library(dplyr)

library(caret)

hitdata <- read.csv("/Users/Hitakshibhatt/Dropbox/Northwood/ML/ML course hitdata download weekk1/studentInfo2.csv") # Replace "your\_hitdataset.csv" with the path to your hitdataset

# Data preprocessing

# Convert categorical variables to factors

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

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

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

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

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

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

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

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

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

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

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

set.seed(123) # For reproducibility

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

train\_hitdata <- hitdata[train\_index, ]

test\_hitdata <- hitdata[-train\_index, ]

# Train the classification model (logistic regression)

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

# Make predictions on the test hitdata

predictions <- predict(model, newhitdata = test\_hitdata)

# Evaluate the model

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