CSE5DMI 2024 Assignment One [20 marks]
Assignment Due: 11:59 PM, Sunday (in Week 7), 22 Sep 2024

GENERAL DESCRIPTION

In this **INDIVIDUAL** assignment, we are going to explore a given dataset and build decision tree to classify customers described by a set of attributes as good or bad credit risks.

Dataset

The data presented for this assignment are a randomly selected subset of the original data[1], one row for each customer.

- The dataset can be found in a CSV file (If you cannot find the file, please let us know ASAP)
- Each student's dataset will be slightly different, but with the same level of difficulty and usability.
- A detailed list of attribute descriptions can be found below.

Attribute	**Description**
Default	Categorical Class Label • 0: Good • 1: Bad
Status of existing checking account	Categorical • A11 : ... < 0 DM • A12 : 0 <= ... < 200 DM • A13 : ... >= 200 DM / salary assignments for at least 1 year • A14: no checking account
Duration in month	Numerical
Credit history	Categorical • A30 : no credits taken/ all credits paid back duly • A31 : all credits at this bank paid back duly • A32 : existing credits paid back duly till now • A33 : delay in paying off in the past • A34 : critical account/ other credits existing (not at this bank)
Purpose	Categorical • A40 : car (new) • A41 : car (used) • A42 : furniture/equipment • A43 : radio/television • A44 : domestic appliances • A45 : repairs • A46 : education • A47 : (vacation - does not exist?)

[1] Statlog (German Credit Data) Data Set, UCI machine learning repository

	• A48 : retraining • A49 : business • A410 : others
Credit amount	Numerical
Savings account/bonds	Categorical • A61 : ... < 100 DM • A62 : 100 <= ... < 500 DM • A63 : 500 <= ... < 1000 DM • A64 : .. >= 1000 DM • A65 : unknown/ no savings account
Present employment since	Categorical • A71 : unemployed • A72 : ... < 1 year • A73 : 1 <= ... < 4 years • A74 : 4 <= ... < 7 years • A75 : .. >= 7 years
Personal status and sex	Categorical • A91 : male : divorced/separated • A92 : female : divorced/separated/married • A93 : male : single • A94 : male : married/widowed • A95 : female : single
Other debtors / guarantors	Categorical • A101 : none • A102 : co-applicant • A103 : guarantor
Present residence since	Numerical
Property	Categorical • A121 : real estate • A122 : if not A121 : building society savings agreement/ life insurance • A123 : if not A121/A122 : car or other, not in attribute 6 • A124 : unknown / no property
Age in years	Numerical
Other instalment plans	Categorical • A141 : bank • A142 : stores • A143 : none
Housing	Categorical • A151 : rent • A152 : own • A153 : for free
Number of existing credits at this bank	Numerical
Job	Categorical • A171 : unemployed/ unskilled - non-resident • A172 : unskilled - resident • A173 : skilled employee / official

	• A174 : management/ self-employed/ • highly qualified employee/ officer
Number of people being liable to provide maintenance for	Numerical
Foreign worker	Categorical • A201 : yes • A202 : no

Requirements and tasks

Your final report and program need to address the following tasks. Marking criteria is given in Appendix.

1. Explore, aggregate and transform the attributes [**3 Marks**]
   i. Write a Python script to read the input CSV file, perform any necessary pre-processing so that the data becomes suitable to be used.
   ii. Describe the pre-processing you carried out with justifications in your report.

   (Hint: Think about missing values and categorical attributes. How to deal with them? Should we turn them all into dummies? Use only some? Create new informative attributes by aggregating some attributes? etc.)
   iii. Submit the pre-processed data in CSV format

2. Classification. [**8 Marks**]

   Using the data exported in Part-1, create a decision tree learner, and perform 10-fold cross-validation to evaluate the performance of decision tree classifier with this data. You must provide
   i. python source code reading the source data, building the learner, and performing 10-fold cross-validation. [3 Marks]
   ii. performance evaluation results including confusion matrix, accuracy, and area under (receiver operating characteristic, ROC) curve (AUC), and cost matrix. [5 Marks]

   Use Cost Matrix as the primary evaluation measure to represent the performance. It is worse to class a record as good when they are bad (10), than it is to class a customer as bad when they are good (1).

   Note: no marks will be given without answer for 2-(i). You also need to follow the standard of coding: make your code elegant and readable by appropriate commenting, documentation, indentation, etc.

