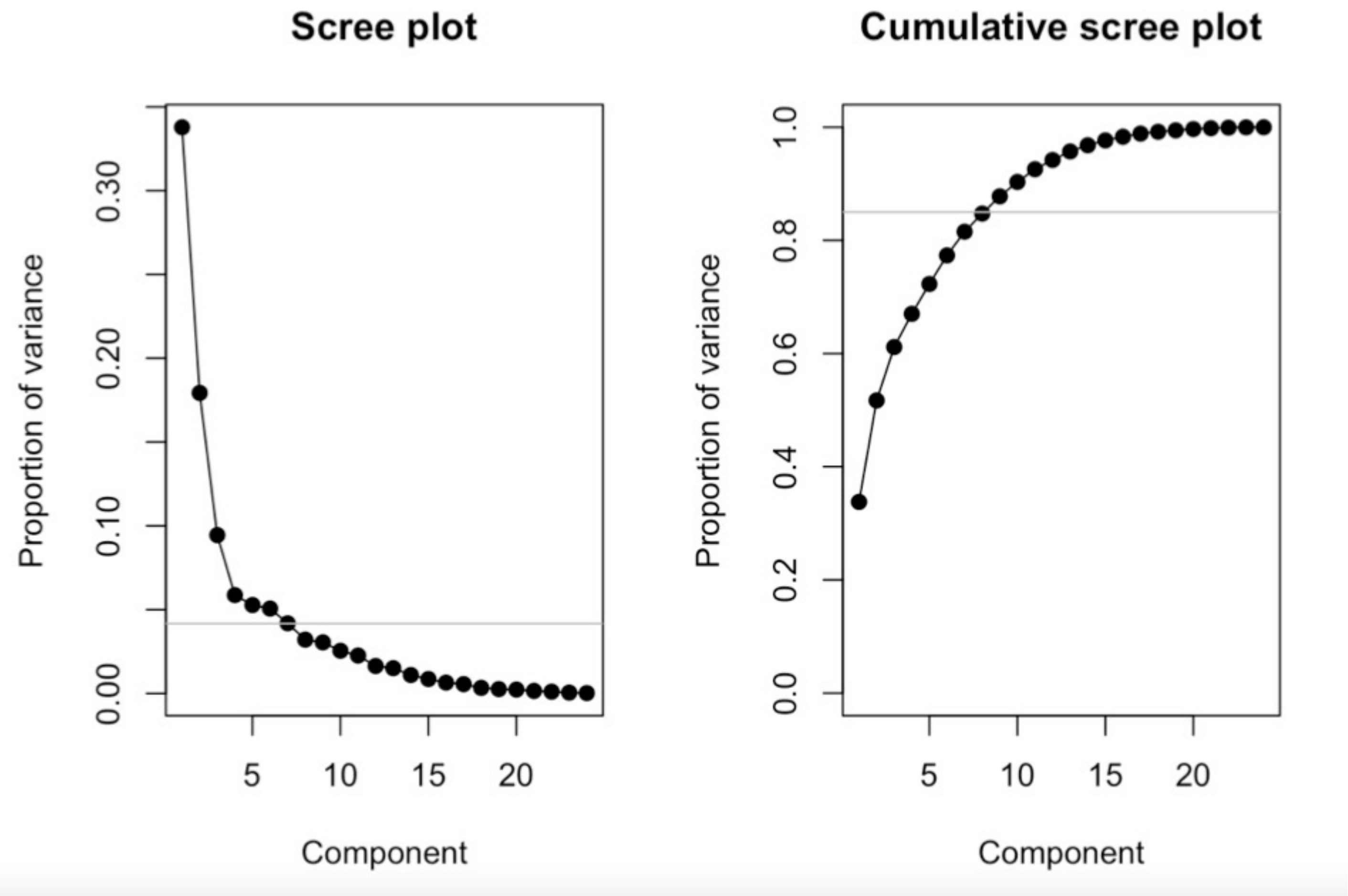
From the correlation matrix, strong positive correlations emerge between different pairs of variables in the dataset

Using a correlation threshold of **0.9**, we have identified the variables that exhibit this high correlation. These include **mean texture, mean perimeter, mean concavity, mean concave points, worst perimeter, and worst area**

As a result, we plan to select these highly correlated variables and remove them from the dataset.



Principal component Analysis



We apply the dimension technique on our balanced and filtered data set (with 25 variables)

The scree plot suggests to keep 7 components, while the cumulative scree plot the desired cumulative variance proportion (in this case, 85%), suggesting to use 8 components as well. (Elbow method)

Our second dataset consists of the **first 8 principal components** along with our response variable.

1 INTRODUCTION

2 EXPLORATORY DATA ANALYSIS

🔍 **MODEL ANALYSIS**

4 EVALUATION AND RESULTS

5 CONCLUSIONS

Diagnosis classification problem

We evaluated four different classification models including **Support Vector Machines**, **K-nearest neighbor, Random Forest and Decision tree** using features selected at different threshold levels to train the models for classifying the two types of breast cancer.

Our dataset comprises numerous features; hence, it can be advantageous to reduce the number of features used for model training. This could:

- Improve computational efficiency,
- Mitigate the risk of overfitting, and
- Simplify result interpretation.

MODEL ANALYSIS

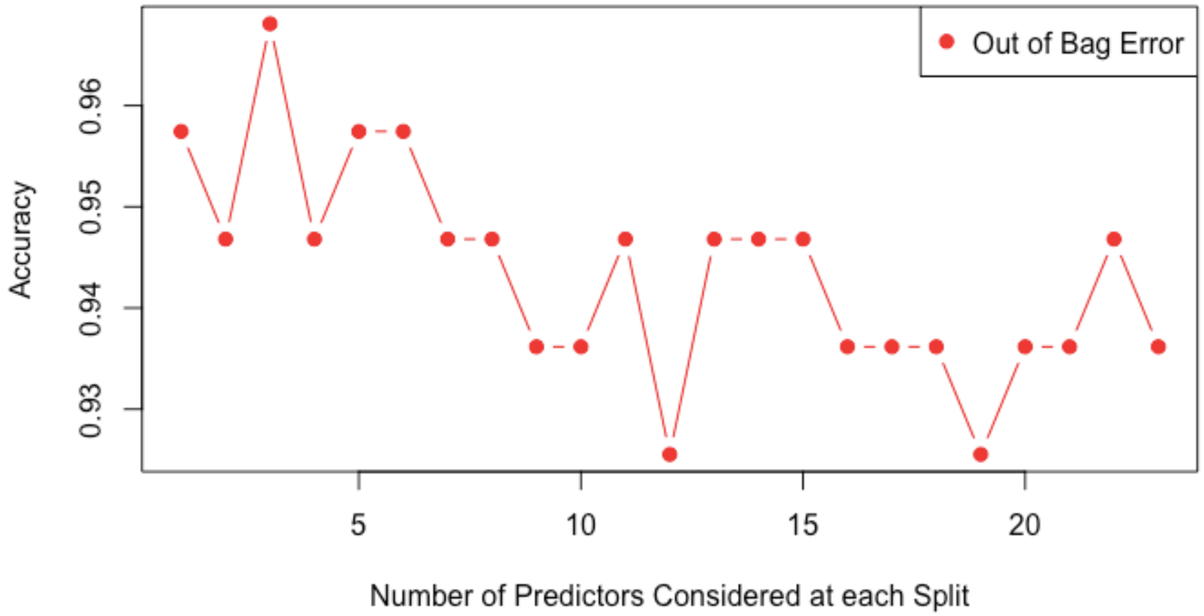
RANDOM FOREST

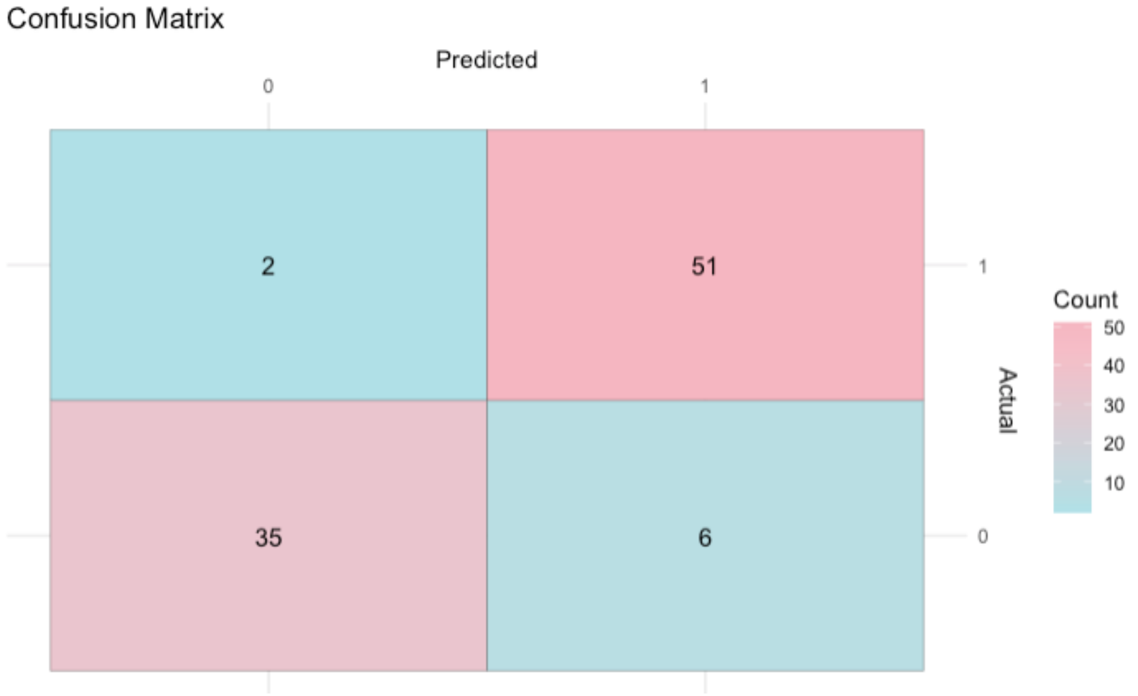
ORIGINAL dataset

Firstly, we tried to implement the Random Forest method, which involves splitting the data into subsets, training decision trees on each subset, and then combining their predictions to obtain an overall prediction.

