COMP7103 Assignment 1

Due date: Oct 27, 2025, 11:59pm

Note: This is a written assignment. You are expected to present your answer in written form unless otherwise specified in the question. Solutions in the form of a program will not be graded.

Question 1 Data preprocessing and distance measure [12%]

Consider the dataset[1] in **Table 1** which shows the average monthly electricity consumptions from several areas of Laâyoune, Morocco, for the period of October 2023 to March 2024. The dataset includes 5 attributes: $Zone1$, $Zone2$, $Zone3$, $Zone4$, and $Zone5$.

Record	Year	Month	$Zone1$	$Zone2$	$Zone3$	$Zone4$	$Zone5$
R_1	2023	10	79	108	164	124	124
R_2	2023	11	78	159	156	126	125
R_3	2023	12	78	104	154	109	129
R_4	2024	1	67	103	155	111	129
R_5	2024	2	68	100	156	110	130
R_6	2024	3	67	95	142	111	127

Table 1 Electricity consumption data set

a) [7%] Using cosine similarity as the similarity measure, find the two consecutive months that have the most similar average electricity consumption across all zones. Show all calculated similarity values between each pair of consecutive months.

b) [5%] Calculate the Pearson's correlation between $Zone4$ and $Zone5$. Based on your calculation, comment on the nature and strength of the linear relationship between these two attributes.

Question 2 Metric Axioms [20%]

Consider a text dataset extracted from a social media platform that analyzes comments on the platform. Comments are normalized in various ways so that each comment is represented as a sequence of word tokens.

Define a distance measure $d(p, q)$ that computes the edit distance between two sequences of word tokens p and q. This measure finds the minimum cost of a set of operations that covert p to q. The operations considered are:

Operations	Cost	Description	Example
Insert	$C_i > 0$	Insert a token	love you → i love you
Delete	$C_d > 0$	Delete a token	i love you → love you
Substitute	$C_s > 0$	Replace a token	i love you → i hate you

[1] Extracted from https://archive.ics.uci.edu/dataset/1158/high-resolution+load+dataset+from+smart+meters+across+various+cities+in+morocco

For example, the distance between p_1 = i always love you and q_1 = love you always is the minimum of $C_i + 2C_d$ and $C_d + 3C_s$ since the minimal set of operations that convert p_1 to q_1 involves either one insertion and two deletions, or one deletion and three substitutions.

Validate whether the distance measure d satisfies each of the properties of a metric (Positivity, Symmetry, and Triangle Inequality). If a property is always satisfied, explain why. If a property is not always satisfied, provide an example illustrating when it fails.

Question 3 Decision Tree Classifier [48%]

Consider the dataset[2] in **Table 2** for a classification dataset predicting the success of crowd funding projects based on numerical attributes such as promotional video length (Vi) and number of related images (Im).

Record	Vi	Im	Class
1	5	7	T
2	42	8	F
3	43	2	F
4	60	8	T
5	60	12	T
6	67	1	T
7	67	6	T

Record	Vi	Im	Class
8	99	2	F
9	115	16	F
10	185	16	T
11	203	0	F
12	215	6	F
13	271	13	F
14	486	7	F

Table 2 Dataset for classification

a) [18%] Build a decision stump to predict the class attribute by selecting the best binary split point. Use the GINI Index as the impurity measure. Show all your steps.

b) [18%] Discretize attributes Vi and Im using the following rules:

Attribute	Range	Label
Vi	$[0,64)$	$Short$
Vi	$[64,194)$	Mid
Vi	$[195,\infty)$	$Long$
Im	$[0,7.5)$	Low
Im	$[7.5,\infty)$	$High$

Then, build a decision tree with binary splits only, to predict the class attribute. Use entropy as the impurity measure and set the pre-pruning criterion to stop splitting if the information gain is less than 0.1. Show all your steps

c) [12%] Evaluate the decision tree in Figure 1 using the dataset in **Table 2** by constructing the confusion matrix. Calculate the precision, recall, and F-measure with respect to class T.

