CSE5DMI 2025 Assignment One [20 marks]
Assignment Due: 11:59 PM, Sunday (in Week 7), 21 Sep 2025

GENERAL DESCRIPTION

In this **INDIVIDUAL** assignment, we are going to explore a given dataset and build decision tree and a neural network (NN) 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 the 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**
Class Label ★	Categorical • 0: Good • 1: Bad
Status of existing checking account ★	Catogorical • A11 : ... < 0 DM deutsche mark 西德货币 • A12 : 0 <= ... < 200 DM • A13 : ... >= 200 DM / salary assignments for at least 1 year • A14: no checking account
Duration in month ★	Numerical
Credit history 信贷 ★	Catogorical • 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 ★	Catogorical • A40 : car (new) • A41 : car (used) • A42 : furniture/equipment • A43 : radio/television • A44 : domestic appliances • A45 : repairs • A46 : education

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

	• A47 : (vacation - does not exist?) • A48 : retraining • A49 : business • A410 : others
Credit amount	Numerical
Savings account/bonds	Catogorical • A61 : ... < 100 DM • A62 : 100 <= ... < 500 DM • A63 : 500 <= ... < 1000 DM • A64 : .. >= 1000 DM • A65 : unknown/ no savings account
Present employment since	Catogorical • A71 : unemployed • A72 : ... < 1 year • A73 : 1 <= ... < 4 years • A74 : 4 <= ... < 7 years • A75 : .. >= 7 years
Installment rate in percentage of disposable income 分期付款占可支配收入的百分比– not exist	Numerical
Personal status and sex	Catogorical • A91 : male : divorced/separated • A92 : female : divorced/separated/married • A93 : male : single • A94 : male : married/widowed • A95 : female : single
Other debtors / guarantors	Catogorical • A101 : none • A102 : co-applicant co–borrowers who responsible for repaying the loan • A103 : guarantor promise to repay the loan if the primary borrower defaults.
Present residence since	Numerical
Property	Catogorical • 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 create catogorical data	Numerical
Other installment plans 其他协定的分期付款方案	Catogorical • A141 : bank • A142 : stores • A143 : none
Housing	Catogorical • A151 : rent • A152 : own • A153 : for free
Number of existing credits at this bank active credit accounts	Numerical
Job	Catogorical

	• 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
Telephone	Catogorical • A191 : none • A192 : yes, registered under the customers name
Foreign worker	Catogorical • A201 : yes • A202 : no

Requirements and marking scheme

Your final report and program need to address the following tasks.

1. Explore, aggregate and transform the attributes [**2 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. Implementation. [**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, and conduct investigations to improve the performance. To answer question 2, you must provide

   i. python source code reading the source data, building the learner, and performing 10-fold cross-validation.

   ii. explanation of how you split the data during 10-fold cross-validation

   iii. performance evaluation results including confusion matrix, accuracy, and area under (receiver operating characteristic, ROC) curve (AUC), and cost matrix.

   Use Cost Matrix as the primary evaluation measure to represent the performance. It is worse to class a customer 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). Explicitly state the mapping between positive/negative and the class labels in your report.

iv. Self-explorations or investigations to improve the performance of the baseline model by fine-tuning the model's parameters or adjusting attributes of the original dataset.

You also need to follow the standard of coding: make your code elegant and readable by appropriate commenting, documentation, indentation, etc.

3. Communication and Discussion [**10 Marks**]

An in-depth report that delivers insights you have found in the implementations and investigations:

- Present the performance of your best classification model and other trails using figures, charts or tables. In your report, explain why and how you fine-tuned the parameters, provide corresponding evaluation results and your findings.
- 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 clients, what suggestions would you make to improve the related marketing strategy?

  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).

Submit your Python source codes in a single Colab Notebook file, the pre-processed data (CSV format), and the 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 FOUR working days after the due date will NOT be accepted, and ZERO mark will be assigned.**
2. 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, 21 Sep 2025 (Week 7).
- Submit the report in word or PDF for Tasks 1, 2, and 3 via **Assignment 1 – Report Submission** portal in LMS before the deadline.

- Submit the pre-processed CSV file and code (Colab notebook file or python source code) via **Assignment 1 – Data and Code Submission** portal before the deadline
  - The code should be able to support your answers to Tasks 1, 2, and 3.
  - **Assignment submitted without code will not be marked.**
- Late submissions will incur a penalty of **5% of the marks per day**.

END