To determine the optimal number of random variables (mtry), we evaluated the accuracy on a dataset containing 24 coefficients. We found that the optimal value for mtry is 3, since it represent the first peak on the graph .

The accuracy of the model is **91,49%**, and it can be seen in confusion matrix, we make few wrong prediction.

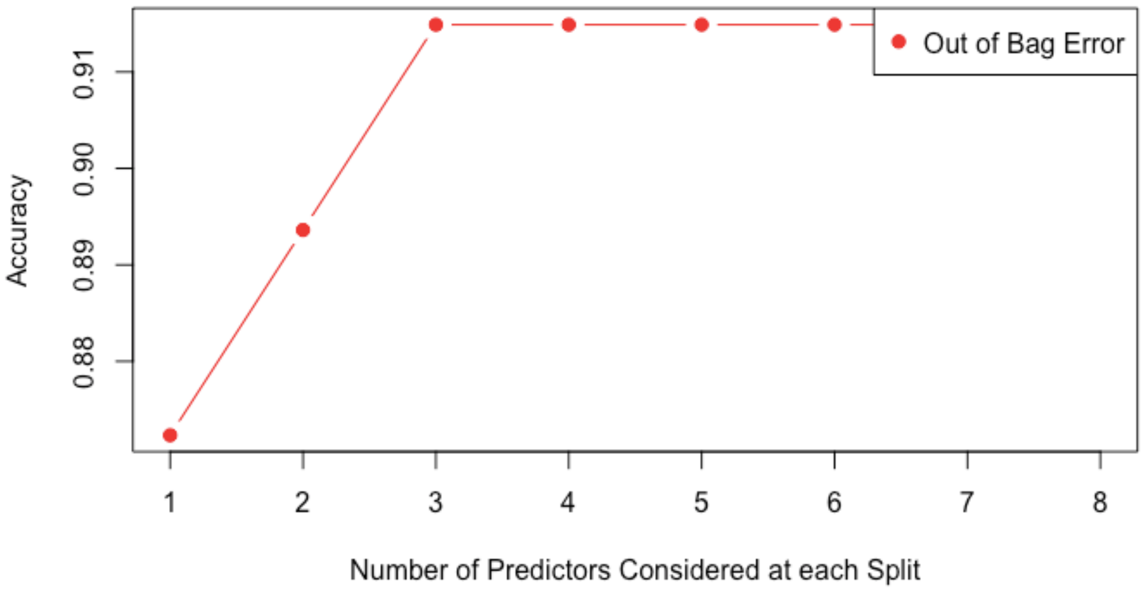




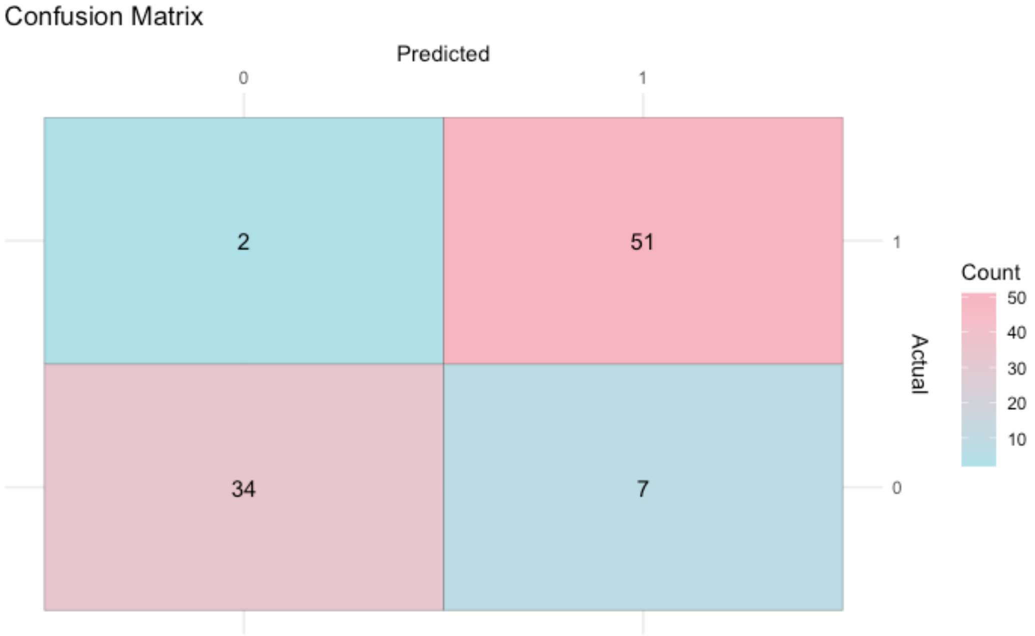
RANDOM FOREST

PCA dataset

Now, we have conducted the same procedure using our dataset reduced through PCA (the first 8 principal components). We did this to assess whether we can utilize a less complex dataset and achieve similar results.
Also in this case, we are using mtry = 3.

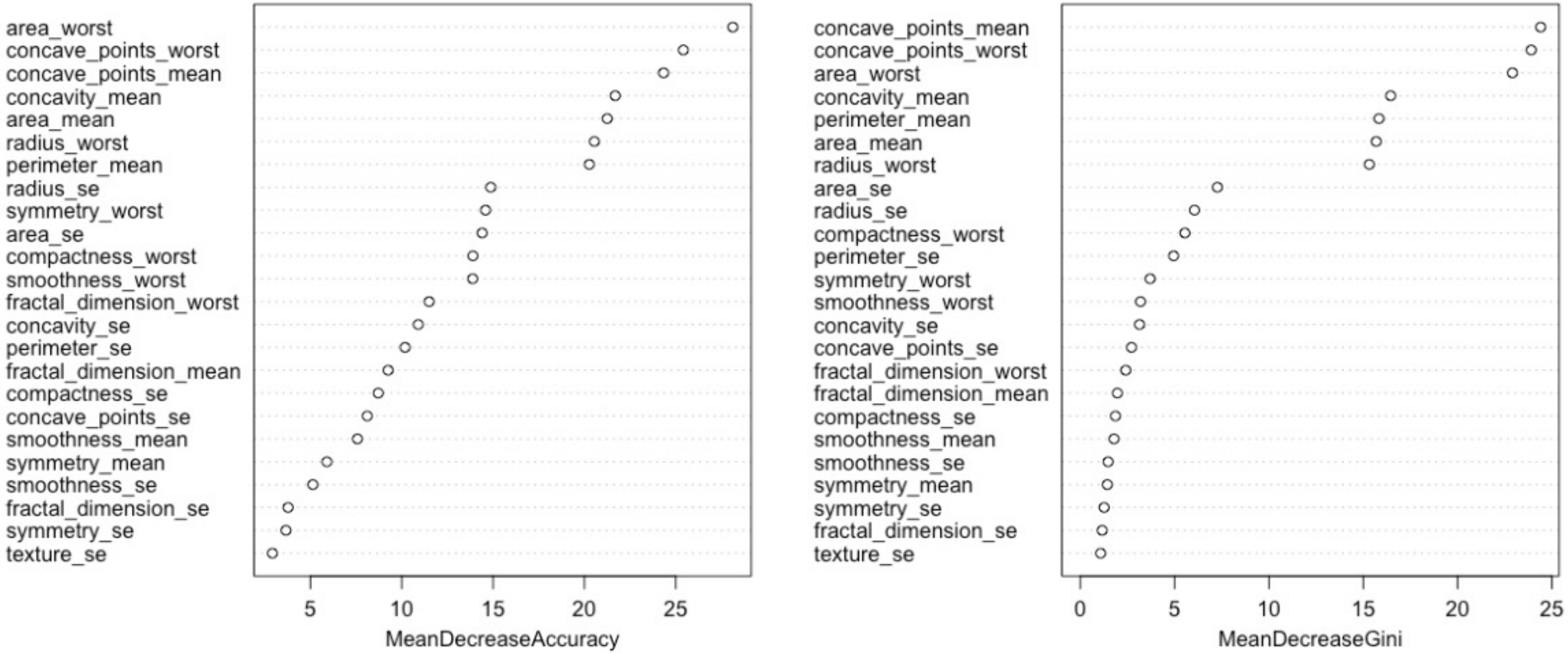


The Random Forest model trained on the dataset reduced through PCA continues to exhibit highly accurate predictions on tumor diagnoses, despite the reduction in dataset complexity.



