



SYMBIOSIS INSTITUTE OF TECHNOLOGY, PUNE

Symbiosis International (Deemed University)

(Established under section 3 of the UGC Act, 1956) **Re-accredited by NAAC**

with 'A' grade (3.58/4) | Awarded Category – I by UGC

Founder: Prof. Dr. S. B. Mujumdar, M. Sc., Ph. D. (Awarded Padma Bhushan and Padma Shri by President of India)

Assignment No. 09	
Subject: Data Science Lab	
Name of Student	Achyut Shukla
PRN No.	20070122005
Branch	CS
Class	A1
Academic Year & Semester	2023-24 _ 7th semester
Date	17th October
Title of Lab Assignment	CLASSIFICATION MODEL

Theory:

Classification Models:

- *Random Forest: An ensemble learning method which constructs a multitude of decision trees during training time and outputs the mode of the classes for classification.*
- *Decision Tree: A flowchart-like structure in which the internal node represents a feature(or attribute), the branch represents a decision rule, and each leaf node represents an outcome.*
- *Logistic Regression: A statistical method for analyzing datasets where the outcome is binary. It predicts the probability of one of the two outcomes.*

Metrics:

- *Mean, Mode, Median: Measures of central tendency.*
- *Confusion Matrix: A table layout that visualizes the performance of an algorithm.*
- *Recall: The ratio of correctly predicted positive observations to the actual positives.*
- *Precision: The ratio of correctly predicted positive observations to the predicted positives.*
- *Entropy: Measure of randomness or unpredictability in the dataset.*
- *R2 Score: Represents the proportion of the variance for the dependent variable that's explained by independent variables in a regression model.*
- *Mean Square Error (MSE): The average of the squares of the errors or deviations, i.e., the difference between estimator and what is estimated.*
- *Mean Absolute Error (MAE): Represents the average of the absolute difference between the observed actual outturns and the forecasted values.*

Answer:

```
# Install and Load Necessary Libraries
install.packages(c("randomForest", "rpart", "caret", "e1071",
"ggplot2"))

library(randomForest)
library(rpart) library(caret)
library(e1071)
library(ggplot2)

# Load mtcars dataset
data(mtcars)

# EDA summary(mtcars) pairs(mtcars)
hist(mtcars$mpg) boxplot(mtcars$mpg, main
= "Boxplot of mpg")

# Let's take 'am' (automatic or manual transmission) as target variable
# Convert it to factor for classification mtcars$am <-
as.factor(mtcars$am)

# Splitting the dataset into training and test set
set.seed(123)
index <- sample(1:nrow(mtcars), nrow(mtcars)*0.7) train
<- mtcars[index,] test
<- mtcars[-index,]

# Classification using Random Forest
rf_model <- randomForest(am ~ ., data = train)
rf_pred <- predict(rf_model, test)

# Classification using Decision Tree
dt_model <- rpart(am ~ ., data = train, method = "class") dt_pred
<- predict(dt_model, test, type = "class")

# Classification using Logistic Regression
lr_model <- glm(am ~ ., data = train, family = binomial)
lr_pred_prob <- predict(lr_model, test, type =
"response") lr_pred <- ifelse(lr_pred_prob > 0.5, 1, 0)
lr_pred <- factor(lr_pred, levels = c(0,1))

# Metrics Calculation
metrics <- function(model_name, actual, predicted)
{ actual_numeric <-
as.numeric(as.character(actual)) cat("\n",
model_name, "\n") cat(" ----- \n")

# Confusion Matrix
conf_matrix <- confusionMatrix(predicted, actual)
print(conf_matrix$table)
```

```

# Basic Metrics cat("Mean:",
mean(actual_numeric), "\n") cat("Median:",
median(actual_numeric), "\n")
cat("Mode:", as.numeric(names(which.max(table(actual))))), "\n")

# Precision and Recall
cat("Recall:", conf_matrix$byClass['Recall'], "\n")
cat("Precision:", conf_matrix$byClass['Precision'], "\n")

# Resampling Metrics resample_metrics <-
postResample(predicted, actual) cat("R2 Score:",
resample_metrics['Rsquared'], "\n") cat("Mean Absolute
Error:", resample_metrics['MAE'], "\n") cat("Mean Square
Error:", resample_metrics['RMSE']^2, "\n")
}
metrics("Random Forest", test$am, rf_pred)
metrics("Decision Tree", test$am, dt_pred)
metrics("Logistic Regression", test$am,
lr_pred)

# For entropy calculation entropy
<- function(data) { prob <-
table(data) / length(data)
-sum(prob * log2(prob))
}
cat("\nEntropy: ", entropy(test$am), "\n")

