MAST5956: Big Data and Machine Learning

Machine Learning Approaches for Breast Cancer Classification and Analysis

Name: Sagari Muraliegaran

Contents

Introduction

One of the most common and fatal diseases that impacts women globally is breast cancer. Breast cancer is a disease where abnormal cells in the breast grow uncontrollably, forming a tumour (*Breast Cancer* n.d.). These tumours can be seen on an X-ray or felt as a lump. It usually starts in the ducts or lobules and can spread to other parts of the body if not detected early.

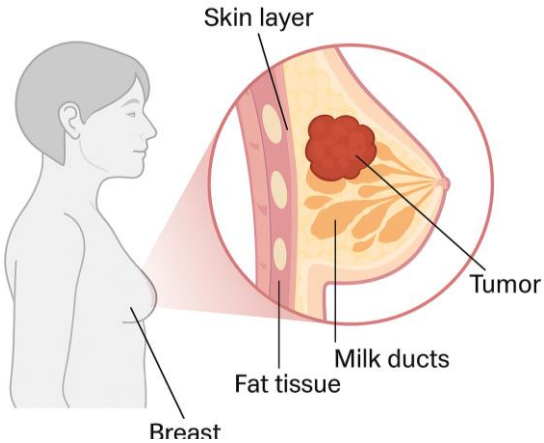


Figure 1 - Cross-sectional illustration of the breast showing the location of a tumour. Breast cancer typically begins in the ducts or lobules and can invade nearby tissues.

Early detection and precise diagnosis are essential to improve patient outcomes and lower death rates. Advances in image processing and machine learning have made it possible to identify breast tumours non-invasively using fine needle aspiration (FNA) samples, whereas before, diagnosis needed invasive surgical procedures like biopsies (*Fine Needle Aspiration (FNA) of the Breast* n.d.).

The Breast Cancer Wisconsin Diagnostic Dataset *(William Wolberg 1993)*, which comprises 569 samples obtained from digital pictures of FNA samples, is used in this paper to investigate the categorisation of breast tumours. 30 numerical parameters, including radius, roughness, smoothness, compactness and fractal dimension, are retrieved from cell nuclei in each sample. These characteristics include measures of the cell nuclei's size, shape and texture, all of which are linked to malignancy.

The dataset was first presented in a 2005 paper titled Nuclear Feature Extraction for Breast Tumour Diagnosis *(Nuclear Feature Extraction for Breast Tumour Diagnosis n.d.)*, which used image processing and a classifier based on linear programming to create a highly accurate diagnosis system. That study showed that a single separation plane could reach a 97% classification accuracy with only 3 features: mean texture, worst area and worst smoothness. This demonstrates how machine learning models may be used to spot trends in medical data and aid in clinical decision-making.

In this project, I aim to build and evaluate predictive models for tumour diagnosis using a combination of methods, which include:

- **Classification model** – such as decision trees, to predict whether a tumour is benign or malignant.
- **Principal Component Analysis (PCA)** – to reduce dimensionality and visualise patterns.
- **Clustering techniques** – to explore natural groupings within the dataset.

The aim is to evaluate these models effectively and contrast their results with the findings of the original study. By doing this, I want to obtain a better understanding of which characteristics are most crucial for diagnosis and how machine learning might help combat cancer.

Method

Dataset

The Breast Cancer Wisconsin Diagnostic Dataset, which includes 569 samples derived from digital images of FNA biopsies, is utilised in this study. 30 numerical characteristics (such as radius, perimeter, compactness, smoothness, and fractal dimension) are included in each sample to characterise the size, shape and texture of the cell nuclei. The diagnostic, which determines whether the tumour is benign (B) or malignant (M), is the target variable.

The dataset that is being analysed was originally in "Nuclear Feature Extraction for Breast Tumour Diagnosis". Using a limited sample of characteristics (mean texture, worst area and worst smoothness), image processing methods and a linear-programming classifier obtained a 97% accuracy.

Data Preprocessing

The dataset was imported into R Studio, and the actions for the preprocessing are:

- Removing the ID column as it is not a predictive feature.
- Changing the diagnostic variable into a binary factor (Benign = 0, Malignant = 1)
- To guarantee comparability across various measurement units, all numerical characteristics should be normalised using min-max scaling.
- Checking there aren't any missing values

Exploratory Data Analysis (EDA)

To fully understand the distribution and class balance, an EDA was carried out. To find trends and identify outliers, visualisations such as bar charts, boxplots and pairwise plots (using ggpairs) were employed. Additionally, correlation analysis was used to investigate feature correlations.

Feature Selection

Only 3 essential characteristics were included in an additional subset analysis to match the results of the original study article:

- Texture mean
- Area worst
- Smoothness worst

This made it possible to compare minimal-feature and full-feature models.

Classification Models - Decision Tree

- A decision tree model was trained using the rpart package
- Cross-validation was used to prune the tree and prevent overfitting.
- The model was visualised to interpret decision paths based on key features.

Principal Component Analysis (PCA)

PCA was used to visualise the feature space and reduce it dimensions. To investigate the natural division between classes, the first two main components were plotted and coloured according to diagnosis.

Clustering

K-means clustering was used to the normalised features in order to examine the data from an unsupervised standpoint. To evaluate alignment, cluster assignments were contrasted with the real diagnostic labels.

Results

Exploratory Data Analysis (EDA)

In this section, EDA is used to comprehend the distribution, linkages and structure of the breast cancer dataset. Identifying class imbalances, outliers, significant trends and possible predictive characteristics that distinguish benign from malignant tumours requires this stage.

Bar Chart

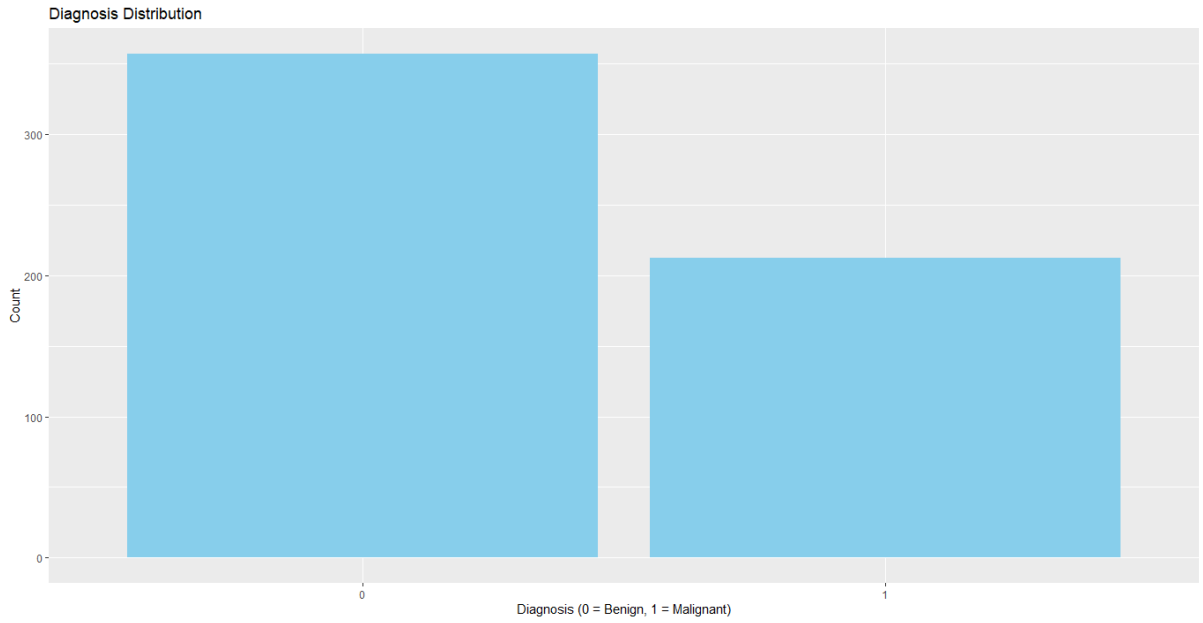


Figure 2 - Diagnosis Distribution Bar Chart. This figure shows the class distribution of tumour diagnoses in the dataset. There is a moderate class imbalance, with more benign instances (label 0) than malignant cases (label 1). (R code is shown in Appendix 2).

