THE HONG KONG UNIVERSITY OF SCIENCE & TECHNOLOGY
Department of Computer Science and Engineering
COMP4331:Introduction to Data Mining
Fall 2020 Mini-Project
Due time and date: 11:59pm, Oct 26 (Mon), 2020.

IMPORTANT NOTES

- **Your grade will be based on the correctness and clarity.**
- **Late submission: 25 marks will be deducted for every 24 hours after the deadline.**
- **ZERO-Tolerance on Plagiarism: All involved parties will get zero mark.**

Objective

In this project, you are given a dataset containing features extracted by a deep network on a 3-class classification problem. Each record corresponds to features generated for a sample. Your objective is to build a decision tree model using these features. Your decision tree will be evaluated by its accuracy on the test set (which is hidden from you). Please write your code in Python and use the library Scikit-learn to create the decision tree model.

Dataset

You are provided with a csv file named data.csv. This dataset contains 2048 features for each record, with their labels in the last column indicating the class ("0", "1" or "'2"). There are a total of 2060 records.

Tasks

1. (Report) In your report, provide details for your whole procedure (including data preprocessing, model creation, training, predictions, etc). Note that you may also use techniques not covered in the lecture notes (e.g., other preprocessing methods, or other decision tree variants). In that case, you have to clearly explain what these techniques are in the report.
2. (Report) For each part of your procedure, include justifications and reasoning for your work. You may also include any graphical representations to support your points.
3. (Report) Choose three splits from your decision tree. For each of them, show the histograms of the three classes based on the selected feature. Show all three histograms in the same plot so that the classes overlap is shown.
4. (Code) Write all your code in a Python notebook named mini-project.ipynb, including **only** your procedure as described in your report.
5. (Outputs) Export your decision tree model as an image (either as decision_tree.png or decision_tree.jpg).

Grading

Your grade will be evaluated by the following criteria:

1. (Report) The logical soundness of your procedure and report. Your explanations should be clear and the audience should be convinced by your reasoning for each step. The plots are clear and communicate effectively.

2. (Code) The code follows closely to the report's description and executes as intended. The generated decision tree image is clear and readable.

3. (Model) The performance of your decision tree model on our test set.

Submission Guidelines

Please submit a report (report.pdf), a Python notebook (mini-project.ipynb) for your code, and the final decision tree diagram (decision_tree.png or decision_tree.jpg). Zip all the files into either [your student ID]_mini-project.zip or [your student ID]_mini-project.tar.gz. Please submit the assignment by uploading the compressed file to Canvas. Note that the assignment should be clearly legible, otherwise you may lose some points if the assignment is difficult to read. Plagiarism will lead to zero point on this assignment.