```

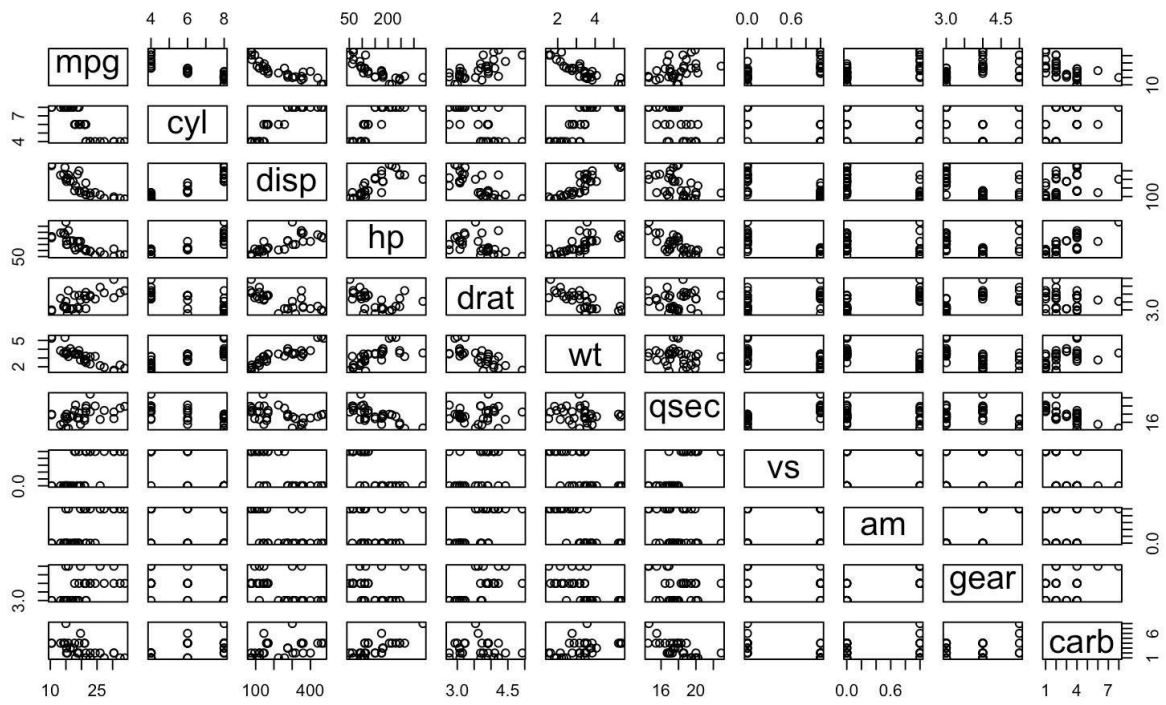
Output:

Answer:

