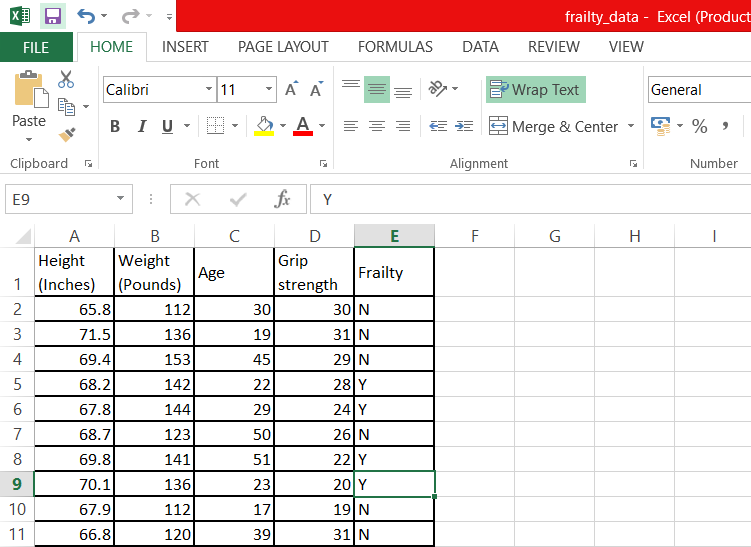
**PDS Assignment**

**Name : Surya Vamsi Pasupuleti**

**Student Id: 16332070**

# Works in Stage-1 (Data Acquisition/Collection):

All project files are often compiled into a single directory, which is then further divided into subdirectories for data, source code, analytical results, etc.



**Folder Structure:**

| - - Frailty\_Project

| | - - raw\_data

|  |  |  |
| --- | --- | --- |
| | | | | | - - dataset.csv |
| | | | | | - - README.txt |

| | - - clean\_data

| | - - results

| | - - src

# Works in stage-2 (Data processing):

We can easily develop a small script that will read the raw table, eliminate the rows with NA yields and those with a field code of N, and save the resulting processed data.

## Folder Structure:

| - - Frailty\_Project

| | - - raw\_data

| | | - - raw\_dataset.csv

| | | - - README.txt

| |- - clean\_data

| | | - - cleaned\_data.csv

| | - - results

| | - - src

| | | - - clean\_data.R

Graphical user interface, table

Description automatically generated with medium confidence

Graphical user interface, application

Description automatically generated

Text

Description automatically generated

Chart, scatter chart

Description automatically generated

There are no missing values after visualization (scatter plot), therefore we may utilize raw data as input to train a prediction model directly.

# Works in stage-3 (Data Analysis):

To predict frailty, we fitted several models (logistic regression, support vector machine, and decision tree) to cleaned and preprocessed data. We divided the data into training and testing sets, fitted the models to the training set, then predicted on the testing set. The models' performance was then tested using confusion matrices.

## Folder Structure:

| - - Frailty\_Project

| | - - raw\_data

|  |  |  |
| --- | --- | --- |
| | | | | | - - raw\_dataset.csv |
| | | | | | - - README.txt |

| | - - clean\_data

| | | - - cleaned\_data.csv

| | - - results

| | | - - test\_results.txt

| | - - src

| | | - - analysis.R

| | | - - clean\_data.R

# R Snippet:

**# Load necessary libraries library(caret)**

**# Load data**

**raw\_yield\_data <- read.csv("C:/Users/DELL/Desktop/frailty\_data.csv") # Remove rows with missing values**

**cleaned\_yield\_data <- na.omit(raw\_yield\_data) # Convert Frailty column to a factor**

**cleaned\_yield\_data$Frailty <- as.factor(cleaned\_yield\_data$Frailty) # Split data into training and testing sets**

**set.seed(123)**

**trainIndex <- createDataPartition(cleaned\_yield\_data$Frailty, p = .7, list = FALSE) train <- cleaned\_yield\_data[trainIndex, ]**

**test <- cleaned\_yield\_data[-trainIndex, ] # Fit logistic regression model**

**lr\_model <- train(Frailty ~ ., data = train, method = "glm", family = "binomial")**

**# Fit support vector machine model**

**svm\_model <- train(Frailty ~ ., data = train, method = "svmRadial")**

**# Fit decision tree model**

**dt\_model <- train(Frailty ~ ., data = train, method = "rpart")**

**# Make predictions on test set**

**lr\_pred <- predict(lr\_model, newdata = test) svm\_pred <- predict(svm\_model, newdata = test) dt\_pred <- predict(dt\_model, newdata = test)**

**# Evaluate performance of models confusionMatrix(lr\_pred, test$Frailty) confusionMatrix(svm\_pred, test$Frailty) confusionMatrix(dt\_pred, test$Frailty)**

**Results:**

Confusion Matrix and Statistics

Reference Prediction N Y

N 1 0

Y 0 1

|  |  |  |
| --- | --- | --- |
| Accuracy : | 1 |  |
| 95% CI : | (0.1581, | 1) |
| No Information Rate : | 0.5 |  |
| P-Value [Acc > NIR] : | 0.25 |  |
| Kappa : | 1 |  |
| Mcnemar's Test P-Value : | NA |  |
| Sensitivity : | 1.0 |  |
| Specificity : | 1.0 |  |
| Pos Pred Value : | 1.0 |  |
| Neg Pred Value : | 1.0 |  |
| Prevalence : | 0.5 |  |
| Detection Rate : | 0.5 |  |
| Detection Prevalence : | 0.5 |  |
| Balanced Accuracy : | 1.0 |  |
| 'Positive' Class : | N |  |

> confusionMatrix(svm\_pred, test$Frailty) Confusion Matrix and Statistics

Reference Prediction N Y

N 1 1

Y 0 0

Accuracy : 0.5

95% CI : (0.0126, 0.9874)

No Information Rate : 0.5 P-Value [Acc > NIR] : 0.75

Kappa : 0 Mcnemar's Test P-Value : 1.00

Sensitivity : 1.0

Specificity : 0.0 Pos Pred Value : 0.5 Neg Pred Value : NaN Prevalence : 0.5 Detection Rate : 0.5

Detection Prevalence : 1.0 Balanced Accuracy : 0.5

'Positive' Class : N

> confusionMatrix(dt\_pred, test$Frailty) Confusion Matrix and Statistics

Reference Prediction N Y

N 1 1

Y 0 0

Accuracy : 0.5

95% CI : (0.0126, 0.9874)

No Information Rate : 0.5 P-Value [Acc > NIR] : 0.75

Kappa : 0 Mcnemar's Test P-Value : 1.00

Sensitivity : 1.0

Specificity : 0.0 Pos Pred Value : 0.5 Neg Pred Value : NaN Prevalence : 0.5 Detection Rate : 0.5

Detection Prevalence : 1.0 Balanced Accuracy : 0.5

'Positive' Class : N

Question 2

Data Visualization

Plot 1: Scatter Plot

Graphical user interface, text, application, email

Description automatically generated

Chart, line chart, scatter chart

Description automatically generated

Plot 2:Bar Graph

Graphical user interface, text, application, email

Description automatically generated

Chart, histogram

Description automatically generated

Plot 3:Box Plot

Graphical user interface, text, application

Description automatically generated

Chart, box and whisker chart

Description automatically generated

Plot 4:Histogram

Graphical user interface, text, application

Description automatically generated

Chart, histogram

Description automatically generated

Plot 5:Pie chart

Graphical user interface, text, application

Description automatically generated

Chart, pie chart

Description automatically generated