Big Data Analytics Assessment 2

Overview

This assessment requires you to implement the tasks below using **PySpark** where applicable (Implementation only using normal Python will not receive credits). Please explain your code in detail in your report. Where appropriate, the output of code execution should be presented that relates your answer to the tasks into your report, as the evidence of working program (e.g., screenshots). Output included without supporting explanation or interpretation will not receive credits.

Where the tasks involve big data analytics with machine learning, you may need to formulate technical solution step by step, e.g., choose appropriate machine learning models / algorithms, justifying appropriateness of models/algorithms/techniques used, applying them in the context of the given task, and practising data visualisation techniques where appropriate. Test your solution and conduct experiments where appropriate. Evaluate the performance of implemented solution and analyse results where appropriate. Delve into deep technical explanations of the results and suggest possible improvement, etc. Make conclusions where appropriate.

You need to present your solution to the tasks into a technical report. The report should be submitted onto Blackboard. **To minimise similarity percentage, please do not copy the questions to your report** (just clearly label your answers with the corresponding question numbers).

Assignment tasks

You will need to download the dataset file "customer_purchases.csv" from the Blackboard. The dataset is about a study of sales measurements whether a customer has certain purchases. The key to the variables in the dataset is explained in the table below. Please note there are null or missing values ('0') in some columns of the dataset.

CustomerID	Customer ID numers
Age	Age(years)
Gender	Gender Identity
AnnualIncome	Annual income in pounds
SpendingScore	Number indicator for customer spending
PurchaseCategory	Category of products
TotalPurchases	Total quantity of purchases
PurchaseAmount	Total amount of purchases in pounds
Outcome	Class variable (0 or 1) for the purchase expectation

1. Load the dataset file into a PySpark DataFrame (1st DataFrame). Describe the structure of the DataFrame. (3 marks)

2. Replace the null/missing values (i.e.,'0' values) in the 'SpendingScore'' feature (or column) and 'TotalPurchase' feature (or column) of the 1st DataFrame with the median values of corresponding features, then save the results to a new PySpark DataFrame (2nd DataFrame). (5 marks)

3. Create a new PySpark DataFrame (3rd DataFrame) by removing rows from the 2nd DataFrame if a row's 'Age' feature, 'AnnualIncome' feature, or 'PurchaseAmount' feature has null/missing values (i.e., '0' values).

4. Compute the total number of rows removed from the 2nd DataFrame. (5 marks). Compute summary statistics of the 'PurchaseAmount' feature in the 3rd DataFrame, including its min value, max value, mean value, median value, variance, and standard deviation. Generate a histogram for the 'PurchaseAmount' feature and describe the distribution of the feature. (5 marks)

5. Display the quartile info of the 'Total purchase' feature in the 3rd DataFrame. Generate a boxplot for the 'Total purchase' feature and discuss the distribution of the feature based on the boxplot. (5 marks)

6. Use a graph to explore and describe the relationship between 'PurchaseAmount' feature and 'SpendingScore' feature in the 3rd DataFrame. Compute the Pearson correlation between the two features. (5 marks)

7. Use Spark SQL query to display the 'Age' feature and 'SpendingScore' feature in the 3rd DataFrame where 'Age' is less than 50 and 'SpendingScore' is great than 100. (5 marks)

8. Use the ‘Outcome’ feature in the 3rd DataFrame as the target label, to build a Decision Tree classifier using all other features as predictors. Conduct performance evaluation for the model and make conclusions. (9 marks)

9. Use the ‘Outcome’ feature in the 3rd DataFrame as the target label, to build a Logistic Regression classifier using all other features as predictors. Conduct performance evaluation for the model and make conclusions. (9 marks)

10. Build a linear regression model to predict ‘PurchaseAmount’ feature in the 3rd DataFrame using ‘AnnualIncome’ feature as the predictor. Conduct performance evaluation for the model and make conclusions. (9 marks)

Note: To read categorical data, you need to convert them into numerical data