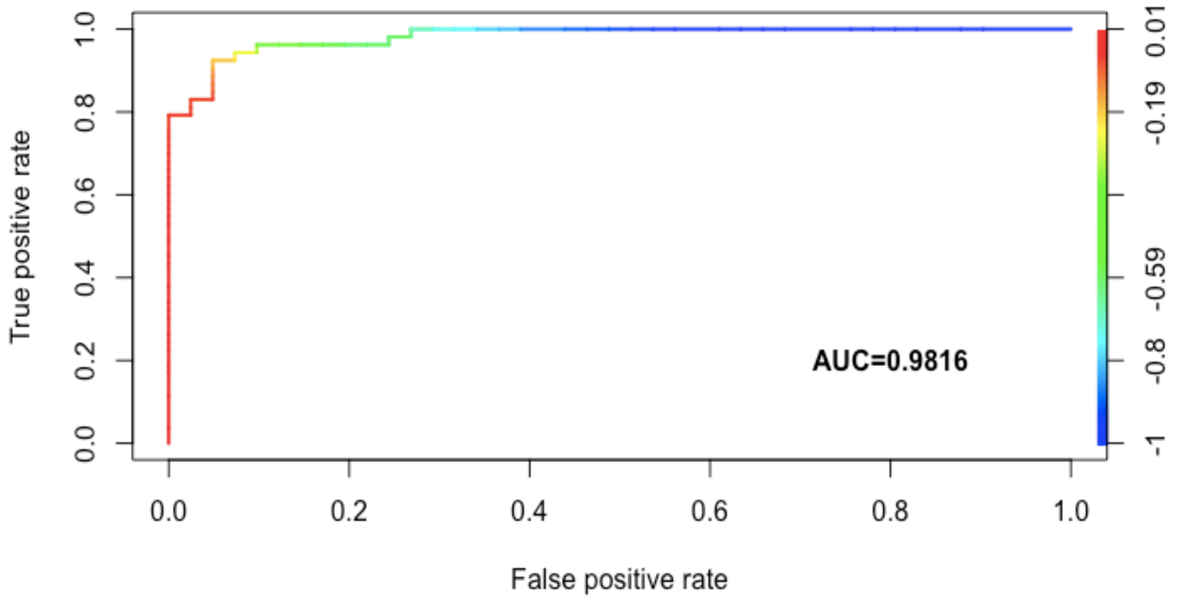
Accuracy : 90,43%

Top variables: Random Forest

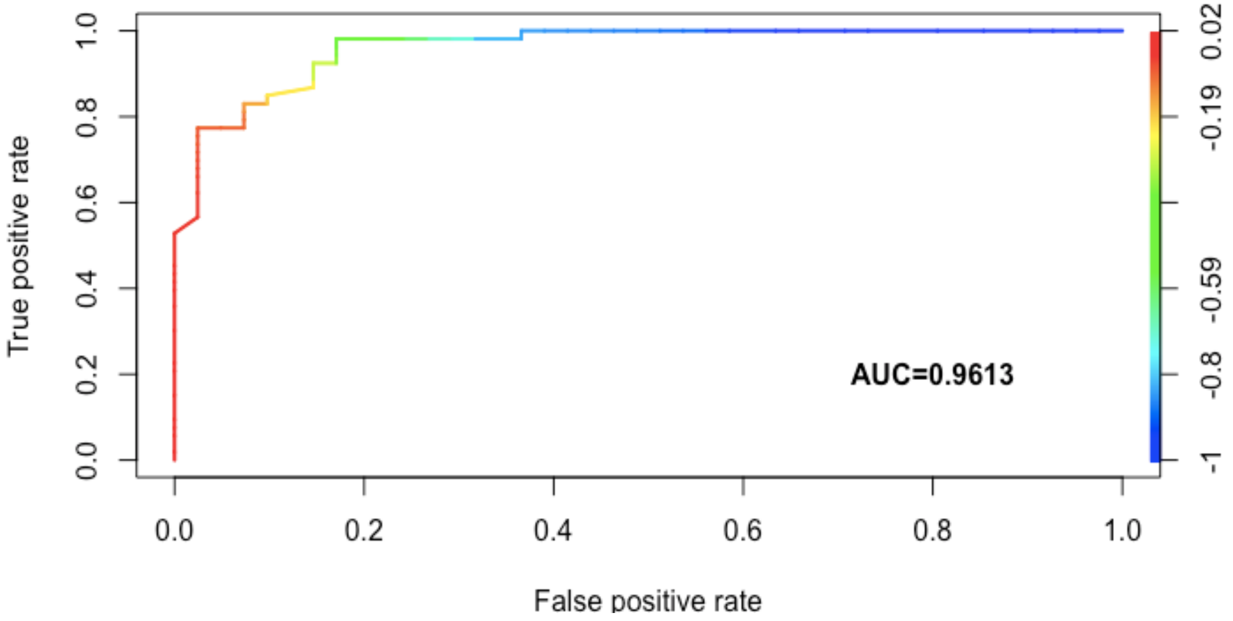


Looking at the results, **'Area Worst', 'Concave Points Worst', 'Concave Points Mean', 'Concavity Mean', 'Area Mean', and 'Radius Worst**' are the most important variables contributing to the accuracy of the random forest model.

ROC Curve: ORIGINAL dataset



ROC Curve: PCA dataset

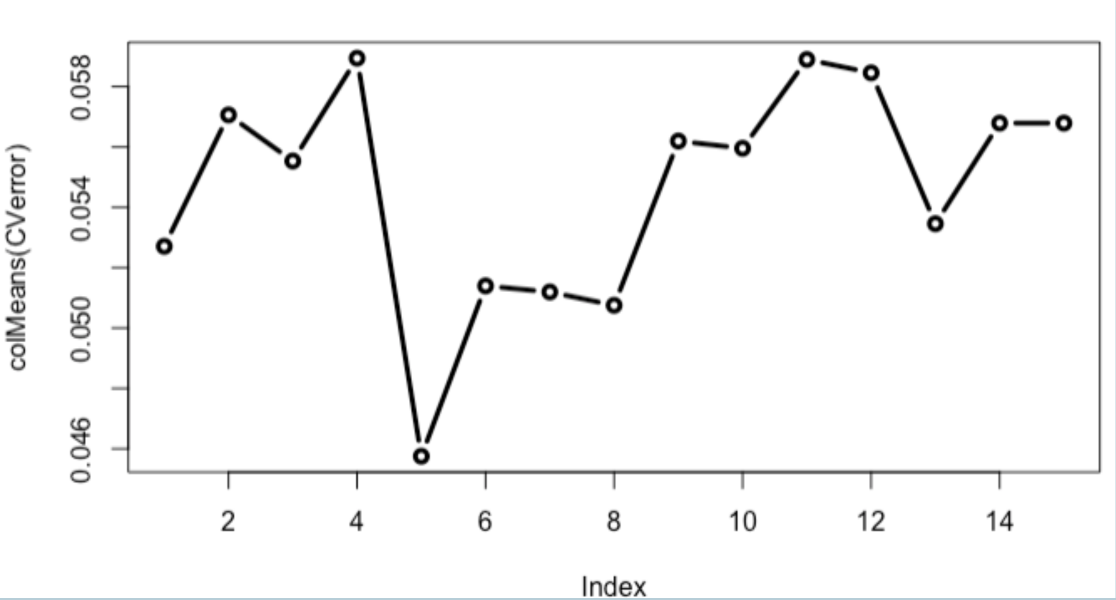


KNN: K-Nearest Neighbor

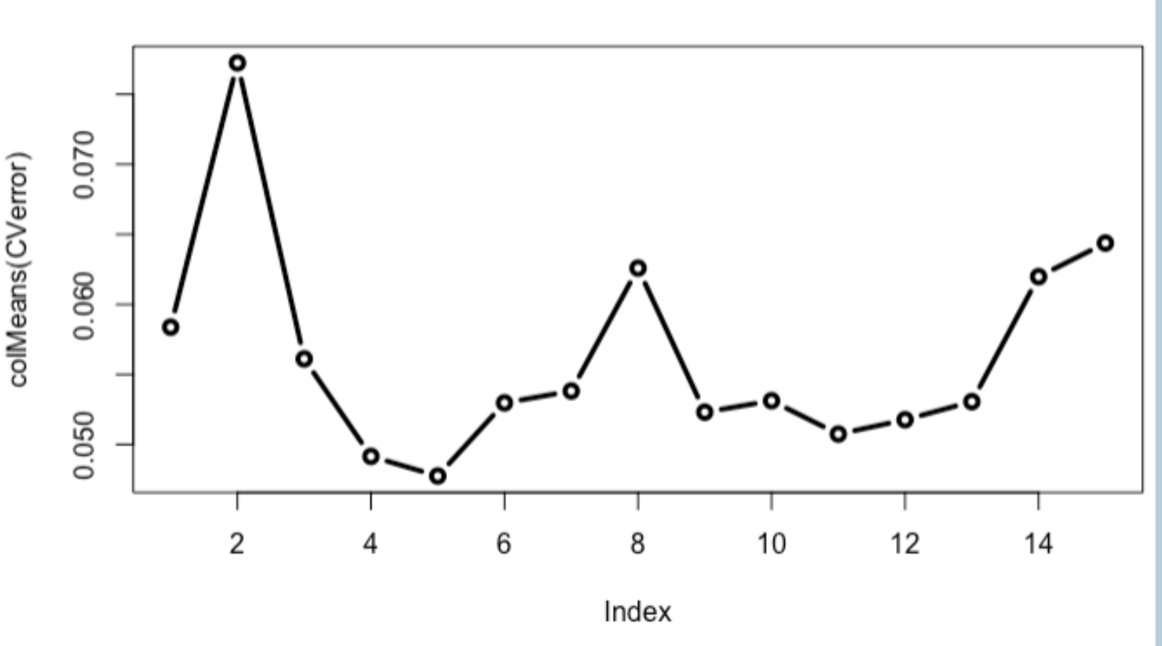
ORIGINAL dataset

- Now, we tried the K-NN approach, which is a form of classification based on the proximity of data points in their feature space (where "k" represents the number of neighbors to consider).
- We based our choice of k on cross-validation over a grid of values for k using the original dataset. We confirmed k = 5 as optimal through multiple iterations with different seeds.

set.seed (24)



set.seed (700)



