NEU COE INFO 6105 Final Project Report

Geetika Barla
002372565

As mentioned in my project proposal I am working on Breast Cancer Wisconsin (Diagnostic) data set in the UCI Machine Learning Repository (https://archive.ics.uci.edu/dataset/17/breast+cancer+wisconsin+diagnostic). I began with analysing the dataset.
It contains 569 records of breast tumor samples, each described by 30 numeric measurements taken from images of cells collected using a fine needle. These features represent characteristics such as radius, texture, perimeter, area, smoothness, and more, computed for the mean, standard error, and worst (largest) measurements of each cell nucleus.
Each instance is labeled as either benign (B) or malignant (M) under the Diagnosis column, which serves as the target variable for classification. This dataset is widely used in medical machine learning research due to its real-world relevance and the diversity of features it provides for predicting cancer diagnosis.
The dataset has:
V1: ID (numeric id, could be used to identify the diagnosis id)
V2: Diagnosis label (M = Malignant, B = Benign) → Target Variable
V3 to V32: 30 numeric features describing characteristics of cell nuclei
Since the dataset does not have proper headers, we name the column headers based on the description given .
The questions that I have addresses are:

1. Which of the classification method - logistic regression, LDA, or decision trees - is best suited to predict breast cancer diagnosis?

CODE:

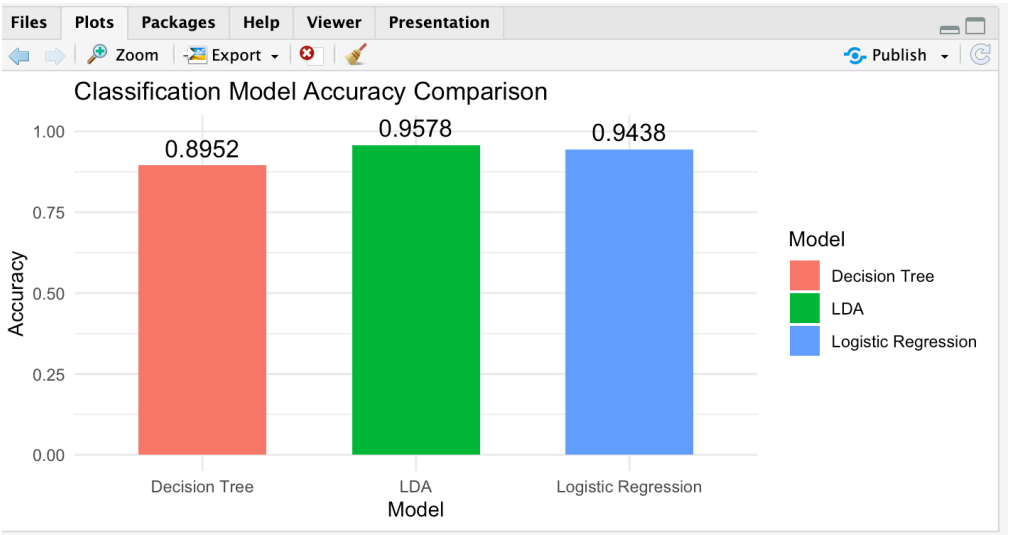
```
#  Load Required Libraries
library(MASS)     # For Linear Discriminant Analysis (LDA)
library(rpart)    # For Decision Trees
install.packages("caret")
library(caret)# For cross-validation and model comparison


# Load and preprocess the dataset

# Load dataset (update the working directory first)
data <- read.csv("wdbc.data", header = FALSE)

#Inspect the dataset
head(data) #View first few rows
dim(data) #Check number of rows and columns
str(data) #Get Column  data types
sum(is.na(data)) # To check missing values

# Assign proper column names based on UCI documentation
colnames(data) <- c("ID", "Diagnosis",
                    "radius_mean", "texture_mean", "perimeter_mean",
"area_mean", "smoothness_mean",
                    "compactness_mean", "concavity_mean", "concave_points_mean",
"symmetry_mean", "fractal_dimension_mean",
                    "radius_se", "texture_se", "perimeter_se", "area_se",
"smoothness_se",
                    "compactness_se", "concavity_se", "concave_points_se",
"symmetry_se", "fractal_dimension_se",
                    "radius_worst", "texture_worst", "perimeter_worst",
"area_worst", "smoothness_worst",
```

```
                    "compactness_worst", "concavity_worst",
"concave_points_worst", "symmetry_worst", "fractal_dimension_worst")
colnames(data)

#Convert Diagnosis to a categorical variable (factor)
data$Diagnosis <- as.factor(data$Diagnosis)


# Cross-validation Setup

set.seed(123) # Ensure reproducible results

# Create 5-fold cross-validation settings
ctrl <- trainControl(method = "cv", number = 5)


# Train the models

# Logistic Regression
log_model <- train(Diagnosis ~ ., data = data, method = "glm", family =
"binomial", trControl = ctrl)

# Linear Discriminant Analysis (LDA)
lda_model <- train(Diagnosis ~ ., data = data, method = "lda", trControl = ctrl)

# Decision Tree
tree_model <- train(Diagnosis ~ ., data = data, method = "rpart", trControl =
ctrl)


# Print accuracy of each model

cat("Logistic Regression Accuracy:", log_model$results$Accuracy, "\n")
cat("LDA Accuracy:", lda_model$results$Accuracy, "\n")
cat("Decision Tree Accuracy:", tree_model$results$Accuracy, "\n")

# Visualize Accuracy Comparison

# Create a data frame to hold model accuracy values
accuracy_df <- data.frame(
  Model = c("Logistic Regression", "LDA", "Decision Tree"),
  Accuracy = c(
    log_model$results$Accuracy,
    lda_model$results$Accuracy,
    mean(tree_model$results$Accuracy)  # in case of multiple values
  )
)

# Load ggplot2 for plotting
library(ggplot2)

# Create bar plot
ggplot(accuracy_df, aes(x = Model, y = Accuracy, fill = Model)) +
  geom_bar(stat = "identity", width = 0.6) +
  ylim(0, 1) +
  geom_text(aes(label = round(Accuracy, 4)), vjust = -0.5, size = 4.5) +
  labs(title = "Classification Model Accuracy Comparison", y = "Accuracy", x =
"Model") +
  theme_minimal()
```

