Big Data Analytics Techniques and Applications

Homework 4

Due Date: 2023/05/24 23:59:59

- **Dataset:**

1. Airline on-time performance datasets

The above files can be downloaded 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.

* Please make sure that you use the correct dataset.

- **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" (e.g., HW4_310456099.pdf) to E3 and describe the following items clearly in your report:

    - 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).

- Submit source code files to E3.
  - You should **zip** all source code files in a file named "HW4_{StudentID}_Code.**zip**" (e.g., HW4_310456099_Code.zip).

- **Penalty for late submission:**
  - If you submit your work within one day after the deadline, you will get **80%** of the original score.
  - If you submit your work within two days after the deadline, you will get **50%** of the original score.
  - If you submit your work more than two days after the deadline, you will get **zero score** on this homework.

- **Penalty for format error:**
  - The report file name has any format error. (-5%)
  - The report is not in pdf. (-5%)
  - The source code file is not in zip. (-5%)