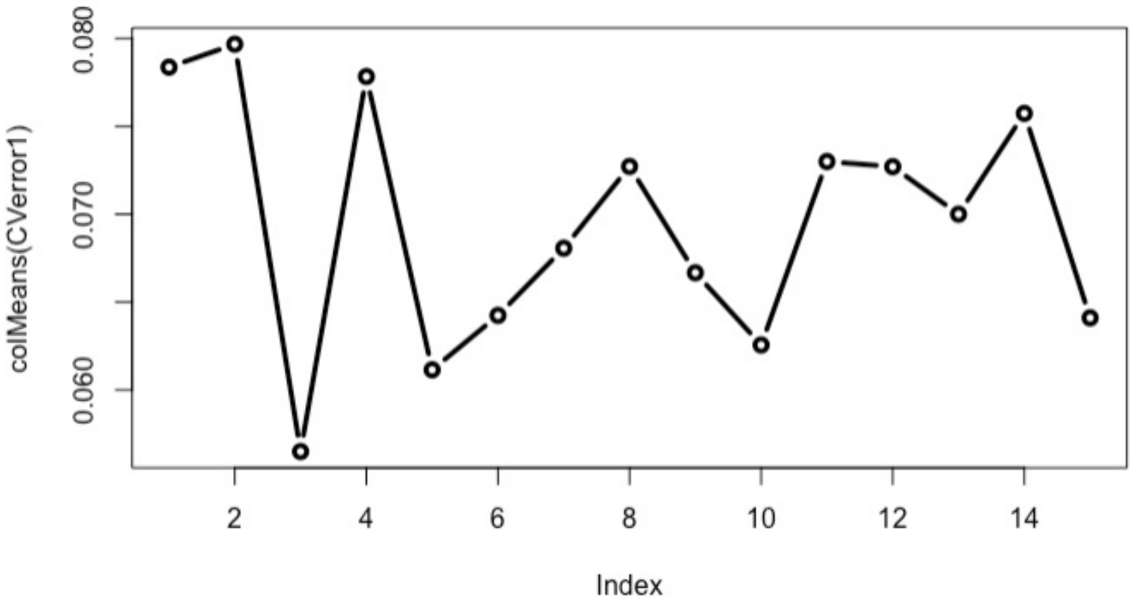
KNN: K-Nearest Neighbor

PCA dataset

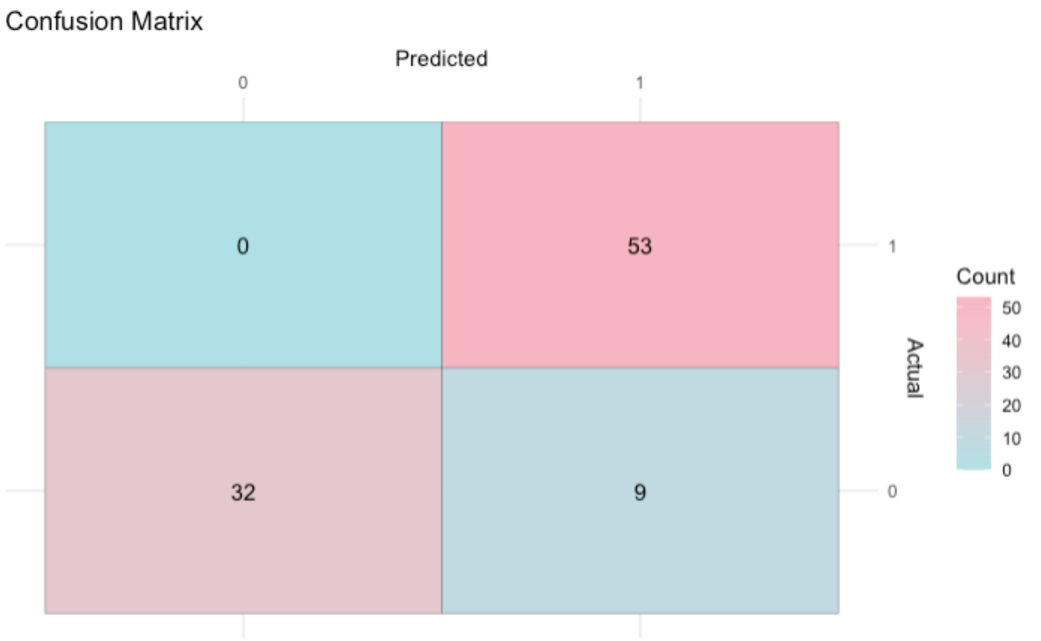
After applying K-NN with k = 3, we observe a slightly higher accuracy when using the original dataset compared to the PCA dataset. The original dataset achieves an accuracy of 90%, while the PCA dataset achieves an accuracy of 87%.

This suggests that while the PCA dataset may have effectively captured essential information in the data by reducing its dimensionality, the original dataset still outperforms it in terms of accuracy.

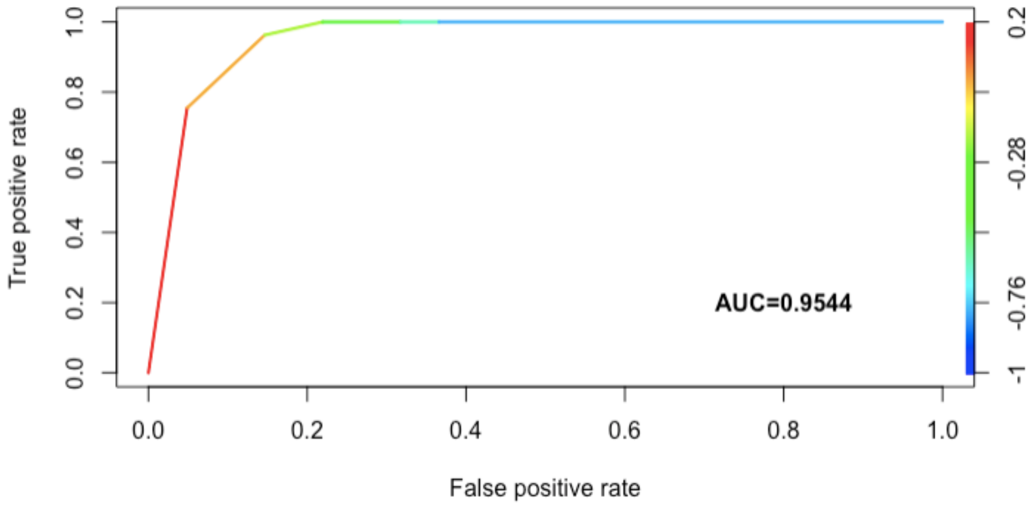
However, despite the slightly lower accuracy with PCA, it's worth noting that the reduction in dimensionality may offer other benefits such as improved computational efficiency and interpretability.