Boxplot

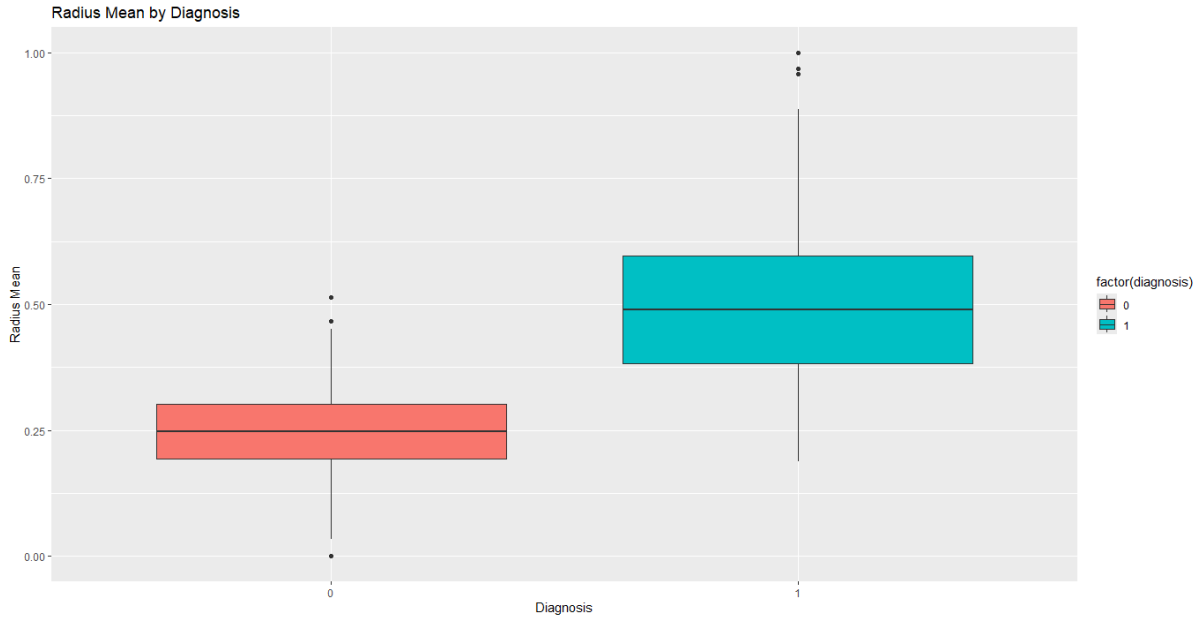


Figure 3 - Boxplot of Radius Mean by Diagnosis. This figure shows the radius mean values from benign and malignant tumours are contrasted in the boxplot. Radius values are often larger for malignant tumours, indicating that this trait might be a helpful classifier. (R code is shown in appendix 3).

Pairwise Plot

Figure 4 - Pairwise Plot of Selected Features. This pairwise plot compares radium mean, texture mean, and area mean. Strong positive correlation between radius mean and area mean is shown by density plots and correlation coefficients, particularly within the malignant group. (R code is shown in appendix 4)

Correlation Heatmap

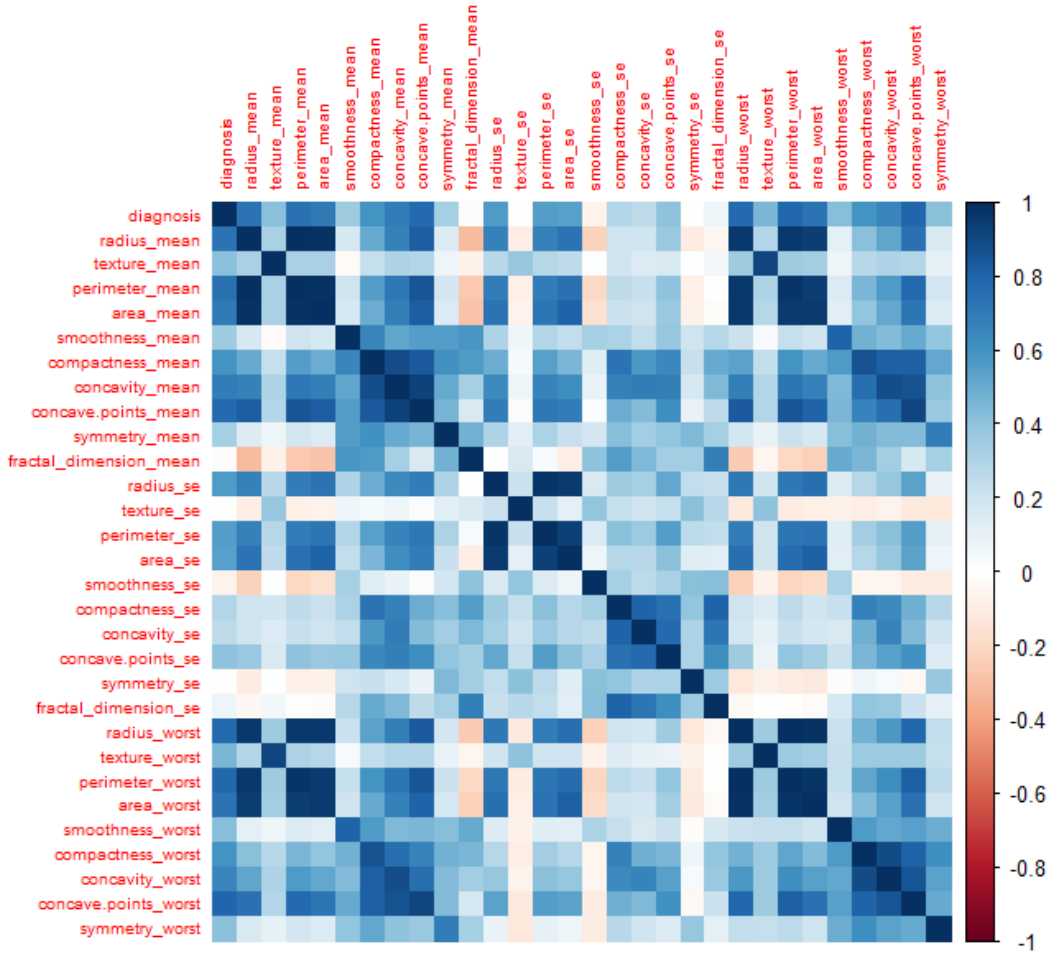


Figure 5 - Correlation Heatmap. The correlation heatmap visualises pairwise Pearson correlations between all features. Features relating to radius, perimeter and area show strong connections with one another, suggesting multicollinearity. Feature selection or dimensionality reduction can be guided by these correlations. (R code is shown in Appendix 5)

In the breast cancer dataset, the exploratory data analysis successfully identifies important trends which provided a solid basis for additional modelling. The dataset is rather unbalanced, with more benign instances than malignant ones, as seen in Figure 2. The radius mean has the potential to be a discriminative characteristic, since Figure 3 shows that it tends to be higher in malignant instances. This is further supported by the paired plots in Figure 4, which show that radius mean, area mean, and texture mean have substantial positive relationships, especially in malignant tumours. Furthermore, multicollinearity across a number of features – particularly those associated with tumour size and shape – was emphasised by the correlation heatmap in Figure 5. These results imply that although certain variables are predictive on their own, others are strongly interconnected, which supports the further application of dimensionality reduction methods such as PCA. All things considered, the EDA stage was essential for comprehending the distribution of the data, spotting significant characteristics and directing the creation of later classification and clustering models.