The data was preprocessed earlier by assigning appropriate column names and converting the target variable (Diagnosis) into a factor variable. 5-fold cross-validation was
done to offer balanced and precise model estimation. Models were trained for all of them using caret in R, and their accuracy values were compared numerically and graphically using a bar plot.

This comparison helped us identify which model provided the highest accuracy in classifying breast cancer using available features.

The visualisation is:



The bar chart illustrates the classification accuracies of three different models—Logistic Regression, Linear Discriminant Analysis (LDA), and Decision Tree—trained using 5-fold cross-validation. Among the models, LDA achieved the highest accuracy of 95.78%, followed by Logistic Regression at 94.38%, and Decision Tree at 89.52%. The visual comparison clearly highlights that both LDA and Logistic Regression outperform the Decision Tree model in terms of classification accuracy for this dataset.

Based on the results of the model comparison, Linear Discriminant Analysis (LDA) emerges as the most effective algorithm for predicting breast cancer diagnosis using this dataset. It not only achieved the highest cross-validated accuracy but also demonstrated consistent performance across different splits of the data. While Logistic Regression also performed well, the Decision Tree model showed comparatively lower accuracy, indicating it may not generalize as effectively in this context. Therefore, LDA is recommended as the preferred model for this classification task.

2. Which of the characteristics (radius, texture, symmetry) are most important is separating malignant from benign tumours?

CODE:

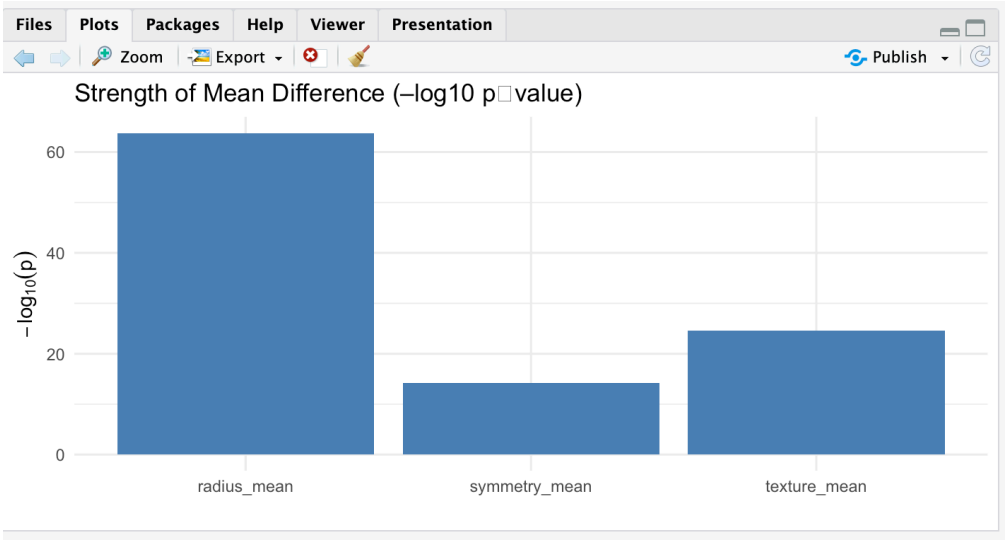
```
# Re-train models
set.seed(123)
ctrl <- trainControl(method = "cv", number = 5)
log_model  <- train(Diagnosis ~ ., data = data, method = "glm",    family =
"binomial", trControl = ctrl)
tree_model <- train(Diagnosis ~ ., data = data, method = "rpart", trControl =
ctrl)

# Step 1: t-tests for mean differences

features_q2 <- c("radius_mean", "texture_mean", "symmetry_mean")

ttest_list <- lapply(features_q2, function(feat) {
```

```
  t <- t.test(data[[feat]] ~ data$Diagnosis)
  data.frame(
    Feature   = feat,
    Mean_B    = round(t$estimate[1], 3),
    Mean_M    = round(t$estimate[2], 3),
    t_stat    = round(t$statistic, 3),
    p_value   = signif(t$p.value, 3)
  )
})
ttest_df <- do.call(rbind, ttest_list)
print(ttest_df)


# Model-based variable importance

# Logistic regression importance
log_imp  <- varImp(log_model, scale = FALSE)$importance
# Decision tree importance
tree_imp <- varImp(tree_model, scale = FALSE)$importance

# Extract our three features
log_imp_q2  <- log_imp[features_q2, , drop = FALSE]
tree_imp_q2 <- tree_imp[features_q2, , drop = FALSE]

# Combine into one data.frame
importance_df <- data.frame(
  Feature    = rep(features_q2, 2),
  Importance = c(log_imp_q2$Overall, tree_imp_q2$Overall),
  Model      = rep(c("Logistic Regression","Decision Tree"), each =
length(features_q2))
)

print(importance_df)


# Visualization

library(ggplot2)

# Bar plot of t-test p-values (log scale)
ggplot(ttest_df, aes(x = Feature, y = -log10(p_value))) +
  geom_bar(stat = "identity", fill = "steelblue") +
  labs(title = "Strength of Mean Difference (-log10 p-value)",
       y = expression(-log[10](p)), x = "") +
  theme_minimal()

# Bar plot of model importance
ggplot(importance_df, aes(x = Feature, y = Importance, fill = Model)) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(title = "Feature Importance from Models",
       y = "Importance (caret::varImp)", x = "") +
  theme_minimal()
```

