MSC-BDT5002 Knowledge Discovery and Data Mining, Fall 2018

Assignment 2

Deadline: Oct. 25th, 11:59pm, 2018

Task Description

The dataset come from 1994 Census database. Prediction task is to determine whether a person makes over 50K a year.

Files Description

1. trainFeatures.csv: 34189 individual's basic information with 14 attributes for training.

2. testFeatures.csv: 14653 individual's basic information with 14 attributes for testing.

3. trainLabels.csv: 34189 individual's incomes, 0: <=50k, 1: >50k.

4. sampleSubmission.csv: The sample submission file you may refer.

5. dataDescription.pdf: 14 attributes information.

Notes

1. **You must use ensemble learning algorithm to do Prediction.**

2. Real-world data contains noise, missing values or even mistakes. Pre-processing is necessary.

3. Your assignment will be graded by the testing accuracy and clarification for your feature engineering (in readme.pdf).

4. TA will check your source code carefully, so your code must be runnable. Keep your code clean and comment it clearly.

5. You can use any programming language. In principle, python is preferred.

6. Cheating is not allowed. Your result **MUST** be reproducible.

7. Plagiarism will lead to zero mark.

Submission Guidelines

1. Assignment should be submitted to mscbdt5002fall18@gmail.com as attachment

2. You need to zip the following three files together:

   - A2_itsc_stuid_readme.pdf. Write your feature engineering in it
   - A2_itsc_stuid_code.zip: The zip file contains all your source codes.
   - A2_itsc_stuid_prediction.csv: The prediction result.

3. Attachment should be named in the format of: A2_itsc_stuid.zip. E.g. A2_lliny_20181234.zip.

4. Submissions after the deadline or not following the rules above are NOT accepted.