Classification Model – Decision Tree

A Decision Tree, which is a type of classification model, was put into place to determine if a tumour is benign or malignant. Decision trees are very interpretable models that create a tree-like structure of decision rules by dividing data according to feature values. Because of this, they are especially helpful in medical datasets where model decision transparency is crucial. Thirty numerical features that describe different tumour characteristics are included in the breast cancer dataset utilised in this experiment. The Decision Tree seeks to appropriately identify new situations by learning from these qualities. The whole dataset was used to train the model, and its structure was shown to show how important features influenced categorisation. The model's performance and accuracy in accurately predicting tumour kinds were assessed using a confusion matrix.

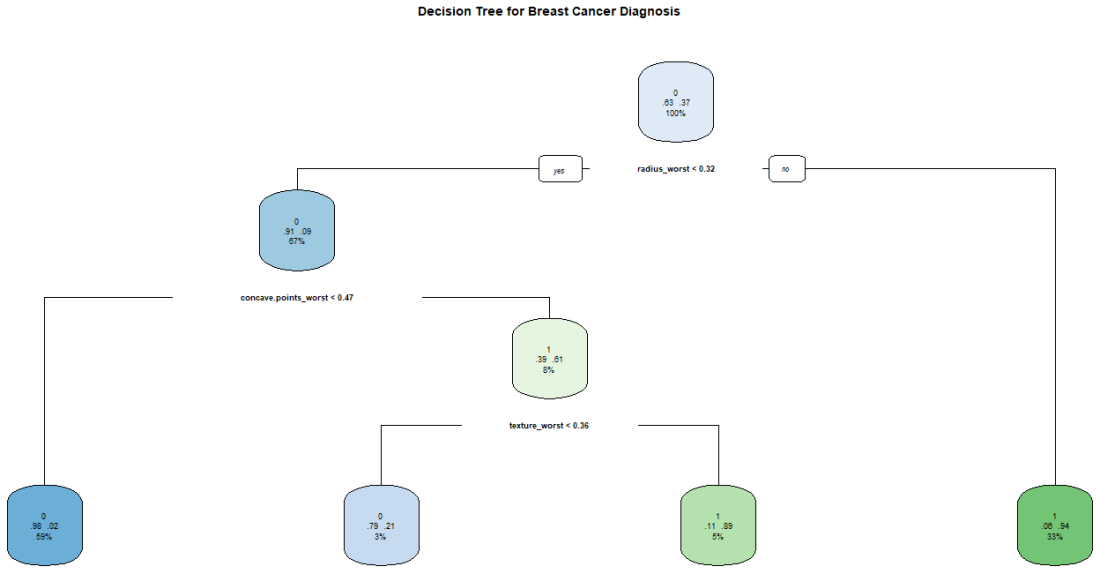


Figure 6 - Decision Tree for Breast Cancer Diagnosis. *This tree visualises the classification process for predicting whether a tumour is benign (0) or malignant (1), based on features from the breast cancer dataset.* (R code is shown in Appendix 6)

Using a few essential characteristics, the decision tree model successfully differentiates between benign and malignant tumours. As the root split, the radius worst is the most significant attribute. All samples in that branch are categorised as benign with 100% certainty if radius worst < 0.32. The model further divides depending on concave points, worst and texture worst when the radius worst is larger, reflecting more subtle differences in tumour structure.

Given that malignant tumours tend to have greater radii and more irregular forms (recorded by concave point worst, this structure is a good agreement with established clinical observations. Additionally, because of its relative shallowness, the tree is accurate and simple to read.

Overall, the model exhibits strong interpretability and predictive ability, providing a clear and simple categorisation technique appropriate for clinical decision support.

Table 1 - Confusion Matrix for the Decision Tree classification model. *This table displays the confusion matrix comparing the predicted diagnoses against the actual tumour classifications. The rows represent the actual classes (benign = 0, malignant = 1), while the columns represent the model's predictions.* (R code is shown in Appendix 7).

	Predicted Benign (0)	Predicted Malignant (1)
Actual Benign (0)	343	14
Actual Malignant (1)	9	203

Table 2 - Performance Metrics for Decision Tree Model. *This table summarises key classification performance metrics derived from the confusion matrix. These metrics that are presented in the table evaluate the model's reliability and effectiveness in correctly identifying malignant and benign tumours.* (R code is shown in Appendix 7).

Metric	Value	Interpretation
Accuracy	95.96 %	Overall correctness of the model
Sensitivity (Recall – Malignant)	95.75 %	Ability to correctly identify malignant cases
Specificity (Recall – Benign)	96.08 %	Ability to correctly identify benign cases
Precision (Positive Predictive Value)	93.55 %	Proportion of predicted malignant cases that were correct
Negative Predictive Value	97.44 %	Proportion of predicted benign cases that were correct
Balanced Accuracy	95.92 %	Average of sensitivity and specificity
Kappa	0.914	Average of sensitivity and specificity
P-value [Acc > NIR]	< 2e-16	Accuracy significantly better than random chance

Tables 1 and 2 demonstrate the exceptional outcomes attained by the Decision Tree classification model. The model showed good prediction performance with an overall accuracy of 95.96%. It is quite successful at differentiating between the two classifications, properly identifying 95.75% of malignant tumours (sensitivity) and 96.08% of benign tumours (specificity). While the negative predictive value of 97.44% indicates great reliability in detecting benign instances, the precision of 93.55% indicates that the majority of the malignant predictions were correct. Furthermore, even with a somewhat unbalanced sample, the balanced accuracy of 95.92% verifies consistent performance across both classes. The model's robustness is further supported by the Kappa statistics of 0.914, which indicates a high degree of agreement between predictions and actual results that goes beyond chance. All things considered, the Decision Tree offers a fair, comprehensible, and precise approach to categorising breast cancer diagnosis.

Principal Component Analysis (PCA)

To find patterns in the breast cancer dataset and lower its dimensionality, PCA was used. Class separation may be visualised, and important components that explain most of the variance across characteristics can be found by PCA. For high-dimensional datasets with 30 numerical variables such as this one, this is really helpful. PCA helps discover the variables that most contribute to the variance between benign and malignant tumours by breaking down the data into a smaller collection of uncorrelated principal components, which allows for easier visual investigation.

