CSCI316 – Big Data Mining Techniques and Implementation
Laboratory 2 (Assessed)
Autumn 2023

6 Marks
Deadline: Refer to the submission link of this assignment on Moodle

One task is included in this laboratory. The specification of each task starts in a separate page.

You must implement and run all your Python code in Jupyter Notebook. ***The deliverables include one Jupyter Notebook source file (with .ipybn extension) and one PDF document for each task.***

To generate a PDF file for a notebook source file, you can either (i) use the Web browser's PDF printing function, or (ii) click "File" on top of the notebook, choose "Download as" and then "PDF via LaTex".

The submitted source file(s) and PDF document(s) must show that all of your code has been executed successfully. Otherwise, they will not be assessed.

This is an individual assessment. Plagiarism of any part of this assessment will result in having 0 mark for this assessment and for all students involved.

The correctness of your implementation and the clearness of your explanations will be assessed.

Task Specification

(6 marks)
Data set: The Abalone Data Set
(Source: https://archive.ics.uci.edu/ml/datasets/abalone)

Data set information
These data consisted of 4,177 observations of 9 attributes, detailed as follows.
Name / Data Type / Measurement Unit / Description

Sex / nominal /-- / M, F, and I (infant)
Length / continuous / mm / Longest shell measurement
Diameter / continuous / mm / perpendicular to length
Height / continuous / mm / with meat in shell
Whole weight / continuous / grams / whole abalone
Shucked weight / continuous / grams / weight of meat
Viscera weight / continuous / grams / gut weight (after bleeding)
Shell weight / continuous / grams / after being dried
Rings / integer /-- / +1.5 gives the age in years

Objective
Use Pandas in Python to clean and pre-process this dataset. No ML library (e.g., Scikit-Learn) is allowed to use.

Requirements

(1) Perform z-score normalization of the values in the attribute Length. Show the mean and variance of the resulted values.
(2) Create five bins for the attribute Diameter such that the bins contain (approximately) equivalent numbers of records. Show the resulted bins.
(3) Apply one-hot-encoding to the Sex attribute. Show the *unique* one-hot-encoding values.
(4) Find out and rank the correlations between the attribute Rings and other attributes with *continuous* values.
(5) Define at least one new attribute based on existing attribute, and justify your reason (e.g., in terms of correlation).

Note. For the requirements (1) – (3) and (5), the new columns should be appended to the existing Pandas dataframe.

Deliverables
- A Jupiter Notebook source file named `your_name_lab2.ipybn` which contains your implementation source code in Python
- A PDF document named `your_name_lab2.pdf` which is generated from your Jupiter Notebook source file, and which includes clear and accurate explanation of your implementation and results.