
	<p style="text-align: center;">Hashemite University Prince Al-Hussein Bin Abdullah II Faculty for Information Technology Department of Information Technology</p>	
---	---	---

For Instructor Use	
Course Name	Introduction to Data Science
Course ID	2010042210
Academic Year	2023/2024
Semester	Second Semester
Assignment	1
Due Date	17/5/2024

For Student Use	
Student Name	
Student ID	

For Instructor Use		
CLO	Max. Score	Student Score
	20	
Total Score	20	

The dataset includes 300 data objects and each data object is described by 10 attributes.

Tasks

Your tasks include:

A. Initial data exploration

- A1. Identify the type of each attribute (nominal, ordinal, interval or ratio).
- A2. Identify the values of the summarising properties for each attribute including frequency, location and spread [e.g. value ranges of the attributes, frequency of values, distributions, medians, means, variances, percentiles, etc. - the statistics that have been covered in the lectures and materials given). Where necessary, use proper visualisations for the corresponding statistics.
- A3. Using Weka, explore your data set and identify any outliers, clusters of similar instances, "interesting" attributes and specific values of those attributes. Note that you may need to 'temporarily' recode attributes to numeric. In the report include the corresponding snapshots from the tools and explanation of what has been identified there. Hint: *please consider scatter plots*.

In the assignment report for each of these techniques you need to illustrate your steps. In your Excel workbook file place the results in separate columns in the corresponding spreadsheet.

B. Data pre-processing

Perform each of the following data preparation tasks (each task applies to the "**Scoring-Dataset**" data):

B1. **Use the following techniques to normalise** the "Age" attribute:

- min-max normalization to transform the values onto the range [0.0-1.0].
- z-score normalization to transform the values.

B2. **Discretise** the Age attribute into the following categories:

- Teenager = 1-16;
- Young = 17-35;
- Mid_Age = 36-55;
- Mature = 56-70;
- Old = 71+.

Provide the frequency of each category in your data set.

B3. **Convert the "Gender" variable into binary variables** [with values "0" or "1"].

In the assignment report provide explanation of the pre-processing procedures. In your Excel workbook file place the results in separate columns in the corresponding spreadsheet.

The deliveries include:

- A report, In the report include a section (starting with a section title) for each of the tasks in this assignment including tasks A and B.
- An Excel workbook file with individual spreadsheets for each task (spreadsheets should be labelled according to the task names, for example, "B"). Each of the results of parts 1 through 3 in task B should be presented in a separate spreadsheet (and respectively table in the assignment report).

Assessment criteria	Task	Percentage
Correctness of the identification of the attribute types.	Task A1	(20%)
Depth of data understanding - how comprehensive are the explanations of your explorative results, appropriateness of illustrations and data visualization.	Task A2 and A3	(30%);
Correctness of the pre-processing procedures, results and explanation of the steps.	Task B	(50%)