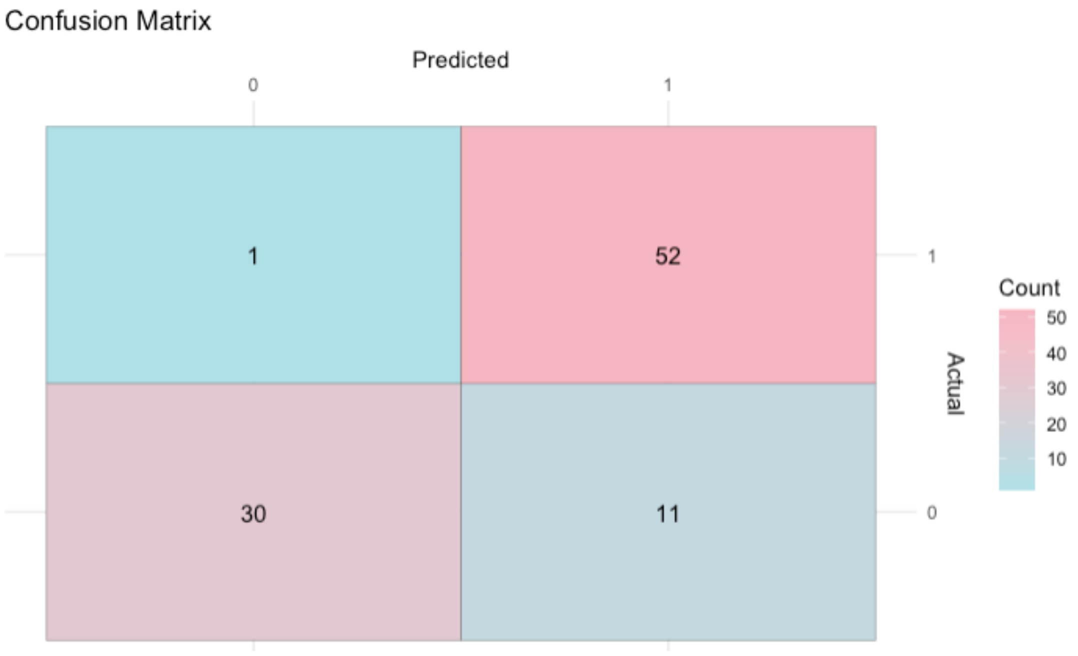
ORIGINAL Dataset



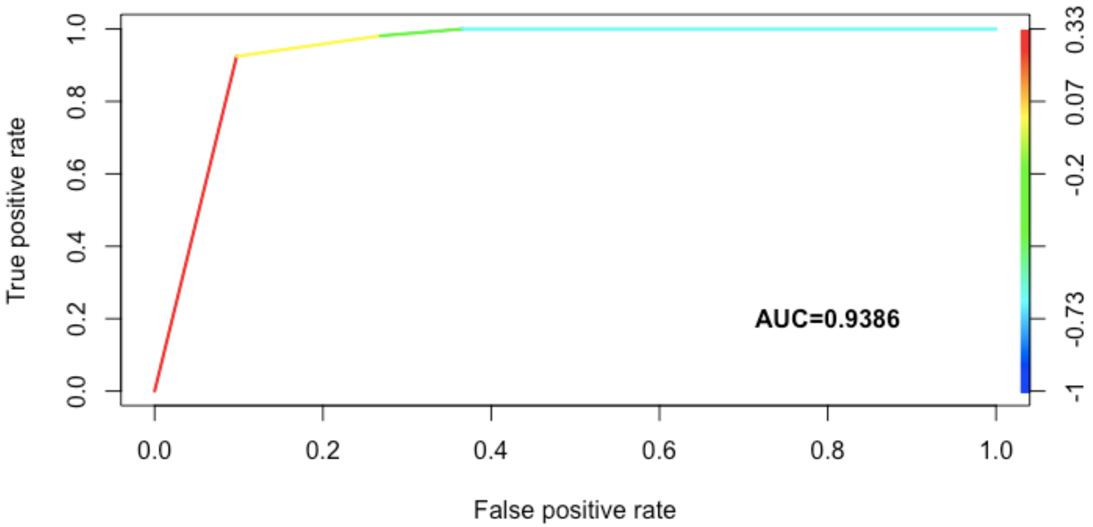
Accuracy: 90,43%



PCA Dataset

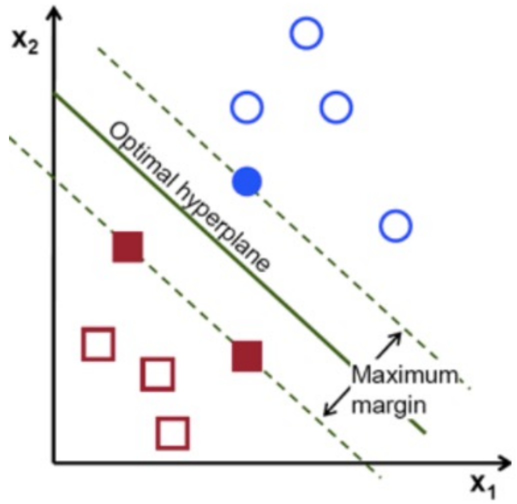


Accuracy: 87,23%



SVM: Support Vector Machine

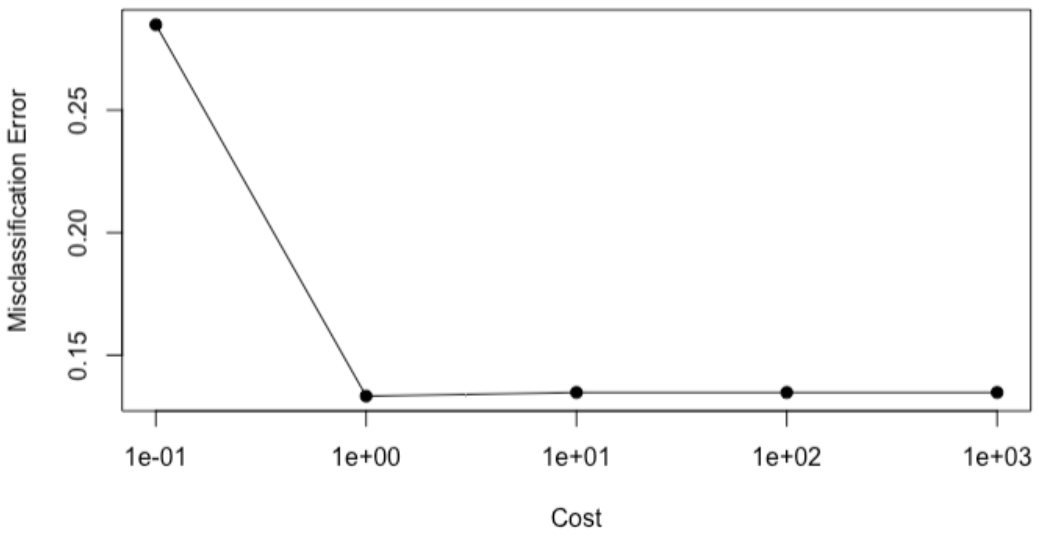
ORIGINAL dataset



- It works by identifying the optimal hyperplane that best separates data points belonging to different classes in a high dimensional space.

- **Parameter Tuning:** we fine-tuned the SVM model's parameters, such as <u>cost</u> and <u>gamma</u>, through cross-validation. This tuning process allowed us to optimize the model's performance, leading to improved classification accuracy.

- The best performing model (according to the tuning criteria) used a **cost value** of **1** and a **gamma value** of **0.5**

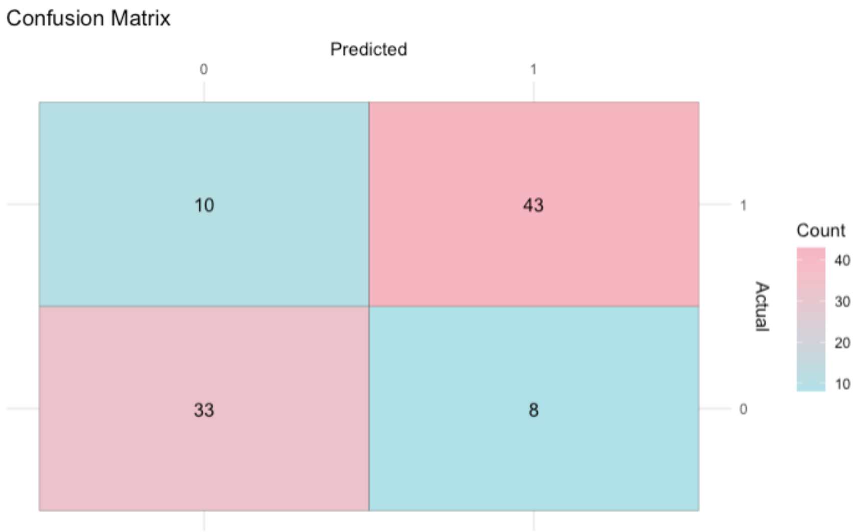


SVM: Support Vector Machine

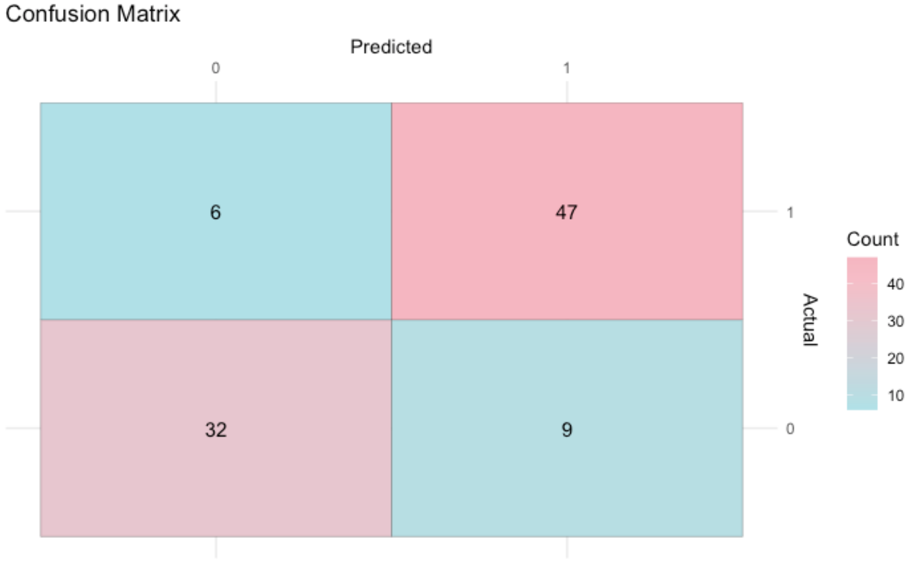
- We conducted the same steps on the PCA dataset as we did on the original dataset, and achieved an **accuracy of 84%. (also in this case, cost = 1 and gamma = 0.5)**

- The results indicate that the use of the **PCA dataset** lead to a slight increase in the **SVM model** accuracy compared to the original dataset **(accuracy = 80%)**, and also improved the model's ability to discriminate between classes, as evidenced by the enhanced **AUC** of the **ROC curve.**