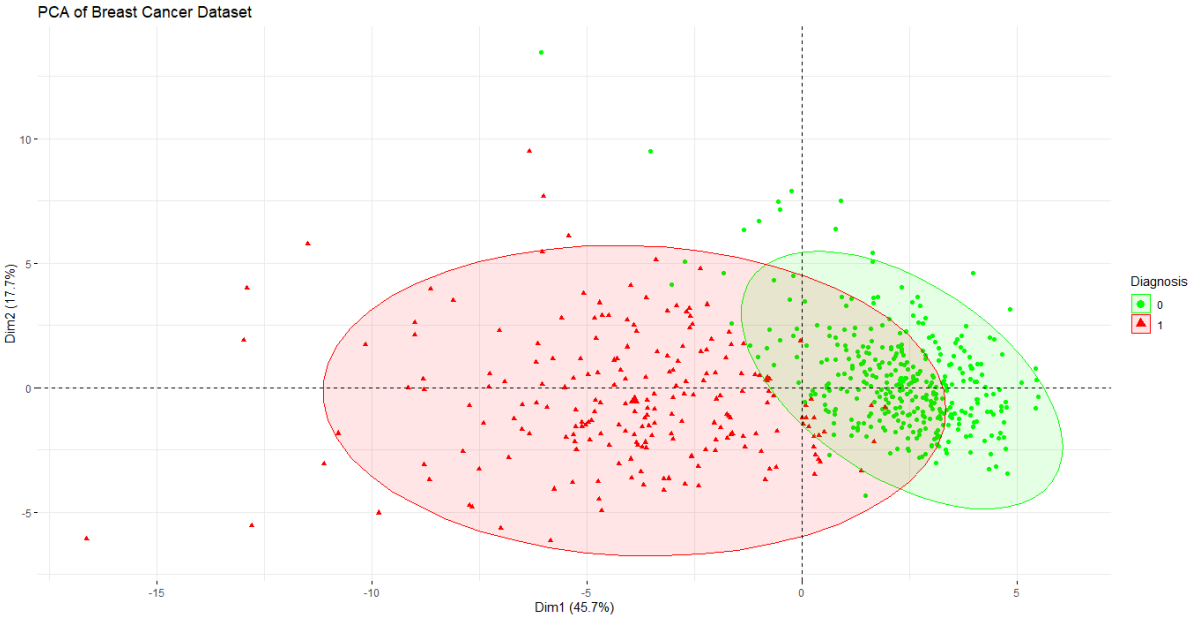


Figure 7a - PCA Plot of Breast Cancer Dataset. Dim1 and Dim2, the first two main components displayed in this PCA plot, account for 45.7% and 17.7% of the total variance, respectively. With some overlap, the graphic clearly distinguishes between benign (green circles) and malignant (red triangles) instances. (R code is shown in Appendix 8)

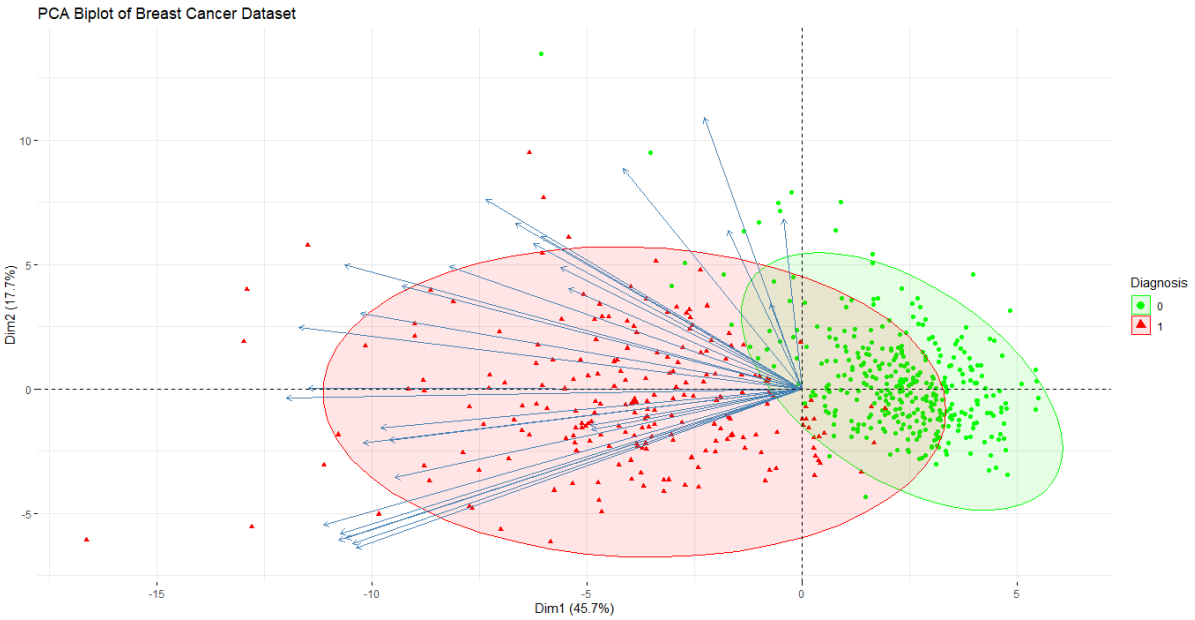


Figure 7b - PCA Biplot with Feature Loadings. This biplot extends the PCA plot by overlaying arrows representing the contribution of the original features to the first two principal components. In accordance with established clinical indications of malignancy, features like radius mean, area mean,

and concave point worst seem to have a significant impact along the Dim1 axis. (R code is shown in Appendix 8)

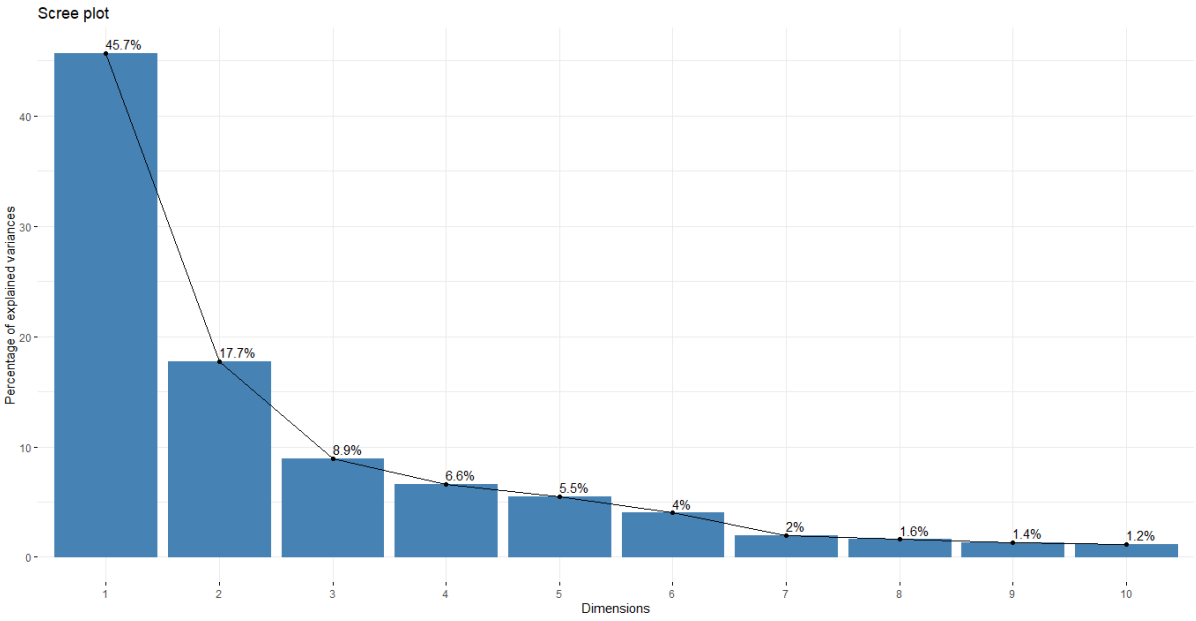


Figure 8 - Sceen Plot of Explained Variance. The percentage of variation accounted for by each of the top 10 main components is displayed in this bar chart. (R code is shown in Appendix 8).

Strong evidence that the dataset has significant structure that distinguishes between benign and malignant tumours is provided by the PCA visualisations. Dim1 and Dim2 (PC1 and PC2, respectively) provide distinct class separability and capture a significant amount of the variation (45.7% and 17.7%), as seen in Figure 7a. The majority of benign and malignant cases cluster in different areas of the plot, despite some class overlap, demonstrating the value of PCA for first categorisation.

The biplot in figure 7b, which shows which traits are mainly responsible for the separation, improves interpretability. Strong projections along Dim1 are shown for features like radius mean, area mean, and concave point worst, indicating their significance in class differentiation. This is consistent with the clinical knowledge that these characteristics are strongly linked to cancer.

Lastly, the scree plot in Figure 8 demonstrates that the majority of the variation in the dataset is contained in the first few components, especially the first two, which together account for 63% of the variance. This demonstrates that dimensionality reduction and visualisation using Dim1 and Dim2 maybe accomplished without sacrificing a lot of information.

