ASSIGNMENT 1: DATA PREPROCESSING & VISUALIZATION

Preparation

Practice data visualization in PowerBI/TABLEAU/Python and basic data exploration in WEKA.

A. (20 points) Identify and characterize a dataset

a. You should select 2 datasets from publicly available data at the following sites: (Try to find a dataset with at least 15 attributes and 1000 instances)

- www.data.gov
- www.data.gov.au
- http://www.kdnuggets.com/datasets/index.html
- https://www.kaggle.com/datasets
- https://data.sa.gov.au/data/dataset
- http://portal.govhack.org/datasets.html

b. Briefly describe what the dataset is about and describe the dataset (e.g. number of tables, number of instances and attributes, etc.). Discuss briefly the challenges in "mining" this dataset.

c. Discuss potential data mining applications for the dataset. Name two types of data mining techniques (classification, clustering, etc.) you think would be relevant and discuss the potential mining results. E.g. if you think clustering is relevant, describe what a likely cluster might contain and what the real-world meaning would be.

B. (40 points) Data exploration and pre-processing in WEKA

a. Select 2 attributes which you think are important from each dataset and explain why by discussing appropriate measures of the central tendency and dispersion for the attribute. For example, you can analyse the attribute by computing the mean, median, mode, range, quartiles, and variance for the attribute (boxplots).

b. Discuss data quality issues of the datasets. Are there (potential) problems with certain data attributes? How will you fix these data quality issues?

c. Discuss any two data pre-processing techniques that are likely required for the dataset (give reasons why). For example, you may need to transform the scale of some attributes or reduce the dimensionality of the dataset. Apply any 3 filters from WEKA or Python to attributes from your dataset and discuss why you picked the attributes and the results.

C. (40 points) TABLEAU

a. Load any **single** dataset into PowerBI or TABLEAU. Explore the data through different charts and identify any 5 interesting insights from the dataset. An example of an insight could be, “Female customers from the north-eastern region in the age group of 28-37 spend at least $45 between June-July”.
b. Create 3 interesting visualizations that provide useful insights from your chosen dataset and select the appropriate display type for the data variables you have decided to include.
c. Create a dashboard with the visualizations that you created and make sure that the charts in the dashboard are related to each other.

Submission

- Submit your reports in pdf to Canvas which include screenshots of your visualizations.
- This will be a turnitin submission.