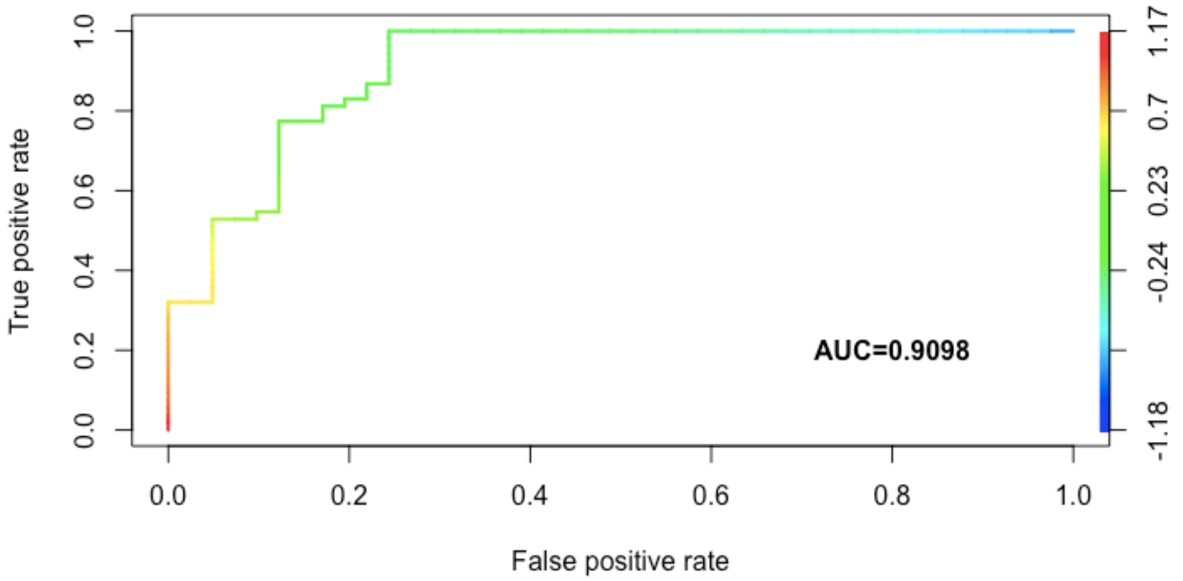
ORIGINAL dataset



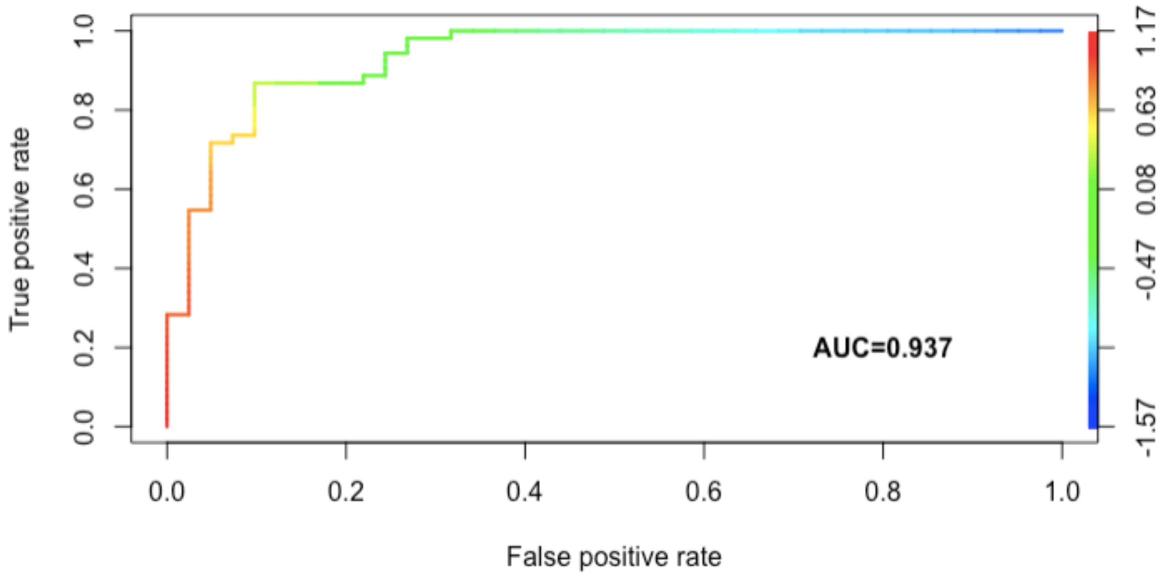
PCA dataset



ROC Curve: ORIGINAL dataset



ROC Curve: PCA dataset



In conclusion, the application of the **SVM model** to both the original dataset and the PCA dataset has been shown to have a good predictive capability for breast cancer diagnosis, with the PCA dataset offering advantages in terms of AUC of the ROC curve.

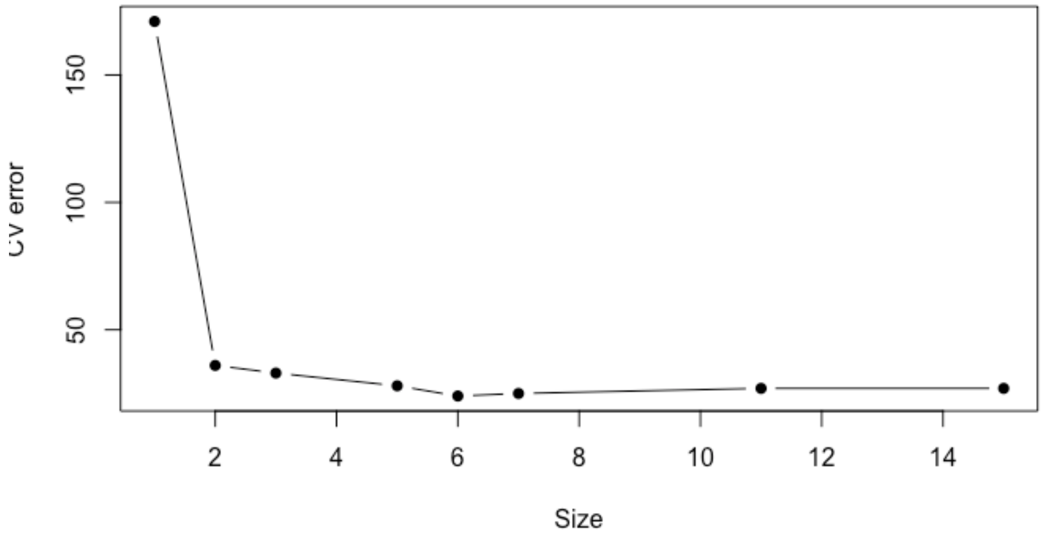
Classification Tree

- It is a predictive model that iteratively divides data into homogeneous groups based on their characteristics, creating a tree structure that classifies observations into distinct categories.

- We perform **10-fold cross-validation** to select the best pruning parameters in order to avoid overfitting.