Overall, the PCA analysis not lonely validates the existence of significant structure in the data but also directs feature selection and shows that class separation with a smaller number of components is feasible.

Clustering

The K-means algorithm, an unsupervised learning method that classifies data according to similarity, was used on the breast cancer dataset. Finding out if the data's natural groups match the actual diagnosis labels (benign vs malignant) was the aim. Clustering aids in assessing the intrinsic separability of data points and locating possible structure unsupervised since the dataset lacks established cluster labels for training.

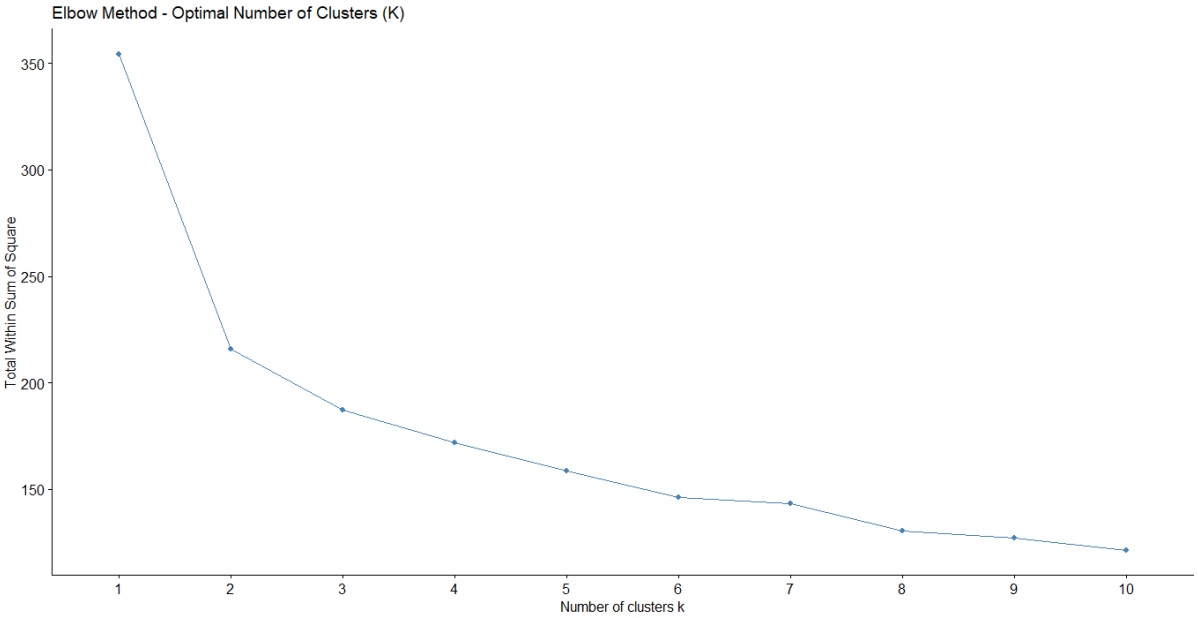


Figure 9 - Elbow Method for Optimal K. This line plot shows the total within-cluster sum of squares (WSS) for different values of K. The "elbow" at K = 2 indicates the optimal number of clusters, suggesting the dataset naturally splits into two groups. (R code is shown in Appendix 9)

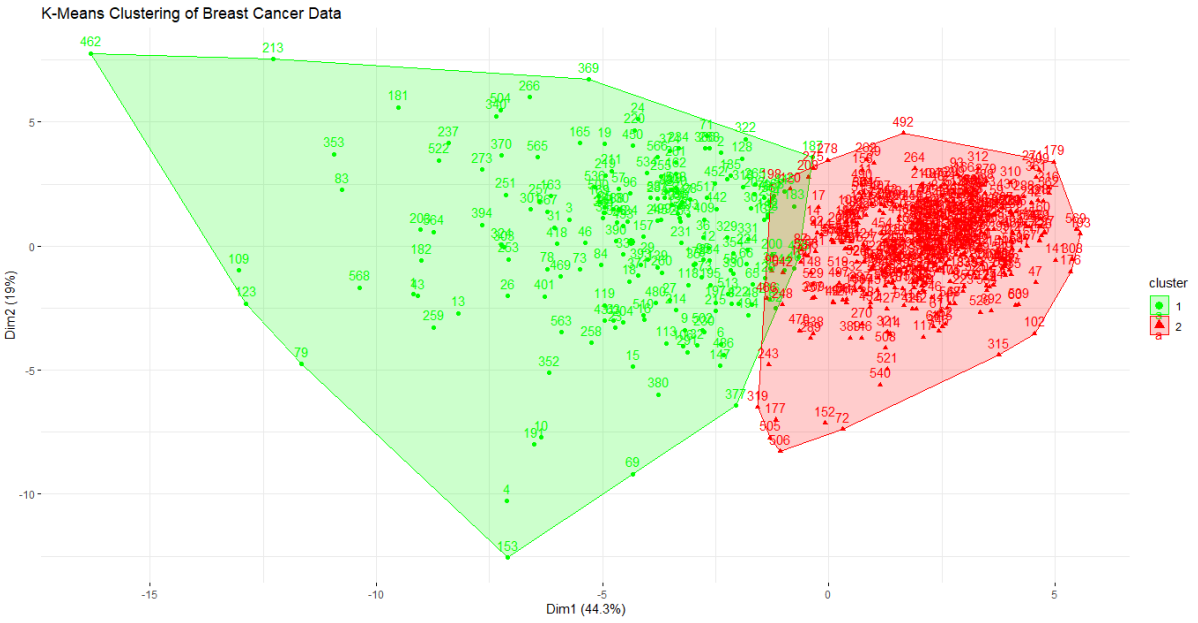


Figure 10- K-Means Clustering Visualisation. With K = 2, the PCA reduced figure shows the outcome of K-means clustering. The red and green coloured area indicated that two groups into which the

data points are grouped. Despite clustering being unsupervised, there is a significant visual alignment between the two groups and the underlying diagnosis. (R code is shown in Appendix 9).

Table 3 - Contingency Table Comparing Clusters with Diagnosis. The K-means cluster assignments and the actual diagnostic labels are contrasted in this table. Although there is considerable misclassification, Cluster 2 encompasses the bulk of benign cases (348), whereas Cluster 1 captures the majority of malignant cases (180). (R code is shown in Appendix 9)

Cluster	Benign (0)	Malignant (1)	Total
1	9	180	189
2	348	32	380
Total	357	212	569

The K-means algorithm's clustering analysis of the breast cancer dataset showed underlying structure. The Elbow Method indicates that K = 2 is the ideal number of clusters, as seen in Figure 9, which is consistent with the data's binary classification (benign vs. malignant). The WSS curve's significant bend indicates the data's underlying division into two major groups.

Using PCA-reduced dimensions, the clustering result is displayed, demonstrating how the data points were divided into two clusters. Despite not having access to the actual diagnostic labels during clustering, the coloured regions – red for Cluster 2 and green for Cluster 1 – showcase a significant distinction and have a strong visual alignment to them.

Table 3 contrasts the fluster allocations with the actual diagnosis. The majority of malignant cases (180) are correctly identified by Cluster 1, whereas the majority of benign cases (348) are included in Cluster 2, despite some misclassification. This demonstrates that, even though K-means clustering is unsupervised, it may find significant groups in the data and could be helpful for pre-classification or exploratory analysis when labels are unavailable.

Discussion

To assess the structure and predictability of the breast cancer dataset, the investigation used a number of machine learning techniques, including K-means clustering, principal component analysis (PCA), and decision tree classification.

