COMP7103 Assignment 2

Due date: Nov 26, 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 Classification [50%]

The **PhiUSIIL Phishing URL Dataset** consists of 134,850 legitimate URLs and 100,945 phishing URLs. The dataset contains several attributes derived from the URL and/or the website content. Both the dataset and its description are available on Moodle.

Link to dataset: https://archive.ics.uci.edu/dataset/967/phiusiil+phishing+url+dataset

Answer the following questions.

a) [35%] Using Weka or any other suitable tools, create a classification model for the class label "label".

   Write a report describing your data mining process. This should include details of the steps followed to create the model, such as data preprocessing, attribute selection, parameter tuning, construction of training and test data, model building, and evaluation. You may omit any steps that are not applicable.

   Reports will be assessed based on the following criteria:

   - **Completeness**: The report must detail the data mining process enough for reproducibility and verification.
   - **Correctness**: The process should be conducted accurately with reference to the dataset description. All models should be properly evaluated.
   - **Quality of the data mining process**: Explore different approaches and iteratively improve the model.
   - **Quality of the report**: The report should be clear, concise, well-organized, and not more than three A4 pages. Appendices can be included for supplementary information but will not be graded.

   Please clearly indicate your final classification model in your report.

b) [15%] Compare, with supportive figures, the usefulness of the derived attributes $CharContinuationRate$, $URLTitleMatchScore$, $URLCharProb$, and $TLDLegitimateProb$ in creating a classification model.

Question 2 Association analysis [30 marks]

a) **Table 1** shows the combinations of 10 orders in a restaurant. In this dataset, $Soup$, $Salad$, $Main$, $Dessert$, and $Drink$ are considered items in the transactions. Perform the following association analysis, assuming the minimal support is 4.

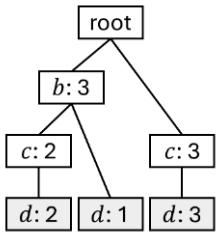
Transaction	Order
1	$\{Salad, Main, Dessert\}$
2	$\{Soup, Salad, Main\}$
3	$\{Main, Dessert, Drink\}$
4	$\{Salad, Main, Dessert\}$
5	$\{Main, Dessert, Drink\}$
6	$\{Salad\}$
7	$\{Soup, Salad, Main, Drink\}$
8	$\{Soup, Main\}$
9	$\{Salad, Dessert, Drink\}$
10	$\{Main\}$

Table 1 Restaurant order dataset for association analysis

i. **Construct an FP-tree** for the orders. Items with the same frequency should be ordered alphabetically. Clearly show the item and its frequency at each node of the tree.

ii. Construct a **conditional FP-tree** for item $Dessert$. You may present the conditional FP-tree either graphically or textually, as illustrated in **Figure 1** below.

Graphical:



Textual:
$\{(b: 3, c: 2), (c: 3)\}|d$

Figure 1 Example conditional FP-tree for item d and the corresponding representation

b) **Table 2** shows the same dataset with associated customer IDs for sequence analysis. Using the **Generalized Sequential Pattern** (GSP) algorithm, find all frequent sequences with a minimal support of 3. Show all frequent sequences, and detail the steps of Candidate Generation, Candidate Pruning, and Support counting.

Customer	Day	Order
1	1	$\{Salad, Main, Dessert\}$
1	4	$\{Soup, Salad, Main\}$
2	2	$\{Main, Dessert, Drink\}$
2	3	$\{Salad, Main, Dessert\}$
2	5	$\{Main, Dessert, Drink\}$
3	2	$\{Salad\}$
4	1	$\{Soup, Salad, Main, Drink\}$
4	2	$\{Main\}$
4	4	$\{Soup, Salad, Dessert, Drink\}$
5	5	$\{Main\}$

Table 2 Restaurant order dataset for sequence analysis

Question 3 Cluster analysis [20 marks]

Table 3 shows the location of 5 data objects.

Object	Location
O_1	$(0, 2)$
O_2	$(0, 5)$
O_3	$(2, 7)$
O_4	$(4, 0)$
O_5	$(4, 6)$

Table 3 Locations of 5 data objects

Perform hierarchical clustering using the **Distance Between Centroids** as the inter-cluster similarity measure. Use **Euclidean distance** as the distance metric.

Show all your steps, including the calculation of distances and cluster merges, and draw the resulting dendrogram with all merge points marked on the y-axis.