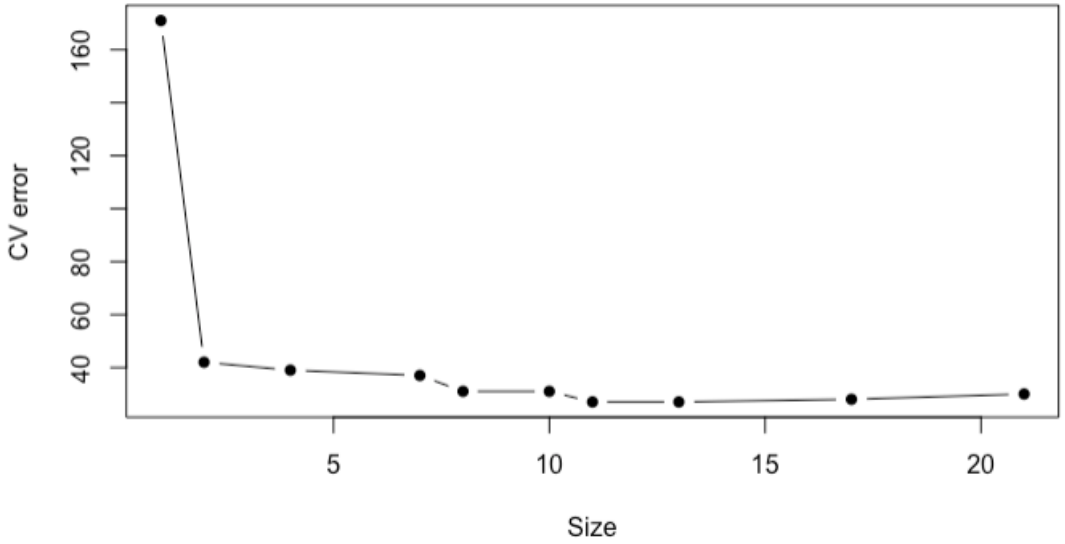
best = 6

ORIGINAL dataset

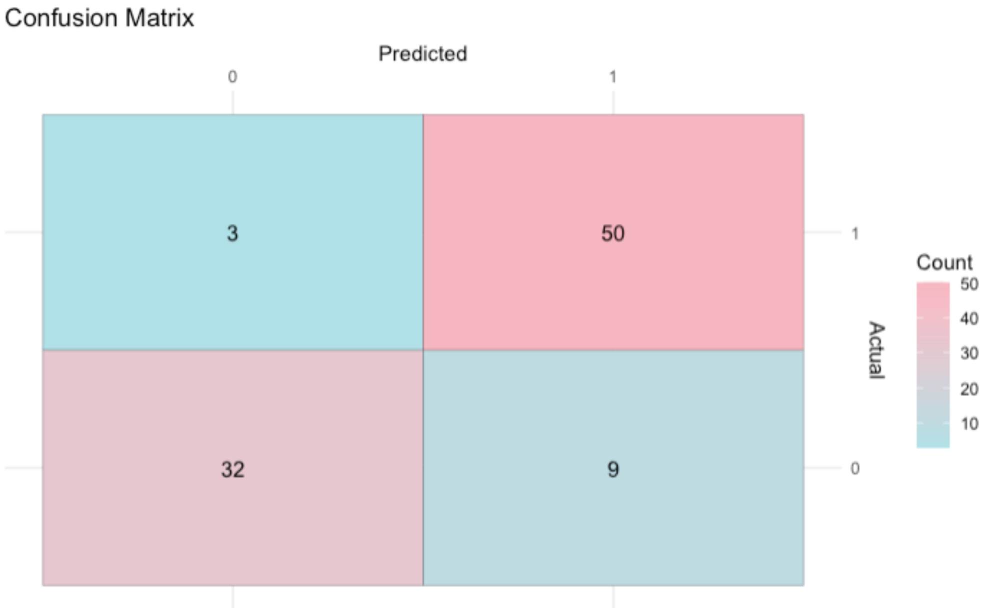


best = 11

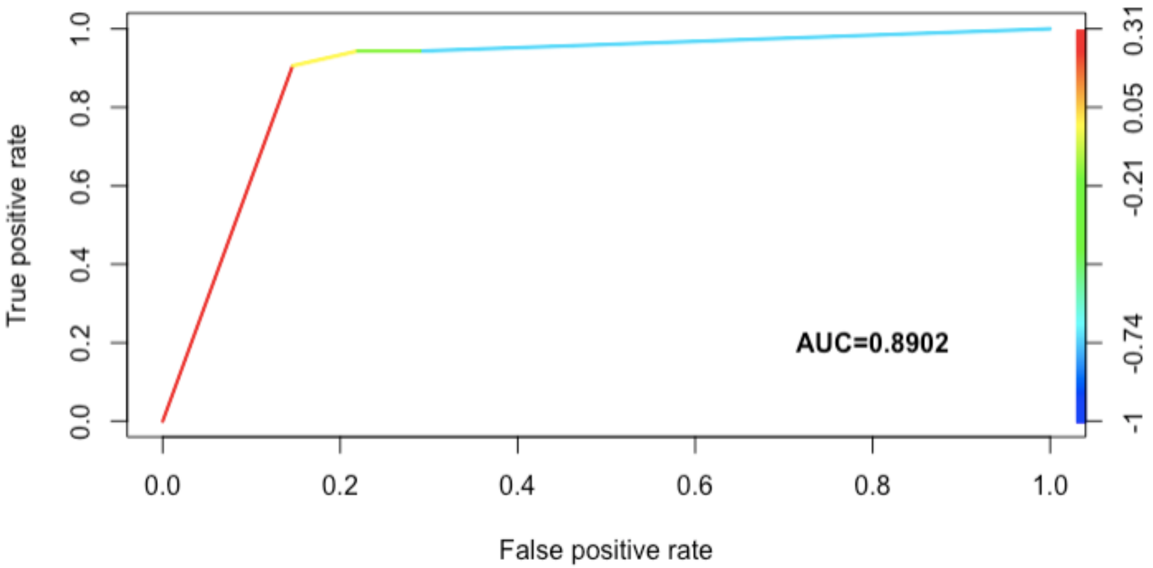
PCA dataset



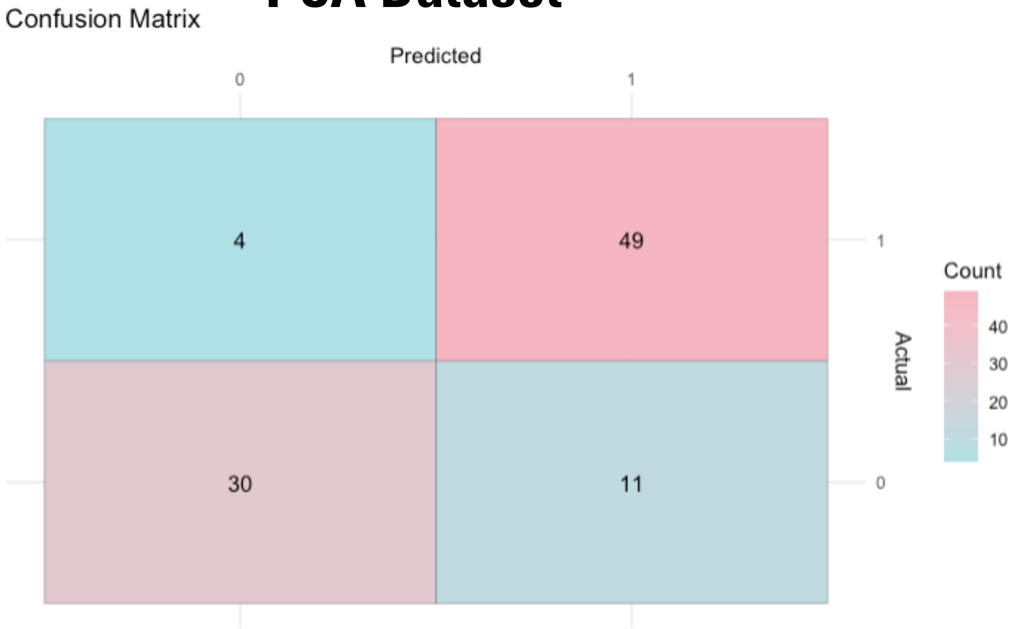
ORIGINAL Dataset



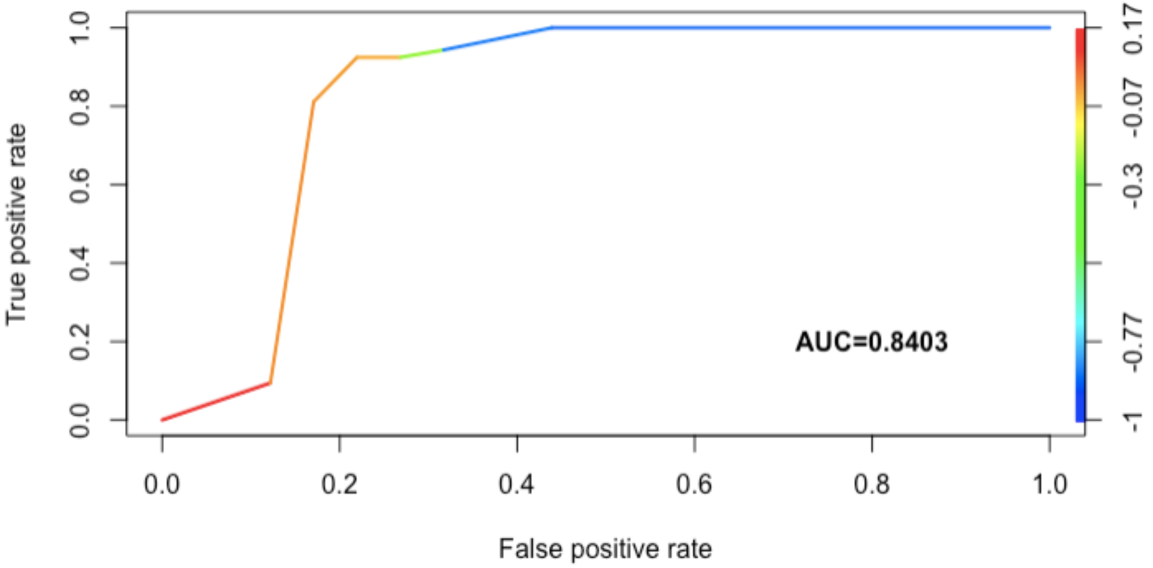
Accuracy: 87,23%



PCA Dataset



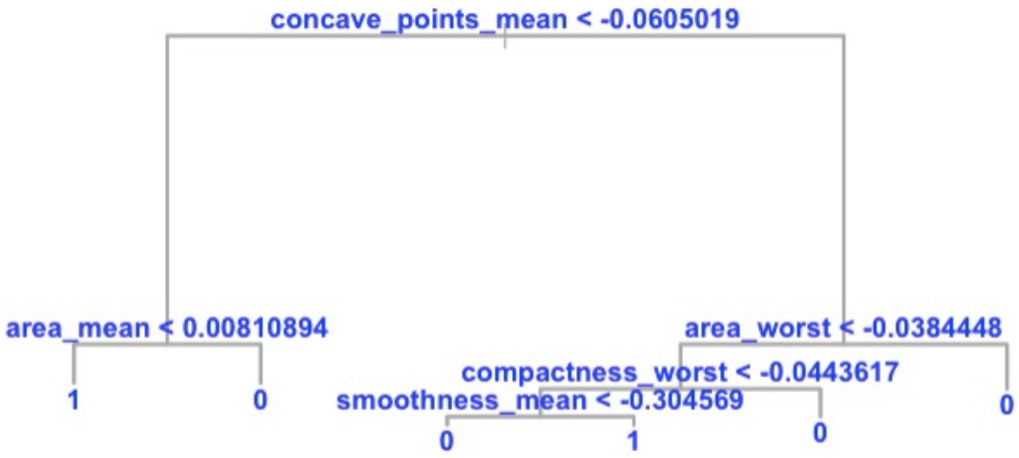
Accuracy: 84,04%



ORIGINAL dataset

Most important variables:

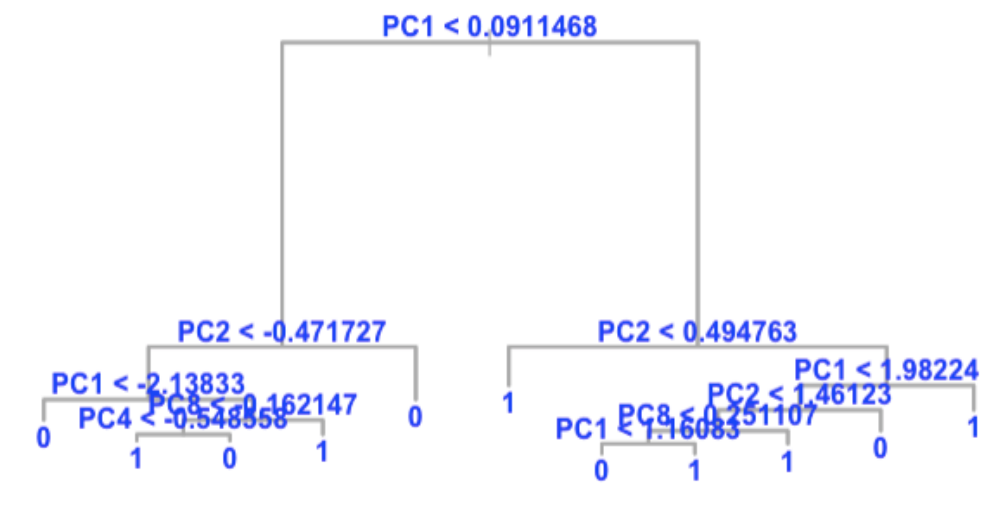
- concave_points_mean
- area_mean
- area_worst
- compactness_worst


concave_points_mean < -0.0605019
area_mean < 0.00810894
area_worst < -0.0384448
compactness_worst < -0.0443617
smoothness_mean < -0.304569
1
0
0
1
0
0

