| | | |
|---|---|---|
| | Hashemite University<br>Prince Al-Hussein bin Abdullah II Faculty for Information Technology<br>Department of Computer Information Systems | |

| For Instructor Use | |
|---|---|
| **Course Name** | Data Mining |
| **Course ID** | 151002351 |
| **Academic Year** | 2023/2024 |
| **Semester** | First Semester |
| **Assignment** | 1 |
| **Due Date** | 12/12/2023 |

| For Student Use | |
|---|---|
| **Section ID** | |
| **Student Name** | **Student ID** |
| | |
| | |
| | |
| | |
| | |

| Max. Score | Student Score |
|---|---|
| **10** | |

# Assignment-1: "Data exploration and preparation"

### *Team:*

You should have four members in your group.
Your dataset will be selected according to group IDs:

- If the majority of students' IDs are even numbers: **Customer_Churn.csv**
- If the majority of students' IDs are odd numbers: **Heart 1.csv**
- Otherwise: **Heart 2.csv**

The description of each dataset is presented in the same folder along with the dataset.

### *Software:*

The suggested software to be used is Weka 3.8. However, feel free to try any ideas you may have to tackle the problem with any other software.

### *Report:*

Your final report must clearly document the following aspects.

### *Tasks*

Your tasks include:

## Data Understanding

1. For each attribute find the following information.

(a). The attribute type, e.g. nominal, ordinal, numeric.

(b). Percentage of missing values in the data.

(c). Max, min, mean, standard deviation.

(d). The type of distribution that the numeric attribute seems to follow (e.g. normal).

(e). Where necessary, use proper visualizations for the corresponding statistics.

(f). Are there any outliers for the attribute under consideration? If you suspect of the existence of outliers for an attribute, you may consider the possibility of using box plots for outlier detection.

2. Switch to the Visualize tab on the upper part of the screen in Weka to visualize 2D-scatter plots for each pair of attributes.

(a). Does any pair of attributes seem to be correlated?

(b). Which attributes seem to be the most/least linked to the class attribute? Summarize in a table your findings concerning the predictive value of each attribute with respect to the class attribute.

(c). Investigate also possible multivariate associations of attributes with the class attribute, i.e. study scatter plots of two attributes X and Y and try to identify possible high(low)-class areas (if any). If you find high(low)-churn areas in any scatter plot then quantify the churning rate in these

areas with respect to the entire data set.

3. Are there any variables that can be eliminated? Justify your answer and motivate the possible benefits of doing so (if any).

(a). Compare your conclusions with the results obtained by using some attribute selection filters in Weka. Do not forget to indicate which filter(s) you have used and to give a brief description.

## Data preprocessing

The second step is to preprocess the data such that the transformed data is in a more suitable form for the mining algorithms. You can find below some aspects you should consider. (*Obviously, you may consider other aspects of data preprocessing, such as creating new attributes from the existing ones.*)

1. Attribute selection. Select one or two subsets of attributes with good predicting capability and motivate your choices (recall your conclusions for points 2 and 3 above).
2. Handling missing values.
3. Eliminating outliers.
4. Discretization of numeric attributes. Do not forget to indicate which discretization method did you use if discretization is used.
5. Normalization. Motivate the need for normalization (if needed at all).

| **Assessment criteria** | **Percentage** | **Mark** |
|---|---|---|
| Correctness of the identification of the attribute types. | (20%) | 2 pts |
| Depth of data understanding - how comprehensive are the explanations of your explorative results, appropriateness of illustrations and data visualization. | (40%); | 4 pts |
| Correctness of the pre-processing procedures, results, and explanation of the steps. | (40%) | 4 pts |

**Deadline: Tuesday 12/12/2023**