3. Open discussion [**9 Marks**]
   - Discuss your best classification model and other trials. In your report, explain why and how you fine-tuned the parameters, provide corresponding evaluation results and your findings. [3 Marks]

- Out of the attributes you have explored, what attribute do you think is the most important factor for deciding the quality of a customer? Is your finding intuitive? If you would report your findings to your manager or boss, what suggestions would you make to improve the related marketing strategy? [3 Marks]

  Hint: you may find it helpful to visualise the tree to assist your discussion.

- Do you think your model suffers from overfitting? What is your evidence? If your model is overfitting, please apply appropriate techniques to alleviate this problem and show improvement. If your model is not overfitting, please justify this conclusion (hint: maybe by drawing learning curves for demonstration). [3 Marks]

Submit your Python source codes in a single Python script file, the pre-processed data (CSV format), and a report.

IMPORTANT NOTES

1. A penalty of **5%** of the marks per day will be imposed on late submissions of assessment up to five (5) working days after the due date. **An assignment submitted more than FIVE working days after the due date will NOT be accepted, and ZERO mark will be assigned.**
2. If you need an extension of **up to five business days**, you should apply using the 'Request for Extension' form **up to three business days** before the due date of your assessment.
3. When you cannot apply for a short extension prior to the deadline, or you need more than a five-day extension, you should review the criteria on this page https://www.latrobe.edu.au/students/admin/forms/special-consideration/eligibility-criteria to see if you are eligible to apply for Special Consideration. More information can be found in the Assessment Procedure – Adjustments.
4. Academic misconduct includes poor referencing, plagiarism, copying and cheating. **Copying, Plagiarism**: Plagiarism is the submission of somebody else's work in a manner that gives the impression that the work is your own. Recall that **the University takes academic misconduct very seriously**. When it is detected, penalties are strictly imposed. You should familiarise yourself with your responsibilities about Academic Integrity. Detailed information can be found here: http://www.latrobe.edu.au/students/learning/academic-integrity

SUBMISSION GUIDELINE

- Submit before 11:59 PM, Sunday, 22 Sep 2024 (Week 7).
- Upload a single .zip archive onto LMS before the deadline. The .zip archive needs to be named with your SID, e.g. if your SID is "12345678", then the archive must be called "12345678.zip". It should contain
  - Python source code to support your answers to Tasks 1, 2, and 3.
    - **Assignment submitted without Python source code will not be marked.**
  - Pre-processed CSV file of Task 1
  - Report in word or PDF for Tasks 1, 2 and 3.
- Late submissions will incur a penalty of **5% of the marks per day**.

Appendix: Assignment 1– Marking guide

Assessment criteria / grading rubric

CRITERIA	A: Excellent (>80%)	B: Very Good (70–79%)	C: Good (60–69%)	D: Acceptable (50–59%)	N: Unacceptable (<50%)
Perform correct data processing for different conditions	Consistently performs correct pre-processing for different conditions.	Mostly performs correct pre-processing for different conditions.	Sometimes performs correct processing for different conditions.	Occasionally performs correct processing for different conditions.	Rarely or does not perform(s) correct processing for different conditions.
Design, implement and evaluate decision trees in correct architecture	Designs, implements and comprehensively evaluates decision tree in correct architecture with no coding errors.	Designs, implements and comprehensively evaluates decision tree in correct architecture with minor, insignificant coding errors.	Designs, implements and evaluates decision tree in correct architecture with some insignificant coding errors.	Designs, implements and evaluates decision tree with significant coding errors.	Designs, implements and evaluates decision tree in incorrect architecture.
Perform critical and thorough analysis and discussion of different data mining models to improve performance.	Performs correct and thorough investigation and adjustment of hyperparameters of data mining models to achieve optimal and improved classification performance. Discusses results with clear logic and draws insightful conclusions.	Mostly performs correct investigation of hyperparameters of data mining models to achieve optimal and improved classification performance. Discusses results with mostly correct logic and draws meaningful conclusions.	Sometimes performs correct investigation of hyperparameters of data mining models to achieve improved classification performance. Discusses results with reference to experiments and draws reasonable conclusions.	Occasionally performs correct hyperparameter settings of data mining models to obtain classification results. Occasionally provides reasonable justifications.	Performs no or insufficient investigations of parameter settings or incorrect parameter settings of data mining models.

END