Lab 1 Assignment

# Load necessary libraries

library(caret)

library(dplyr)

library(ggplot2)

# Read data from the file path

file\_path <- "C:/Users/navee/OneDrive/Desktop/Business Analytics/Machine Learning/Lab1/oulad-students.csv"

data <- read.csv(file\_path)

# Describe the variables

summary(subset\_data)

# Remove rows with missing values

data <- na.omit(data)

# Convert necessary variables to factors

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

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

data$gender <- as.factor(data$gender)

data$region <- as.factor(data$region)

data$highest\_education <- as.factor(data$highest\_education)

data$imd\_band <- as.factor(data$imd\_band)

data$age\_band <- as.factor(data$age\_band)

data$num\_of\_prev\_attempts <- as.factor(data$num\_of\_prev\_attempts)

data$disability <- as.factor(data$disability)

data$final\_result <- as.factor(data$final\_result)

# Visualization with ggplot2

ggplot(data, aes(x=highest\_education, fill=final\_result)) +

geom\_bar(position="dodge") +

theme\_minimal() +

labs(title="Final Result by Highest Education",

x="Highest Education", y="Count") +

theme(axis.text.x = element\_text(angle = 45, hjust = 1)) # Rotate x labels for better readability

# Split the data into training and testing sets (80% training, 20% testing)

set.seed(100) # For reproducibility

train\_index <- createDataPartition(data$final\_result, p = 0.8, list = FALSE)

train\_data <- data[train\_index, ]

test\_data <- data[-train\_index, ]

# Train the classification model (logistic regression)

model <- train(final\_result ~ ., data = train\_data, method = "glm", family = "binomial")

# Make predictions on the test data

predictions <- predict(model, newdata = test\_data)

# Evaluate the model

confusionMatrix(predictions, test\_data$final\_result)

Output:

> # Load necessary libraries

> library(caret)

> library(dplyr)

> library(ggplot2)

> # Read data from the file path

> file\_path <- "C:/Users/navee/OneDrive/Desktop/Business Analytics/Machine Learning/Lab1/oulad-students.csv"

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

> # Describe the variables

> summary(subset\_data)

Error: object 'subset\_data' not found

> # Remove rows with missing values

> data <- na.omit(data)

> # Convert necessary variables to factors

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

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

> data$gender <- as.factor(data$gender)

> data$region <- as.factor(data$region)

> data$highest\_education <- as.factor(data$highest\_education)

> data$imd\_band <- as.factor(data$imd\_band)

> data$age\_band <- as.factor(data$age\_band)

> data$num\_of\_prev\_attempts <- as.factor(data$num\_of\_prev\_attempts)

> data$disability <- as.factor(data$disability)

> data$final\_result <- as.factor(data$final\_result)

> # Visualization with ggplot2

> ggplot(data, aes(x=highest\_education, fill=final\_result)) +

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

+ theme\_minimal() +

+ labs(title="Final Result by Highest Education",

+ x="Highest Education", y="Count") +

+ theme(axis.text.x = element\_text(angle = 45, hjust = 1)) # Rotate x labels for better readability

> # Split the data into training and testing sets (80% training, 20% testing)

> set.seed(100) # For reproducibility

> train\_index <- createDataPartition(data$final\_result, p = 0.8, list = FALSE)

> train\_data <- data[train\_index, ]

> test\_data <- data[-train\_index, ]

> # Train the classification model (logistic regression)

> model <- train(final\_result ~ ., data = train\_data, method = "glm", family = "binomial")

There were 50 or more warnings (use warnings() to see the first 50)

> # Make predictions on the test data

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

> # Evaluate the model

> confusionMatrix(predictions, test\_data$final\_result)

Confusion Matrix and Statistics

Reference

Prediction Fail Withdrawn

Fail 0 4

Withdrawn 1 1953

Accuracy : 0.9974

95% CI : (0.9941, 0.9992)

No Information Rate : 0.9995

P-Value [Acc > NIR] : 0.9994

Kappa : -8e-04

Mcnemar's Test P-Value : 0.3711

Sensitivity : 0.0000000

Specificity : 0.9979561

Pos Pred Value : 0.0000000

Neg Pred Value : 0.9994882

Prevalence : 0.0005107

Detection Rate : 0.0000000

Detection Prevalence : 0.0020429

Balanced Accuracy : 0.4989780

'Positive' Class : Fail

Ggplot2