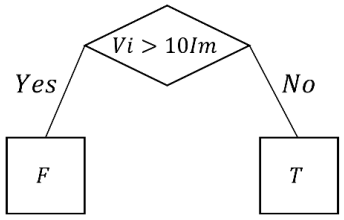

Figure 1 decision tree

[2] Extracted from https://archive.ics.uci.edu/dataset/1025/turkish+crowdfunding+startups
(platform_adi, kategori, fon_sekli, tanitim_videosu, web_sitesi, sosyal_medya)=(fongogo, diğer, ya hep ya hiç, var, var, var)

Question 4 Classification in Weka [10%]

Download the **Gallstone** dataset from https://archive.ics.uci.edu/dataset/1150/gallstone-1 and read the description. A copy of the dataset (`gallstone.xlsx`) and an extract of the description are also available on Moodle.

Pre-process the dataset by extracting the following attributes from it:

Attribute	Type	Description
Gender	Nominal (Male/Female)	Attribute "**Gender**" in the dataset, where **0** = "**Male**" and **1** = "**Female**".
Comorbidity	Nominal (Yes/No)	Attribute "**Comorbidity**" in the dataset, where **0** = "**Yes**" and all other values = "**No**".
CAD	Nominal (Yes/No)	Attribute "**Coronary Artery Disease**" in the dataset, where **0** = "**No**" and **1** = "**Yes**".
Hypothyroidism	Nominal (Yes/No)	Attribute "**Hypothyroidism**" in the dataset, where **0** = "**No**" and **1** = "**Yes**".
Hyperlipidemia	Nominal (Yes/No)	Attribute "**Hyperlipidemia**" in the dataset, where **0** = "**No**" and **1 =** to "**Yes**".
DM	Nominal (Yes/No)	Attribute "**Diabetes Mellitus**" in the dataset, where **0** = "**No**" and **1** = "**Yes**".
HFA	Nominal (Yes/No)	Attribute "**Hepatic Fat Accumulation**" in the dataset, where **0** = "**No**" and all other values = "**Yes**".
Class	Class label (Yes/No)	Attribute "**Gallstone Status**" in the dataset, where **0** = "**No**" and **1** = "**Yes**".

Answer the following questions.

a) [4%] Prepare an ARFF file for the pre-processed dataset. Show all sections before the "**@DATA**" section in the ARFF file. You do not need to submit the ARFF file itself.

b) [6%] Download **4b.model** from Moodle. Load your ARFF file from part a) into Weka, then in the "**Classify**" tab, right-click on the "Result list" section to load **4b.model** into Weka. Generate classifier output using your ARFF file with this model, use 10-fold cross-validation as the test option.

   1) Give the final model (the decision tree) built.

   2) Give the confusion matrix.

   3) Examine the "Classifier output" in Weka. List, in the order of execution, the algorithm(s) used to build the final model.

Question 5 Classification in Python [10%]

The file "gallstone`.csv`" is a CSV file containing the **Gallstone** dataset from Question 4. Write a Python program that:

1) Reads "`gallstone.csv`" from the current folder, then builds a decision tree for the dataset using **DecisionTreeClassifier** from **scikit-learn**. The program should:
   - Use all instances of data as training data
   - Use only the 7 attributes (excluding the class label) listed in Question 4, along with the "**Gallstone Status**" attribute as the class label.
   - Limit the decision tree to at most 7 leaf nodes.

2) Evaluate the model built in part 1) using **cross_validate** (or similar) in **scikit-learn**. Perform a 10-fold cross validation and compute the average accuracy.

Answer the following questions.

a) [2%] Give the Python code segment that prepares the dataset before building the decision tree.

b) [2%] Give the Python code segment that builds and visualize the decision tree in text form.

c) [2%] Give the visualized decision tree in text form.

d) [2%] Give the python code segment that performs the cross-validation and calculates the average accuracy.

e) [2%] List the accuracy scores for each fold of the cross-validation, as well as the overall average accuracy.