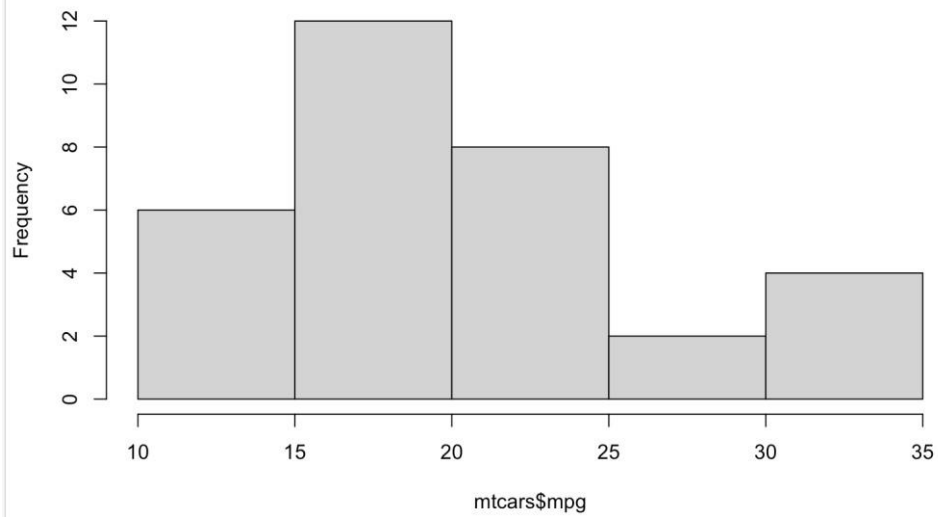
```

> # EDA
> summary(mtcars)
      mpg          cyl          disp         hp          drat
Min.   :10.40   Min.    :4.000   Min.    : 71.1   Min.    : 52.0   Min.    :2.760
1st Qu.:15.43   1st Qu.:4.000   1st Qu.:120.8   1st Qu.: 96.5   1st Qu.:3.080
Median :19.20   Median :6.000   Median :196.3   Median :123.0   Median :3.695
Mean   :20.09   Mean   :6.188   Mean   :230.7   Mean   :146.7   Mean   :3.597
3rd Qu.:22.80   3rd Qu.:8.000   3rd Qu.:326.0   3rd Qu.:180.0   3rd Qu.:3.920
Max.   :33.90   Max.   :8.000   Max.   :472.0   Max.   :335.0   Max.   :4.930
      wt          qsec          vs          am          gear
Min.   :1.513   Min.   :14.50   Min.   :0.0000   Min.   :0.0000   Min.   :3.000
1st Qu.:2.581   1st Qu.:16.89   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:3.000
Median :3.325   Median :17.71   Median :0.0000   Median :0.0000   Median :4.000
Mean   :3.217   Mean   :17.85   Mean   :0.4375   Mean   :0.4062   Mean   :3.688
3rd Qu.:3.610   3rd Qu.:18.90   3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:4.000
Max.   :5.424   Max.   :22.90   Max.   :1.0000   Max.   :1.0000   Max.   :5.000
      carb
Min.   :1.000
1st Qu.:2.000
Median :2.000
Mean   :2.812
3rd Qu.:4.000
Max.   :8.000
> pairs(mtcars)
> hist(mtcars$mpg)
> boxplot(mtcars$mpg, main = "Boxplot of mpg")

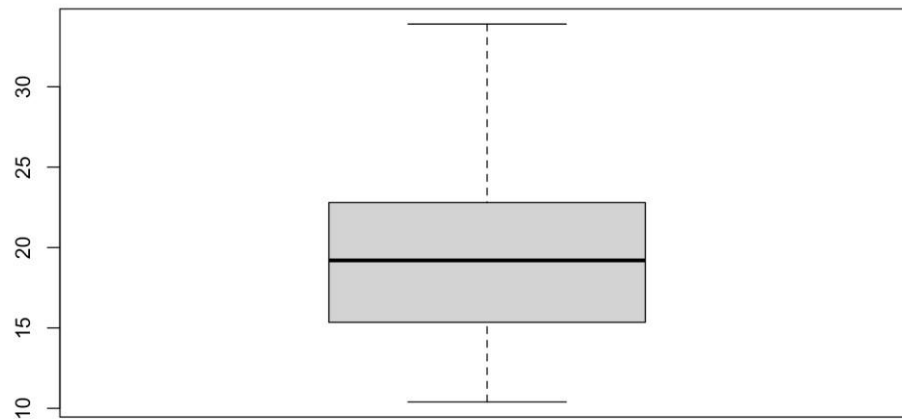
```



Histogram of mtcars\$mpg



Boxplot of mpg




```
Console Terminal x Background Jobs x
R 4.3.1 · ~/
> metrics("Random Forest", test$am, rf_pred)

Random Forest
-----
          Reference
Prediction 0 1
          0 7 0
          1 1 2
Mean: 0.2
Median: 0
Mode: 0
Recall: 0.875
Precision: 1
R2 Score: NA
Mean Absolute Error: NA
Mean Square Error: NA
> metrics("Decision Tree", test$am, dt_pred)

Decision Tree
-----
          Reference
Prediction 0 1
          0 7 0
          1 1 2
Mean: 0.2
Median: 0
Mode: 0
Recall: 0.875
Precision: 1
R2 Score: NA
Mean Absolute Error: NA
Mean Square Error: NA

> metrics("Logistic Regression", test$am, lr_pred)

Logistic Regression
-----
          Reference
Prediction 0 1
          0 7 0
          1 1 2
Mean: 0.2
Median: 0
Mode: 0
Recall: 0.875
Precision: 1
R2 Score: NA
Mean Absolute Error: NA
Mean Square Error: NA

> # For entropy calculation
> entropy <- function(data) {
+   prob <- table(data) / length(data)
+   -sum(prob * log2(prob))
+ }
> cat("\nEntropy: ", entropy(test$am), "\n")

Entropy: 0.7219281
>

Files Packages Help Viewer Presentation
```

Conclusion: In this lab assignment, students explored the mtcars dataset and practiced fundamental classification techniques. Through EDA, they understood the data's distributions and relationships. They then applied Random Forest, Decision Tree, and Logistic Regression models to classify whether a car had an automatic or manual transmission. Performance metrics were calculated for each model to evaluate and compare their effectiveness.