**DATA MINING AND MACHINE LEARNING (EBUS537)**

**Coursework**

*Set by Professor Dongping SONG & Dr Eric Leung*

**Date of issue**: 30th October 2021

**Date of submission**: Thursday 14th January 2022 before 12 noon (ELECTRONIC SUBMISSION ONLY.)

**Contribution**: 100%.

**Report length:** 3500 words (excluding references and appendix).

**Assignment Requirements:**

**Question 1. (maximum 2000 words; 50%)**

You are given a training dataset, "**myCarTrainDataset.csv**", and a testing dataset, "**myCarTestDataset.csv**", which will be provided in electronic form in Canvas. You are required to apply decision tree classification technique to the above case appropriately. Specifically, you are required to:

1. Use the training dataset, apply the Greedy strategy combined with the Gini impurity measure to build a fully-grown decision tree. If the attribute has multiple attribute values, please use multiway split (do not use binary split). Leaf nodes should be declared as a single class label (do not use probability/fraction). Samples of the calculations and explanations should be provided to demonstrate the application process of the Greedy strategy and Gini impurity measure.
2. Perform the post-pruning activities to the fully-grown decision tree that was built in the previous step by applying the following rule: prune the sub-tree if all of its leaf nodes have the same class label. Test the decision tree using the test dataset. Discuss the results.
3. Beyond the case context, discuss the applications of decision tree-based classification methods in practices for management purposes. Support your arguments with relevant references.

**Question 2. (maximum 2000 words; 50%)**

Data mining and machine learning tools have a wide variety of applications. In this open-ended question, you are free to choose any dataset and any particular tool associated with clustering, association rule learning and fuzzy logic (e.g. K-means clustering, Apriori algorithm, etc.). Imagine that you are a data analyst who is responsible for mining hidden patterns from datasets. Specifically, you are required to:

1. Suggest an applicable area where a tool can be applied –
    a. Brainstorm and suggest a real-life scenario where the data mining or machine learning tools taught in the module can be applied to generate insights from a dataset (see Remark (i))

b. Discuss why this tool is selected for this scenario

2. Identify and discuss a dataset – The dataset can be an open dataset from various open sources (see Remark (ii)). Alternatively, if no dataset is available for the application/scenario you brainstormed in Step 1, you may create a virtual dataset, which contains at least 50 data objects/observations so that a data mining tool can further be applied.
    a. Introduce the dataset: what the dataset contains (types of data, data variables, etc.), the source of the dataset (if obtained from a database website) (see Remark (iii))
    b. the potential insights that can be generated through applying the selected tool to mine the data from the selected dataset

3. Apply the selected tool on the dataset you picked/created in Step 2 –
    a. Discuss and interpret the data mining results after applying the tool
    b. Discuss the novelty and significance of this application

Remark:

(i) Your chosen application should be new. In other words, you should not select an application area exactly the same as any existing ones in the literature. This exercise is to allow you to think out of the box to identify any promising areas of a data mining and machine learning tool.

(ii) There are numerous open database websites that enable you to identify and retrieve a free, public dataset. They include, but not limited to, **Google Dataset Search**, **Kaggle**, **Datahub.io**, **UCI Machine Learning Repository**, **Earth Data**, **Global Health Observatory Data Repository**. Try to google the rest of them and identify a dataset, or create a virtual dataset!

(iii) Your submission should include the dataset (in excel format) you picked.