CSE5DMI 2020 Assignment One [20 marks]
Completed Assignment Due: 11:00 PM, Monday (in Week 8), 14 Sep 2020

GENERAL DESCRIPTION

This **INDIVIDUAL** assignment consists of **TWO PARTS** and is worth **20%** of the assessment of this subject. This assignment aims at consolidating your knowledge base and developing practical skills to build a decision tree.

PART I (12 MARKS)

In this project, we are going to build a data mining model using decision tree to predict wine quality based on physicochemical tests.

- **Dataset**
  Each student will be assigned a dataset with wine samples collected from a public dataset[1].
  - The dataset can be found in the CSV file *part1_data*. (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[2].
  - The first six columns are attributes of wine samples, and the last column is the class label denoting the wine quality.
    i. The wine sample attributes, including alcohol, density and chlorides, are all numeric values.
    ii. Quality (class label) is of the categorial type where the value can be 5 or 7 denoting wine of average or excellent qualities respectively.

- **Requirements and marking scheme**
  Your final report and program need to address the following three subtasks.
  a. **Data pre-processing [2 marks]**
  There are one or more attributes with missing values.
  - Write a Python script to read the input CSV file, replace the missing values with the mean value for that attribute, and then export the entire dataset to another CSV file.
  - Briefly describe the pre-processing you performed in your final report.

  (Hint: You may need to check how your data import and/or export functions handle missing values and might require further pre-processing before step (b)).

[1]UIUC Wine Quality Data Set: https://archive.ics.uci.edu/ml/datasets/wine+quality. P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis. Modelling wine preferences by data mining from physicochemical properties. In Decision Support Systems, Elsevier, 47(4):547-553, 2009.
[2] The original dataset has been modified for this assignment. Modifications included random sampling, feature selection, etc.

b. **Wine quality prediction [7 marks]**
Using the data exported in part (a), create an Orange3 decision tree learner, and perform 10-fold cross-validation to evaluate the performance of decision tree classifier with this data. To answer question b, you need to 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). Note: no marks will be given without answer for (i). **[3 marks]**

   You also need to follow standard of coding: make your code elegant and readable by appropriate commenting, documentation, indentation, etc. **[1 mark]**

c. **Open discussion [3 marks]**
There is an argument that not all the given attributes contribute to the wine quality prediction results by decision tree. Do you agree with this claim? Justify your answer with both explanation and computational demonstration in your final report.
(Hint: you may want to investigate the accuracy and AUC of decision tree using fewer numbers of attributes).

PART II [8 MARKS]

In this part, we are going to build a decision tree **MANUALLY**.

a. A training dataset with 20 instances can be found in the PDF file *part2_data*. (If you cannot find the file, please let us know ASAP)
b. Build a decision tree for classification problem using GINI INDEX. Notes: no marks will be given without step by step calculation from the root node to the growth of the tree. **[6 marks]**
c. Will you be including Customer ID as an attribute? Justify your answer. **[2 marks]**

IMPORTANT NOTES

1. A penalty of **5%** of the marks per day will be imposed on late submissions of assessment up to four (4) 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. If you would like to seek extensions for submission, please apply for formal **Special Consideration**. To do this or find detailed information, please go to http://www.latrobe.edu.au/special-consideration
3. Recall that **the University takes academic misconduct very seriously**. Academic misconduct includes poor referencing, plagiarism, copying and cheating. You should familiarise yourself with your responsibilities about Academic Integrity. Detailed information can be found on the website at: http://www.latrobe.edu.au/students/learning/academic-integrity

SUBMISSION GUIDELINE

- Submit before 11:00 PM (Australian Eastern Standard Time), Monday, 14 Sep 2020 (Week 8).
- Late submissions will incur a penalty of **5% of the marks per day**.

1. PART I (12 MARKS)

- Code submission
  - You **MUST** upload a single *.zip* archive onto LMS before the deadline.
    - The *.zip* archive must be named with your SID, e.g. if your SID is "12345678", then the archive must be called "12345678.zip".
    - It should contain your program (**ALL the Python source code)** to support your answers of Part I.
  - The submission portal 'Assignment 1 – Code submission' can be found in LMS under "Assignment 1 & 2"
- Report submission
  - You must submit a documentation (word or PDF) of your answers to data-pre-processing, performance evaluation, and open discussions through submission portal 'Assignment 1 – Part 1 Report'.
  - In addition to the document, all answers must be supported with Python script file.

2. PART II (8 MARKS)

- Electronic submission of your answers through 'Assignment 1-Part 2 Report' submission portal in LMS
  - If you like, you are welcome to use pen and paper to complete Part II exercise. You may attach photos or scanned copies to a word document.
- Full marks will not be awarded without showing the steps of your calculation

END