With an accuracy of 95.96%, sensitivity of 95.75% and specificity of 96.08%, the decision tree model fared especially well. These results show that both benign and malignant tumours may be accurately identified by the model. The high Kappa value (0.914) confirms that the model's predictions are not the result of chance but rather closely match actual results. Furthermore, the model's interpretability and resilience were emphasised by its shallow structure and dependence on a few number of crucial variables, most notably radius mean, area mean, and concave points worst. Their biological and statistical relevance was further supported by the fact that these variables were often chosen at the top splits in the decision tree and showed up as dominating loadings in PCA.

The PCA gave important information on data structure and feature redundancy. Effective class separation visualisation in a lower-dimensional space was made possible by the first two main components (Dim1 and Dim 2), which accounted for 63.4% of the total variance. Features such as

radius mean and area mean showed substantial directional loading along the first principal component, according to the biplot, indicating that they are the predictive power of these characteristics was further reinforced by the distinct, if somewhat overlapping, division between benign and malignant patients.

Despite being unsupervised, the K-means clustering method did a respectable job of matching actual diagnoses. The contingency table (Table 3) shows that Cluster 2 has the majority of benign cases (348) with a few malignant instances (32) that were incorrectly diagnosed, whereas Cluster 1 had 180 malignant and only 9 benign cases. The clustering was able to reveal important structure in the data even when no prior knowledge of diagnostic labels was present. However, when used without prior labelling or supervision, K-means showed more misclassification than the decision tree, showing its limitations.

The decision tree was meticulously adjusted to prevent undue depth in terms of model complexity and overfitting, resulting in an accurate and comprehensible model. Although improper reduction can cause decision trees to overfit, the model used in this study successfully struck a compromise between simplicity and performance. On the other hand, because clustering is unsupervised, it does not overfit in the same manner; nonetheless, it may be affected by inadequate initialisation or incorrect assumptions (such as spherical clusters).

In conclusion, PCA facilitated feature significance analyses and dimensionality reduction, the decision tree was the most accurate and comprehensible model, and clustering brought out the dataset's natural structure even in the face of substantial misclassification. When combined, these methods offered a thorough comprehension of the breast cancer dataset and validated the benefits of integrating supervised and unsupervised methods in biomedical data analysis.

Conclusion

The results of the published paper on nuclear feature extraction in breast cancer are confirmed by this study. Radius mean, area mean, and concave point worst were shown to be important characteristics for differentiating between benign and malignant tumours in both the analysis and the paper. These factors were crucial to the decision tree model's 95.96% accuracy, confirming its clinical applicability.

By demonstrating distinct class separation and a significant variance contribution from these variables, PCA further validated their significance. Furthermore, even though K-means clustering was unsupervised, it produced groups that matched diagnostic labels.

Overall, this work showed that nuclear morphological characteristics may successfully enable accurate and interpretable breast cancer categorisation, validating the article's finding through real-world machine learning applications.

Reference

Breast Cancer [Online]. Available at: https://www.cancerresearchuk.org/about-cancer/breast-cancer [Accessed: 3 April 2025].

Fine Needle Aspiration (FNA) of the Breast [Online]. Available at: https://www.cancer.org/cancer/types/breast-cancer/screening-tests-and-early-detection/breast-biopsy/fine-needle-aspiration-biopsy-of-the-breast.html [Accessed: 3 April 2025].

Nuclear Feature Extraction for Breast Tumour Diagnosis [Online]. Available at: https://minds.wisconsin.edu/bitstream/handle/1793/59692/TR1131.pdf;jsessionid=07B0FDD0450F3A59F7102361F10DCD41?sequence=1 [Accessed: 31 March 2025].

William Wolberg, O.M. (1993). Breast Cancer Wisconsin (Diagnostic). [Online]. Available at: https://archive.ics.uci.edu/dataset/17 [Accessed: 31 March 2025].

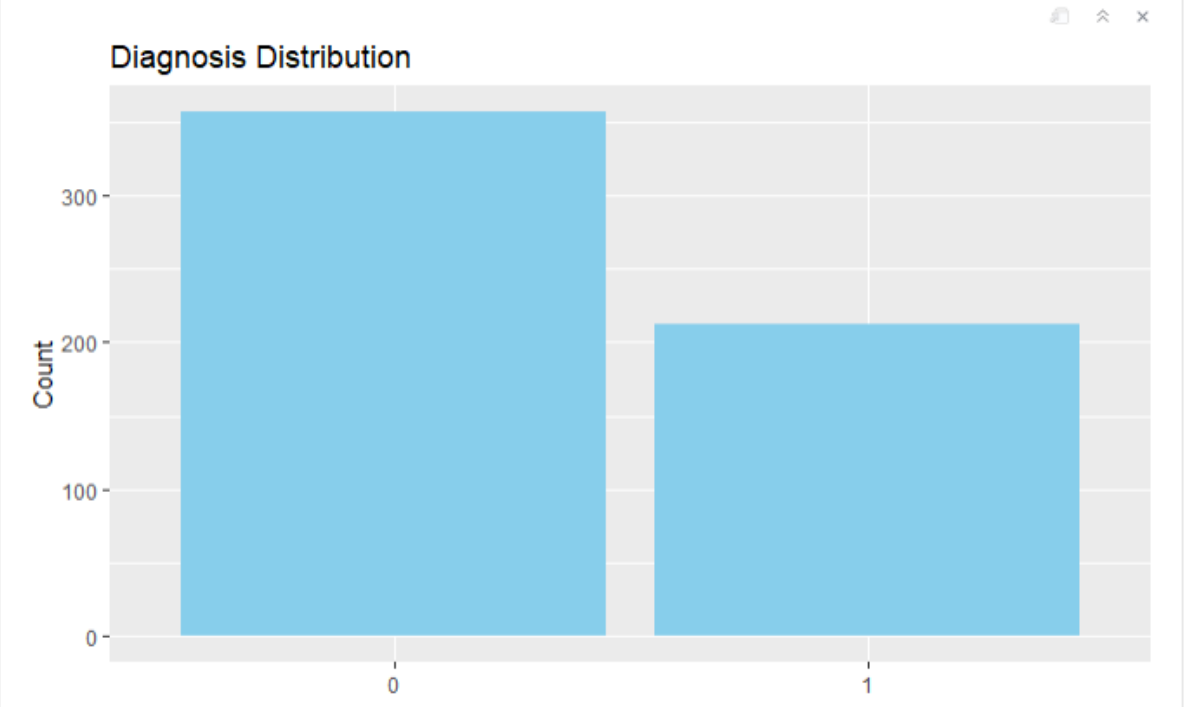
Appendix

Appendix 1 -Load Required Packages and Organising the dataset.

```{r}
# Load Required Packages
install.packages(c("readxl", "tidyverse", "GGally", "caret"))
library(readxl)
library(tidyverse)
library(GGally)
library(caret)

# Load Dataset
cancer1 <- read_excel("Data Analytics/Big Data and Machine Learning/CW/cancer1.xlsx")

# Remove ID column
cance1 <- cancer1 %>% select(-id)

# Convert diagnosis to binary factor (B = 0, M = 1)
cancer1$diagnosis <- as.numeric(cancer1$diagnosis == "M")

# Check for missing values
colSums(is.na(cancer1))  # Should all be 0

# Normalize numerical features (excluding diagnosis)
preproc <- preProcess(cancer1[, -1], method = "range")
cancer_scaled <- predict(preproc, cancer1[, -1])

# Add diagnosis back
cancer_scaled$diagnosis <- cancer1$diagnosis

# Clean column names to avoid issues
names(cancer_scaled) <- make.names(names(cancer_scaled))
```

