TUTORIAL
STQD6324 Data Management
SEMESTER 2 2024/2025

In this tutorial, you will use Spark MLlib to perform classification on the Iris dataset. The Iris dataset can be downloaded online.

The following steps will guide you in completing the tutorial:

- Load the Iris dataset into a Spark DataFrame.
- Split the dataset into training and testing sets.
- Choose a classification algorithm (e.g., Decision Trees, Random Forest, Logistic Regression) from Spark MLlib.
- Use techniques such as cross-validation and grid search to fine-tune the hyperparameters of the selected 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.
- Compare the predicted labels with the actual labels to assess the model's performance.

You are encouraged to explore creative approaches (e.g., Jupyter Notebook or Apache Zeppelin) in completing this tutorial.