Big Data Analytics Techniques and Applications

Homework 4
Due Date: 2022/05/25 23:59:59

- Dataset

  Airline on-time performance datasets. You can find these datasets on E3.

  Analyze the Airline Dataset by using Spark MLlib. You may choose either one language from Java, Scala, Python, or R to implement it. Build a predictive framework for predicting whether each flight in 2005 **will delay or not** by using the data from 2003 to 2004 as training data.

- Questions:
  - Q1: Show the predictive framework you designed.

    Hint: What features do you extract? What algorithms do you use in the framework?
  - Q2: Explain the validation method you use.

    Hint: Leave-one-out, Holdout, k-fold, or other methods?
  - Q3: Explain the evaluation metric you use.

    Hint: Don't just show the prediction results, you should show the effectiveness of your framework (e.g., using a confusion matrix).
  - Q4: Show the validation results and give a summary of results.

- Requirements
  - Submit a report named **"HW4_StudentID.pdf" and your source code** to E3 and describe clearly the following items:
    - You can use **Google Colab** or other platforms.
    - The execution results by using Spark (Attach source code)
    - Descriptions of how you solve each question in detail.
    - Some figures or tables to illustrate your analyzed answers to each question.

        - Anything else worth mentioning (e.g., other valuable observations, or difficulties encountered in this work and how you resolve them).

- Penalty for late submission
    - If your work is submitted within one day after the deadline, a penalty of 20 percentage marks will be applied.
    - If your work is submitted within two days after the deadline, a penalty of 50 percentage marks will be applied.
    - If your work is submitted over two days after the deadline, you will get 0 in this homework.