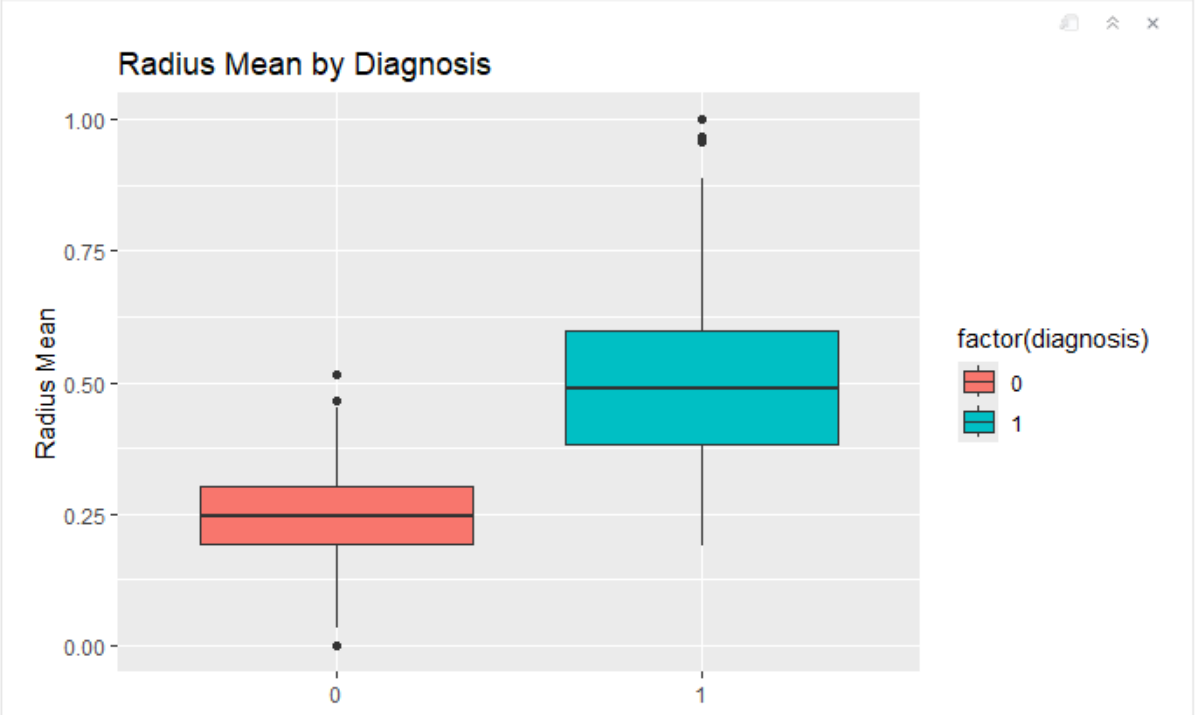
Appendix 2 – EDA (Bar chart)

```r
{r}
# Bar chart of class distribution
  ggplot(cancer_scaled, aes(factor(diagnosis))) +
  geom_bar(fill = "skyblue") +
  labs(title = "Diagnosis Distribution", x = "Diagnosis (0 = Benign, 1 = Malignant
)", y = "Count")
```



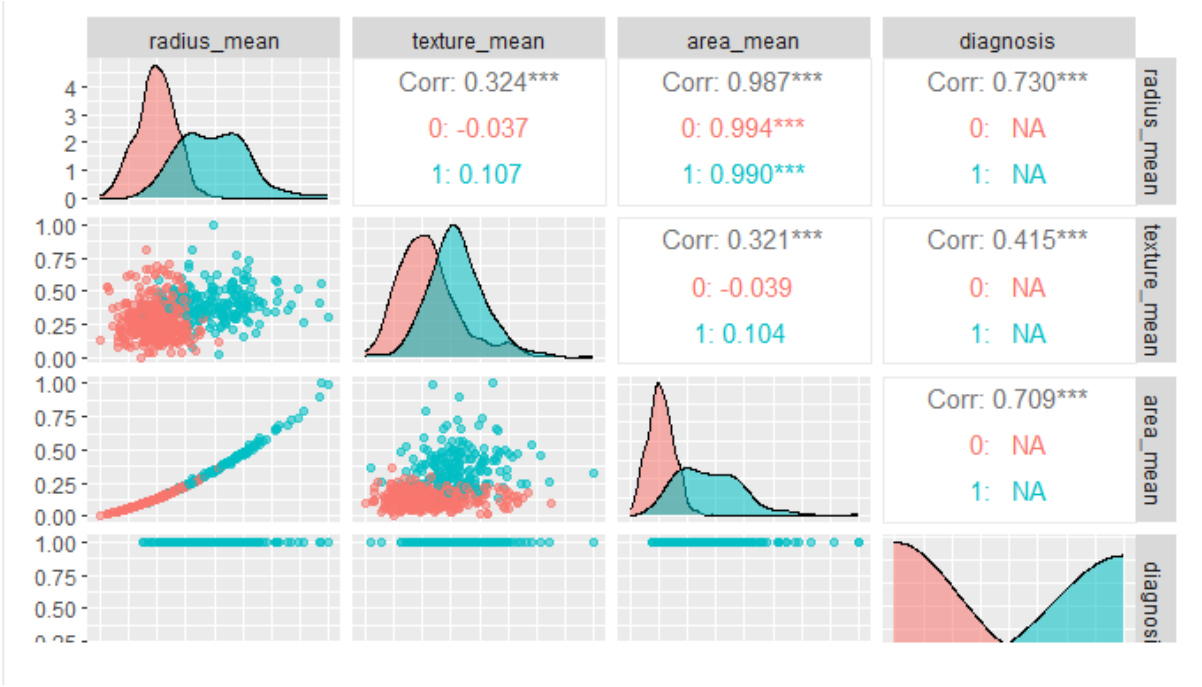
Appendix 3 – EDA (Box plot)

```r
{r}
# Boxplot of one feature by diagnosis (example: radius_mean)
ggplot(cancer_scaled, aes(x = factor(diagnosis), y = radius_mean, fill = factor
(diagnosis))) +
  geom_boxplot() +
  labs(title = "Radius Mean by Diagnosis", x = "Diagnosis", y = "Radius Mean")
```



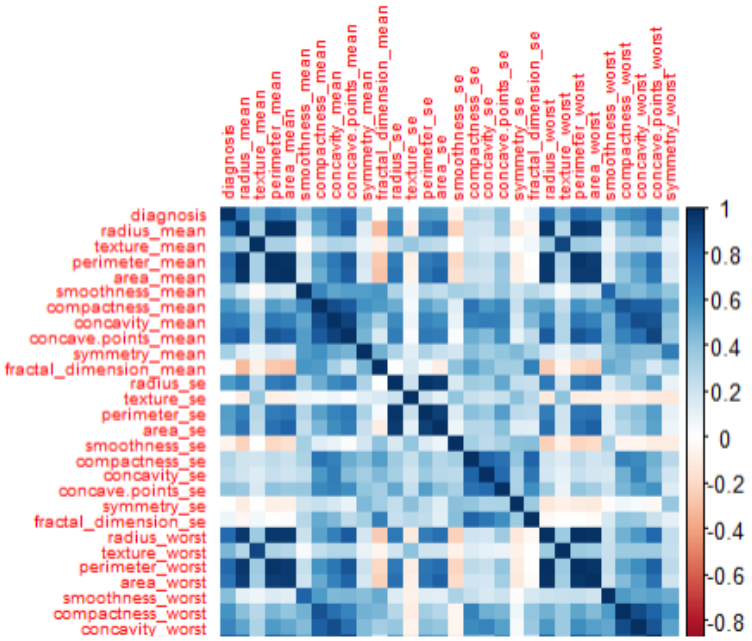
Appendix 4 – EDA (Pairwise Plot)

```r
{r}
# Pairwise scatterplots (only a few features to keep it readable)
ggpairs(cancer_scaled[, c("radius_mean", "texture_mean", "area_mean", "diagnosis")]
,
        aes(color = factor(diagnosis), alpha = 0.5))
```



