CSCI316 – Big Data Mining Techniques and Implementation
Individual Assignment 1
2025 Session 3 (SIM)

15 Marks
Deadline: Refer to the submission link of this assignment on Moodle

Two (2) tasks are included in this assignment. The specification of each task starts in a separate page.

You must implement and run all your Python code in Jupyter Notebook. ***The deliverables include one Jupyter Notebook source file (with .ipybn extension) and one PDF document for each task.***

Note: To generate a PDF file for a notebook source file, you can either (i) use the Web browser's PDF printing function, or (ii) click "File" on top of the notebook, choose "Download as" and then "PDF via LaTex".

All results of your implementation must be reproducible from your submitted Jupyter notebook source files. In addition, the submission must include all execution outputs as well as clear explanation of your implementation algorithms (e.g., in the Markdown format or as comments in your Python codes).

Submission must be done online by using the submission link associated with assignment 1 for this subject on MOODLE. The size limit for all submitted materials is 20MB. DO NOT submit a zip file.

This is an individual assignment. Plagiarism of any part of the assignment will result in having 0 mark for the assignment and for all students involved.

Task 1

(5 marks)
Dataset: Credit Score Classification Dataset
Source: https://www.kaggle.com/datasets/parisrohan/credit-score-classification

Objective
Use Pandas in Python to clean and pre-process this dataset. You cannot use any ML library (including Sci-kit Learn) for this task, otherwise no mark for this task will be provided.

Requirements

(1) Create one Pandas data frame for this data set.
(2) Identify the attributes with missing values. Select one attribute and propose a method to clean the missing values in that attribute.
(3) Perform z-score normalization of the values in the attribute "Amount_invested_monthly". Show the mean and variance of the normalized values.
(4) Create four bins for the attribute "Amount_invested_monthly" such that the bins contain (approximately) equivalent numbers of records (samples).
(5) Apply one-hot-encoding to the attribute "Credit_Mix".

For the requirements (2) – (5), append the new columns to the existing Pandas dataframe.

Deliverables

- A Jupiter Notebook source file named `<your_name>_task1.ipybn` which contains your implementation source code in Python
- A PDF document named `<your_name>_task1.pdf` which is generated from your Jupiter Notebook source file, and presents clear and accurate explanation of your implementation and results.

Task 2

(10 marks)
Dataset: The Drug Classification Dataset
Source: https://www.kaggle.com/datasets/prathamtripathi/drug-classification

Objective
The objective of this task is to implement *from scratch* a Decision Tree classifier to predict the drug type. Note that the primary goal is to implement a DT classifier correctly, while the performance is less of a concern in this task.

You must not use any ML library for this task, otherwise no mark for this task will be provided.

Requirements

(1) Use binning to transform continuous attributes (if any) into discrete values.
(2) Use 80% samples for training, and 20% samples for test.
(3) It is recommended that your implementation includes a "tree induction function" and a "classification function".
(4) Note. You can (but not must) use any reasonable pre-processing methods. You also can (but not must) use pre-pruned technique. If you do so, you must explain your reasons clearly.

Deliverables

- A Jupiter Notebook source file named `<your_name>_task2.ipybn` which contains your implementation source code in Python
- A PDF document named `<your_name>_task2.pdf` which is generated from your Jupiter Notebook source file.