Code:

# Load necessary libraries

library(caret)

library(dplyr)

# Read data from the Github link

file\_path <- "C:/Users/MSI/OneDrive/Desktop/ML/oulad-students.csv"

data <- read.csv(file\_path)

# Assuming 'data' is your data frame

subset\_data <- data[, c("id\_student", "date\_registration", "module\_presentation\_length", "studied\_credits", "num\_of\_prev\_attempts")]

# Describe the variables

summary(subset\_data)

# View the first few rows of the data

head(data)

# Remove rows with missing values

data <- na.omit(data)

# Convert categorical 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)

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

set.seed(120) # 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:

> # Read data from the Github link

> file\_path <- "C:/Users/MSI/OneDrive/Desktop/ML/oulad-students.csv"

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

> # Assuming 'data' is your data frame

> subset\_data <- data[, c("id\_student", "date\_registration", "module\_presentation\_length", "studied\_credits", "num\_of\_prev\_attempts")]

> # Describe the variables

> summary(subset\_data)

id\_student date\_registration module\_presentation\_length studied\_credits

Min. : 3733 Min. :-322.00 Min. :234 Min. : 30.00

1st Qu.: 508573 1st Qu.:-100.00 1st Qu.:241 1st Qu.: 60.00

Median : 590310 Median : -57.00 Median :262 Median : 60.00

Mean : 706688 Mean : -69.41 Mean :256 Mean : 79.76

3rd Qu.: 644453 3rd Qu.: -29.00 3rd Qu.:268 3rd Qu.:120.00

Max. :2716795 Max. : 167.00 Max. :269 Max. :655.00

NA's :45

num\_of\_prev\_attempts

Min. :0.0000

1st Qu.:0.0000

Median :0.0000

Mean :0.1632

3rd Qu.:0.0000

Max. :6.0000

> # View the first few rows of the data

> head(data)

code\_module code\_presentation id\_student gender region highest\_education

1 AAA 2013J 11391 M East Anglian Region HE Qualification

2 AAA 2013J 28400 F Scotland HE Qualification

3 AAA 2013J 30268 F North Western Region A Level or Equivalent

4 AAA 2013J 31604 F South East Region A Level or Equivalent

5 AAA 2013J 32885 F West Midlands Region Lower Than A Level

6 AAA 2013J 38053 M Wales A Level or Equivalent

imd\_band age\_band num\_of\_prev\_attempts studied\_credits disability final\_result

1 90-100% 55<= 0 240 N Pass

2 20-30% 35-55 0 60 N Pass

3 30-40% 35-55 0 60 Y Withdrawn

4 50-60% 35-55 0 60 N Pass

5 50-60% 0-35 0 60 N Pass

6 80-90% 35-55 0 60 N Pass

module\_presentation\_length date\_registration date\_unregistration

1 268 -159 NA

2 268 -53 NA

3 268 -92 12

4 268 -52 NA

5 268 -176 NA

6 268 -110 NA

> # Remove rows with missing values

> data <- na.omit(data)

> # Convert categorical 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)

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

> set.seed(120) # 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 47 warnings (use warnings() to see them)

> # 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 1

Withdrawn 1 1956

Accuracy : 0.999

95% CI : (0.9963, 0.9999)

No Information Rate : 0.9995

P-Value [Acc > NIR] : 0.9197

Kappa : -5e-04

Mcnemar's Test P-Value : 1.0000

Sensitivity : 0.0000000

Specificity : 0.9994890

Pos Pred Value : 0.0000000

Neg Pred Value : 0.9994890

Prevalence : 0.0005107

Detection Rate : 0.0000000

Detection Prevalence : 0.0005107

Balanced Accuracy : 0.4997445

'Positive' Class : Fail