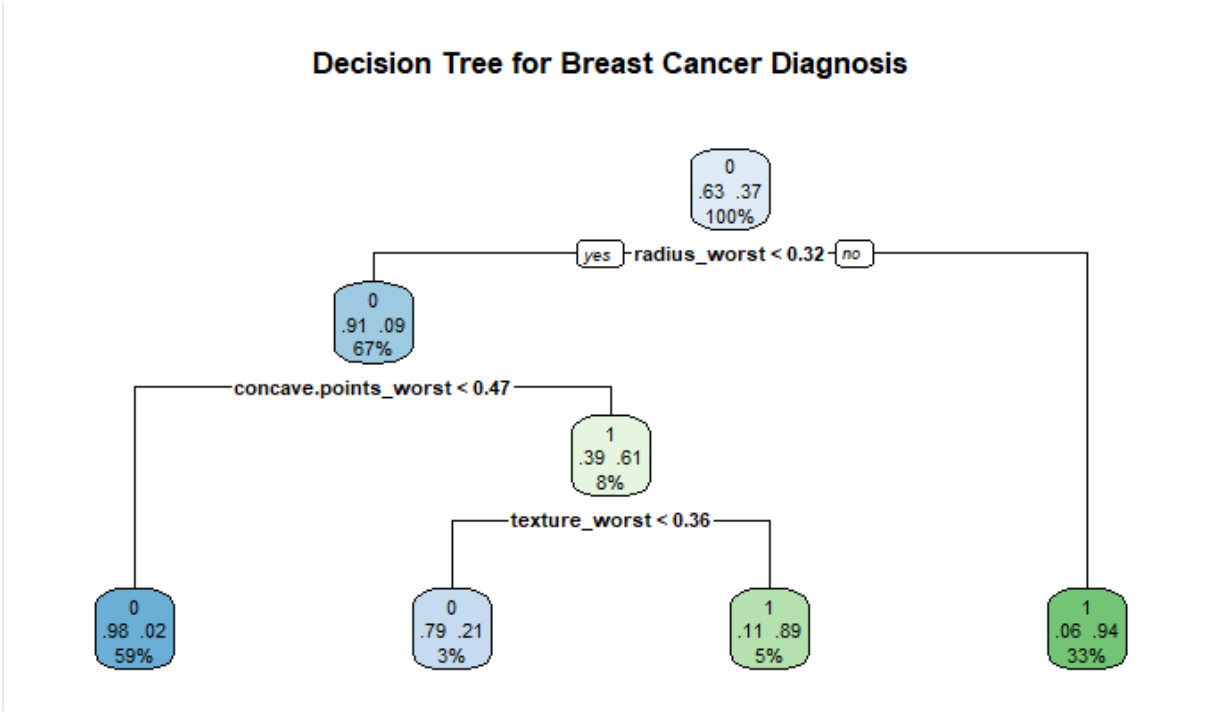
Appendix 5 – EDA (Correlation Heatmap)

```r
# Correlation heatmap
cor_matrix <- cor(cancer_scaled[, -ncol(cancer_scaled)])
corrplot::corrplot(cor_matrix, method = "color", tl.cex = 0.6)
```



Appendix 6 – Classification Model – Decision Tree

```r
{r}
install.packages("rpart")
install.packages("rpart.plot")
install.packages("caret")

library(rpart)
library(rpart.plot)
library(caret)

# Prepare the Data (Make sure diagnosis is a factor for classification)
cancer_tree <- cancer_scaled
cancer_tree$diagnosis <- as.factor(cancer_tree$diagnosis)

# Train the Decision Tree Model
set.seed(123)
tree_model <- rpart(diagnosis ~ ., data = cancer_tree, method = "class")


# Visualize the Tree
rpart.plot(tree_model,
           extra = 104,        # Show probabilities and percentages
           type = 2,           # Draw split labels below branches
           main = "Decision Tree for Breast Cancer Diagnosis")
```



Appendix 7 – Confusion Matrix and Statistics

This r code shows Table 1 and Table 2.

```
# Make Predictions and Evaluate
tree_pred <- predict(tree_model, cancer_tree, type = "class")

# Confusion Matrix
confusionMatrix(tree_pred, cancer_tree$diagnosis, positive = "1")
```

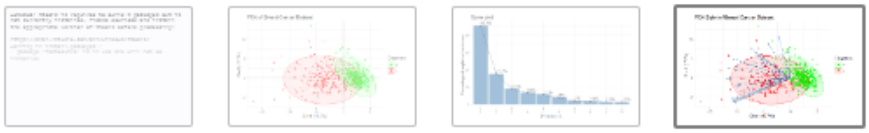
```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 343   9
         1  14 203

               Accuracy : 0.9596
                 95% CI : (0.94, 0.9742)
    No Information Rate : 0.6274
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.914

 Mcnemar's Test P-Value : 0.4042

            Sensitivity : 0.9575
            Specificity : 0.9608
         Pos Pred Value : 0.9355
         Neg Pred Value : 0.9744
             Prevalence : 0.3726
         Detection Rate : 0.3568
   Detection Prevalence : 0.3814
      Balanced Accuracy : 0.9592

       'Positive' Class : 1
```

Appendix 8 – PCA

This r code shows for Figures 7a, 7b and 8.

```r
{r}
# Install & Load Required Packages

install.packages("factoextra")    # For clean PCA visuals
library(factoextra)

#  Clean Column Names (if needed)

names(cancer_scaled) <- make.names(names(cancer_scaled))

# Run PCA (excluding target column)

pca_result <- prcomp(cancer_scaled[, -ncol(cancer_scaled)], scale. = TRUE)

# Visualize PCA (with diagnosis as color)
fviz_pca_ind(pca_result,
             geom.ind = "point",
             col.ind = as.factor(cancer_scaled$diagnosis),  # Color by diagnosis
             palette = c("green", "red"),
             addEllipses = TRUE,
             legend.title = "Diagnosis") +
  labs(title = "PCA of Breast Cancer Dataset")

fviz_eig(pca_result, addlabels = TRUE)
```

```r
fviz_pca_biplot(pca_result,
                label = "none",                          # 🔷 Removes point labels
                col.var = "steelblue",                   # Variable arrows in blue
                col.ind = as.factor(cancer_scaled$diagnosis),  # Color by diagnosis
                palette = c("green", "red"),
                addEllipses = TRUE,
                legend.title = "Diagnosis") +
  labs(title = "PCA Biplot of Breast Cancer Dataset")
```



Appendix 9 – Clustering

This R code shows Table 3 and figures 9 and 10.

```r
{r}
# Install & Load Required Packages

install.packages("factoextra")    # For clean PCA visuals
library(factoextra)

#  Clean Column Names (if needed)

names(cancer_scaled) <- make.names(names(cancer_scaled))

# Run PCA (excluding target column)

pca_result <- prcomp(cancer_scaled[, -ncol(cancer_scaled)], scale. = TRUE)

# Visualize PCA (with diagnosis as color)
fviz_pca_ind(pca_result,
             geom.ind = "point",
             col.ind = as.factor(cancer_scaled$diagnosis),  # Color by diagnosis
             palette = c("green", "red"),
             addEllipses = TRUE,
             legend.title = "Diagnosis") +
  labs(title = "PCA of Breast Cancer Dataset")
# Visualize the Clusters
fviz_cluster(kmeans_result, data = cancer_clustering,
             palette = c("green", "red"),
             ellipse.type = "convex",
             ggtheme = theme_minimal(),
             main = "K-Means Clustering of Breast Cancer Data")

# Compare Clusters to Actual Diagnosis
table(Cluster = kmeans_result$cluster,
      Diagnosis = cancer_scaled$diagnosis)
```