Lab – 2 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)
```

Interpretation:

Here is my interpretation of the logistic regression model I developed using the OULAD dataset. This interpretation covers various aspects of model evaluation, variable importance, model diagnostics, comparative metrics, interpretability techniques, practical insights, and statistical tests. Below, I outline the specific methods and insights derived from my analysis:

Confusion Matrix: Here in this code, I have generated a confusion matrix to understand how well my model performed, observing the number of correct and incorrect predictions for each class. This helped me compute further metrics like precision, recall, and F1-score.

Precision, Recall, and F1-Score:

Precision: I calculated precision to see how accurate my model's positive predictions were.

Recall: I computed recall to understand my model's ability to capture all relevant cases.

F1-Score: This score helped me balance precision and recall, especially when class distribution was uneven.

ROC Curve and AUC Score: Although not included in my analysis, these metrics would provide insights into how well my model distinguishes between classes.

Kappa Statistic: This metric wasn't initially analyzed, but it could measure the agreement between predicted and actual classifications, considering chance agreement.

2. Variable Importance

By examining the coefficients obtained from the logistic regression model I determined the importance and influence of each predictor on the outcome.

3. Model Diagnostics

Residual Analysis: While more relevant for regression tasks, I could still examine deviance residuals to assess the model's performance.

Cross-Validation Results: Although not implemented in my initial script, I could use caret's resampling techniques to ensure my model's stability and generalizability.

4. Comparative Metrics

I compared my logistic regression model's performance against simpler or more complex models to evaluate if the added complexity provided significant value.

5. Interpretability Techniques

While not utilized in my code, techniques like Partial Dependence Plots (PDPs) and SHAP Values would help me understand how each feature influences my model's predictions.

6. Practical and Operational Insights

Considering the deployment of my model in real-world scenarios, I would need to plan for updates and monitoring to ensure its continued effectiveness.

7. Statistical Tests

I could conduct tests to compare coefficients for significance or evaluate the fit of different logistic models, providing deeper insights into the robustness of my model's predictors.