**Lab 2 – Assignment**

# Load Libraries

library(caret)

library(randomForest)

library(ggplot2)

# Read data from the file path

file\_path <- "C:/Users/navee/OneDrive/Desktop/Business Analytics/Machine Learning/Lab 2/oulad-assessments.csv"

data <- read.csv(file\_path)

# Remove rows with missing values

data <- na.omit(data)

# Convert factors

data$code\_module <- as.factor(data$code\_module)

data$code\_presentation <- as.factor(data$code\_presentation)

data$assessment\_type <- as.factor(data$assessment\_type)

# Creating a new target variable for classification based on submission timeliness

data$early\_submission <- as.factor(ifelse(data$date\_submitted <= data$date, "OnTime", "Late"))

# Split data into training and testing

set.seed(123) # for reproducibility

training\_indices <- createDataPartition(data$early\_submission, p = 0.8, list = FALSE)

train\_data <- data[training\_indices, ]

test\_data <- data[-training\_indices, ]

# Build a Classification Model (Random Forest)

model\_rf <- randomForest(early\_submission ~ code\_module + code\_presentation + assessment\_type, data = train\_data)

# Predict and Evaluate the Model

predictions\_rf <- predict(model\_rf, test\_data)

conf\_matrix <- confusionMatrix(predictions\_rf, test\_data$early\_submission)

print(conf\_matrix)

# Importance of Variables

importance(model\_rf)

varImpPlot(model\_rf)

# Visualization of Factor Levels Impact using ggplot2

ggplot(train\_data, aes(x=code\_module, fill=early\_submission)) +

geom\_bar(position="fill") +

labs(title="Submission Timeliness by Module", y="Proportion", x="Code Module") +

theme\_minimal()

Output:

> # Load Libraries

> library(caret)

> library(randomForest)

> library(ggplot2)

> # Read data from the file path

> file\_path <- "C:/Users/navee/OneDrive/Desktop/Business Analytics/Machine Learning/Lab 2/oulad-assessments.csv"

> data <- read.csv(file\_path)

> data <- na.omit(data)

> # Convert factors

> data$code\_module <- as.factor(data$code\_module)

> data$code\_presentation <- as.factor(data$code\_presentation)

> data$assessment\_type <- as.factor(data$assessment\_type)

> # Creating a new target variable for classification based on submission timeliness

> data$early\_submission <- as.factor(ifelse(data$date\_submitted <= data$date, "OnTime", "Late"))

> # Split data into training and testing

> set.seed(123) # for reproducibility

> training\_indices <- createDataPartition(data$early\_submission, p = 0.8, list = FALSE)

> train\_data <- data[training\_indices, ]

> test\_data <- data[-training\_indices, ]

> # Build a Classification Model (Random Forest)

> model\_rf <- randomForest(early\_submission ~ code\_module + code\_presentation + assessment\_type, data = train\_data)

> # Predict and Evaluate the Model

> predictions\_rf <- predict(model\_rf, test\_data)

> conf\_matrix <- confusionMatrix(predictions\_rf, test\_data$early\_submission)

> print(conf\_matrix)

Confusion Matrix and Statistics

Reference

Prediction Late OnTime

Late 6099 67

OnTime 3737 24271

Accuracy : 0.8887

95% CI : (0.8853, 0.892)

No Information Rate : 0.7122

P-Value [Acc > NIR] : < 2.2e-16

Kappa : 0.6945

Mcnemar's Test P-Value : < 2.2e-16

Sensitivity : 0.6201

Specificity : 0.9972

Pos Pred Value : 0.9891

Neg Pred Value : 0.8666

Prevalence : 0.2878

Detection Rate : 0.1785

Detection Prevalence : 0.1804

Balanced Accuracy : 0.8087

'Positive' Class : Late

> # Importance of Variables

> importance(model\_rf)

MeanDecreaseGini

code\_module 10352.214

code\_presentation 1097.544

assessment\_type 8834.910

> varImpPlot(model\_rf)

> # Visualization of Factor Levels Impact using ggplot2

> ggplot(train\_data, aes(x=code\_module, fill=early\_submission)) +

+ geom\_bar(position="fill") +

+ labs(title="Submission Timeliness by Module", y="Proportion", x="Code Module") +

+ theme\_minimal()

Ggplot2:

