Assignment 3 (25%)
STQD6324 Data Management
SEMESTER 2 2023/2024

In this assignment, you will use Spark MLlib to perform classification on the Iris dataset. The Iris dataset can be accessed from the R environment.

The following steps would be helpful in completing the assignment:

- Load the Iris dataset into a Spark DataFrame.
- Split the dataset into training and testing sets.
- Select a classification algorithm (e.g., Decision Trees, Random Forest, Logistic Regression) from Spark MLlib.
- Employ techniques such as cross-validation and grid search to fine-tune the hyperparameters of the chosen algorithm.
- Evaluate the performance of the tuned model using relevant evaluation metrics (e.g., accuracy, precision, recall, F1-score).
- Use the tuned model to generate predictions on the testing data.
- Conduct a comparative analysis between the predicted labels and the actual labels to assess the model's performance.

You are encouraged to explore creative approaches and leverage online resources in completing this assignment.

Please share your completed work via GitHub before **2024-06-04**.