PCA dataset

Most important variables:

- PC1
- PC2
- PC8
- PC4



1 INTRODUCTION

2 EXPLORATORY DATA ANALYSIS

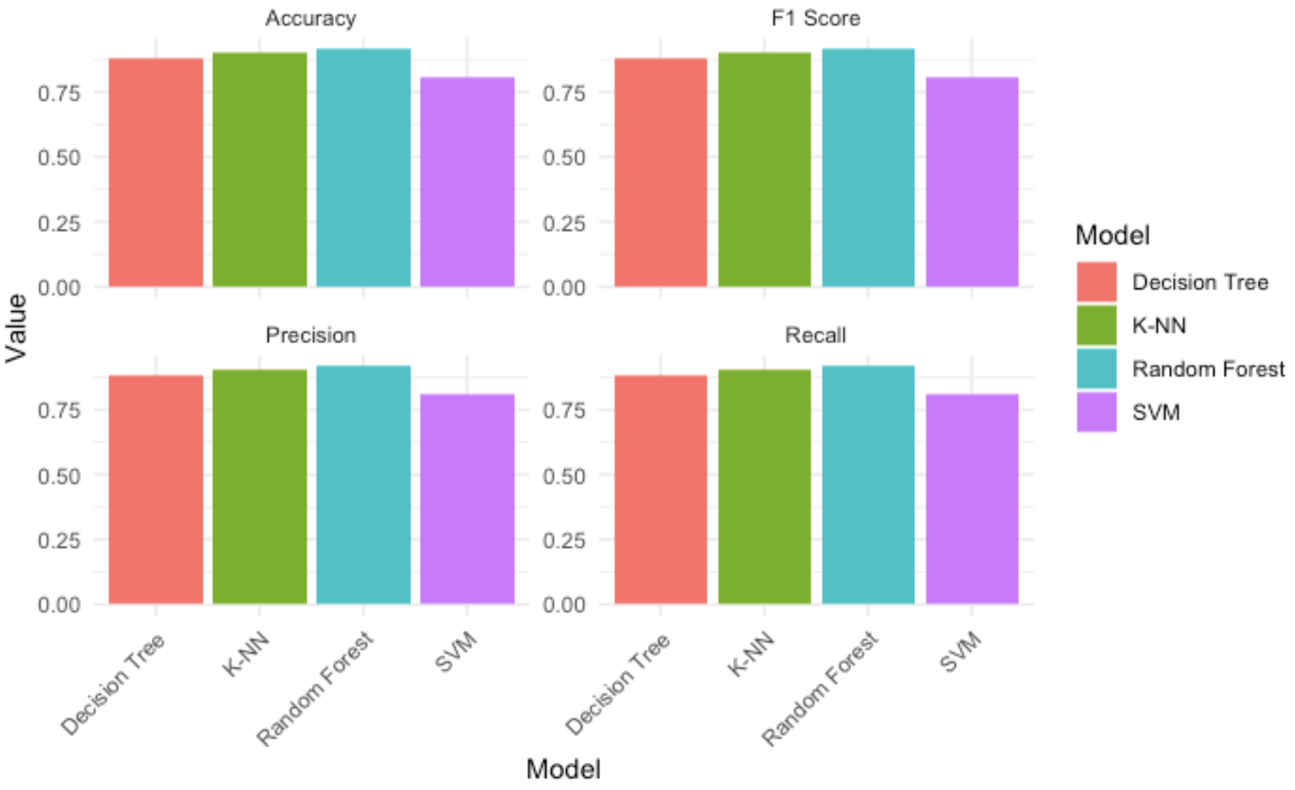
3 MODEL ANALYSIS

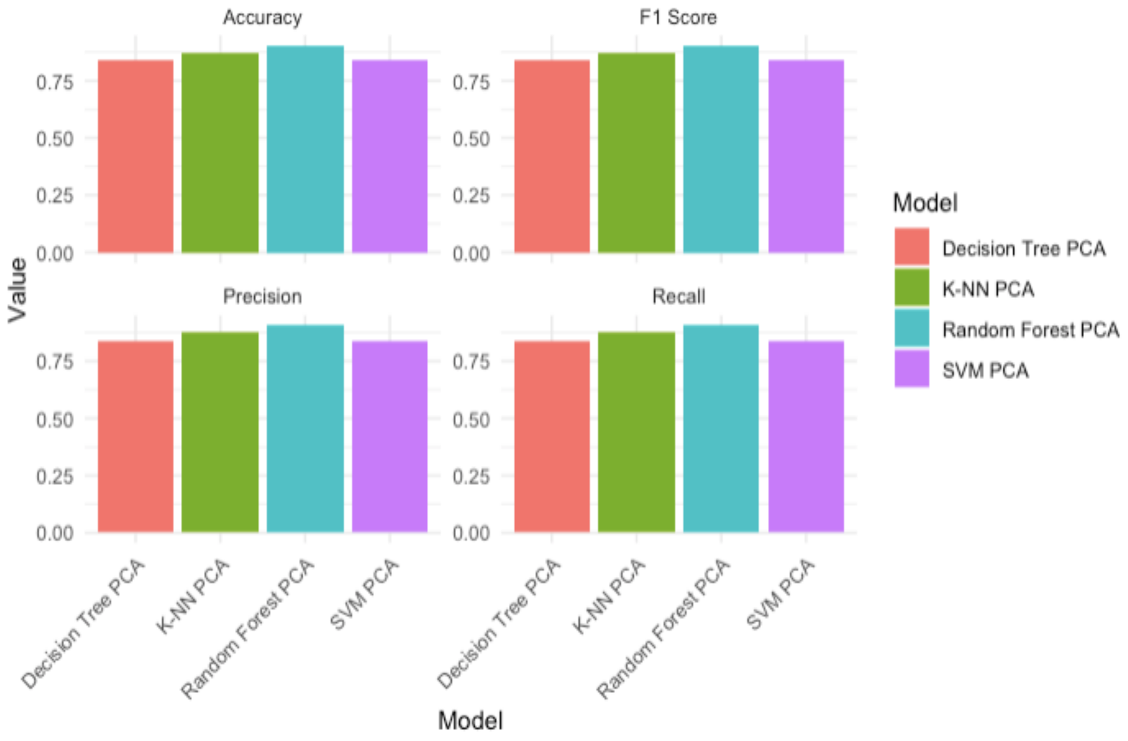
🔍 **EVALUATION AND RESULTS**

5 CONCLUSIONS

EVALUATION AND RESULTS

We can now compare and evaluate the results obtained before





EVALUATION AND RESULTS

- **Consideration of Dimensionality:**

While models trained on the original dataset exhibit slightly higher performance, they operate on a larger number of variables.

Models derived from PCA offer reduced dimensionality while maintaining high accuracy, making them attractive for practical implementation.

- **Practical Implications:**

In real-world applications, computational efficiency and interpretability are essential factors alongside accuracy.

PCA-derived models strike a balance between accuracy and efficiency, offering a practical solution for breast cancer diagnosis.

1 INTRODUCTION

2 EXPLORATORY DATA ANALYSIS

3 MODEL ANALYSIS

4 EVALUATION AND RESULTS

🔍 **CONCLUSIONS**

CONCLUSIONS

The Random Forest model demonstrates superior performance in the original dataset and even after applying the PCA highlighting its ability to handle data complexity and identify the most relevant features for breast cancer diagnosis.

Impact of Variables on Breast Cancer Diagnosis

1. **Area Worst:** area of the widest zone of a tumor in breast tissue.
A larger area might indicate the presence of a larger mass, potentially correlated with a more advanced tumor

2. **Concave Points Worst:** number and arrangement of the most pronounced concave points in the tumor's contour.
Greater presence of concave points might suggest a more irregular or jagged structure, indicating a higher likelihood of malignancy.

3. **Concave Points Mean:** reflects the average of concave points present in breast tissue.
A higher average of concave points could indicate greater morphological complexity of the tumor, correlated with a higher probability of cancer.

CONCLUSIONS

4. Concavity Mean: This is a measure of the average concavity index of the tumor's contour. A higher value might indicate greater deviation from regular shape, which is often associated with malignant tumors.

5. Concavity Mean: This is a measure of the average concavity index of the tumor's contour. A higher value might indicate greater deviation from regular shape, which is often associated with malignant tumors.

6. Area Mean: This parameter represents the average area of a tumor in breast tissue. A larger area could indicate the presence of a larger mass, which could be correlated with a more advanced tumor.

7. Radius Worst: This is the radius of the widest zone of a tumor in breast tissue. A larger radius might indicate a more extensive mass, which could be associated with a more invasive and advanced tumor.

CONCLUSIONS

Stakeholder Consideration:

- For researchers and developers, PCA-derived models offer streamlined algorithms that are easier to implement and interpret.
- Clinicians can benefit from PCA-derived models due to their ability to provide accurate diagnoses while requiring less computational resources.

For researchers, a model that offers high precision and sensitivity, along with clear interpretation of significant variables, would be preferable. In this context, models based on algorithms like Random Forest might be more suitable as they provide insights into variable importance and can be easily interpreted.

For clinicians, besides accuracy, interpretability and computational efficiency are paramount. Simpler and easier-to-interpret models such as k-Nearest Neighbors (KNN) or Support Vector Machine (SVM) could be preferable. These models can provide reliable predictions with more intuitive interpretation, enabling clinicians to make more informed treatment decisions more efficiently.

Thanks for your attention!