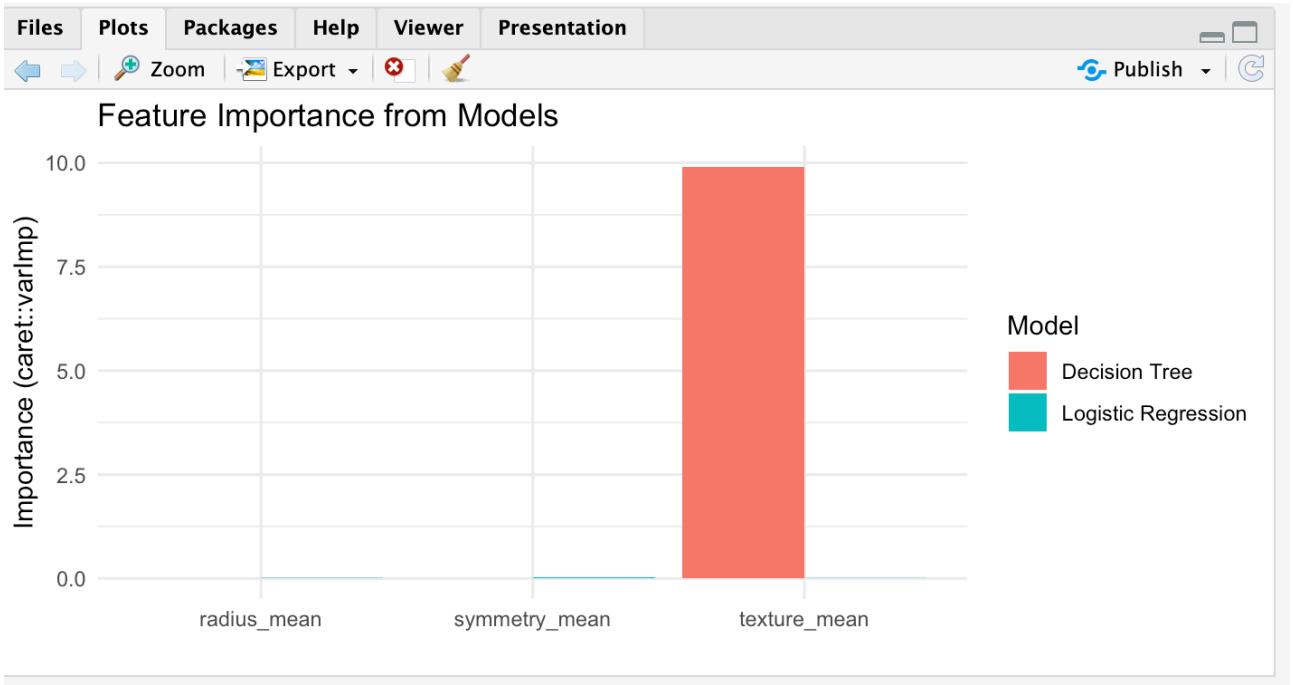
I analyzed which features among radius_mean, texture_mean, and symmetry_mean are most effective in distinguishing between malignant (M) and benign (B) tumors. To accomplish this, I applied both statistical methods and model-based approaches.
I first conducted independent two-sample t-tests to check for differences in average values of each feature between the two diagnosis groups. The sizes of these differences were represented by a bar chart based on the $-\log_{10}$ of the p-values so that I can identify which features have the greatest statistical discrimination.
Then I trained a Logistic Regression and a Decision Tree model with the entire dataset. employed the varImp() function to obtain the feature importance values from both models in order to see what among the three features had the highest influence on prediction. Then compared the importance values graphically using a bar plot.This two-pronged approach—statistical testing and model-

based evaluation—allowed us to identify features that are both statistically significant and practically useful for breast cancer diagnosis.

Visualisations:



The bar chart displays the statistical strength of the mean differences between malignant and benign tumors for three selected features, based on $-\log_{10}$(p-value) from independent t-tests. A higher bar indicates a more statistically significant difference between the two groups. Among the three features, radius_mean shows the strongest separation, followed by texture_mean and symmetry_mean. This suggests that radius_mean is the most statistically significant feature in distinguishing between malignant and benign tumors.



This bar chart compares the importance of three features—radius_mean, texture_mean, and symmetry_mean—in two classification models: Decision Tree and Logistic Regression. The chart reveals that the Decision Tree model relies heavily on texture_mean, while Logistic Regression assigns low and nearly equal importance to all three features, with slightly higher weight to

symmetry_mean. This indicates that texture_mean plays a key role in decision tree-based classification, whereas Logistic Regression distributes its attention more evenly across the inputs.

To determine which features are most important in distinguishing between malignant and benign tumors, we analyzed radius_mean, texture_mean, and symmetry_mean using both statistical testing (t-tests) and model-based feature importance.

The t-test results showed that radius_mean had the strongest statistical difference between the two classes, indicating it is highly significant in separating benign from malignant tumors. However, when evaluating practical feature importance through models, the Decision Tree model relied almost entirely on texture_mean, while Logistic Regression assigned slightly higher importance to symmetry_mean.

This contrast highlights that statistical significance does not always align with model preference, and both perspectives are valuable. Overall, the combination of t-tests and model insights suggests that while radius_mean is statistically distinct, texture_mean is the most influential feature in decision-based prediction, making it a key factor for model-driven breast cancer diagnosis.

Overall, the project demonstrated how statistical analysis and machine learning models can complement each other in medical data exploration. It also emphasized the importance of both accuracy evaluation and feature interpretability when building predictive models in a real-world healthcare context.

VIDEO LINK:

R-CODE IMPLEMENTATION