CST8390 Assignment 1

Due: Jun 13, 2021 at 11:59 PM Sharp!!!

(Late submissions will not be accepted)

Goal: The goal of this lab is to explore and analyze **2 datasets** and select one of them, and then preprocess & clean the selected one, find statistics, view visualization and perform classification using kNN with various settings.

Steps

1. Select 2 datasets from the following list of 8 datasets:

Default Task - Undo	Name	Data Types	Default Task	Attribute Types	# Instances	# Attributes	Year
Classification (26)	Breast Cancer Wisconsin (Original)	Multivariate	Classification	Integer	699	10	1992
Regression (7)	Breast Tissue	Multivariate	Classification	Real	106	10	2010
Clustering (2)	Breast Cancer Coimbra	Multivariate	Classification	Integer	116	10	2018
Other (1)	ILPD (Indian Liver Patient Dataset)	Multivariate	Classification	Integer, Real	583	10	2012
Attribute Type - Undo	Fertility	Multivariate	Classification, Regression	Real	100	10	2013
Categorical (6)	Algerian Forest Fires Dataset	Multivariate	Classification, Regression	Real	244	12	2019
Numerical (26)	Heart failure clinical records	Multivariate	Classification, Regression, Clustering	Integer, Real	299	13	2020
Mixed (7)	HCV data	Multivariate	Classification, Clustering	Integer, Real	615	14	2020
Data Type							
Multivariate (24)							
Univariate (2)							
Sequential (0)							
Time-Series (0)							
Text (0)							
Domain-Theory (0)							
Other (0)							
Area - Undo							
Life Sciences (26)							
Physical Sciences (6)							
CS / Engineering (10)							
Social Sciences (4)							
Business (5)							
Game (0)							
Other (8)							
# Attributes - Undo							
Less than 10 (11)							
10 to 100 (26)							
Greater than 100 (5)							
# Instances - Undo							
Less than 100 (3)							
100 to 1000 (26)							

To see the details of these datasets, navigate to https://archive.ics.uci.edu/ml/datasets.php. Apply same filters as mine (highlighted in yellow - Classification, Numerical, life sciences, 10 to 100 attributes, 100 to 1000 instances, sort based on the number of attributes) on the left-hand side to narrow down the list of datasets. Consider the first 8 with less than 15 attributes, which is in the above picture and select 2 datasets.

2. Explore and analyse data and provide a brief description (10 lines) about each dataset. Then **select one** of them and specify the reason for your selection (10 lines). From the description, you may be able to find the performance of some classification methods applied on those datasets. Altogether, you need to write 3 paragraphs (one paragraph each for each dataset, third paragraph with reason for selection).
3. Thoroughly analyze your data to have a clear understanding of your data and their attributes and types.

4. Load your file to Weka. Double check the type of your attributes in Weka. If they are not as expected, apply filters to convert them to the right types.

5. Tabulate statistics and counts (whichever apply) for each attribute. Provide that information in **one** table (like the summarized table for Iris in Lab 1).

6. Perform preprocessing, data cleaning, remove duplicates, handle missing information etc. Specify which all filters you applied and the corresponding reason.

7. Once you are done with data preparation, navigate to Visualize tab to visualize your data. Include 3 interesting charts in your submission. You need to specify how those charts are interesting (you may have clusters, or classes are separable, or classes have too much of overlapping etc.) (at least 5 lines)

8. Now, perform kNN classification with 10-fold cross validation for various k's ranging from 3, 5, …, 11 and tabulate the percentage of correctly classified instances. For the worst and best k's (in terms of accuracy) **only**, tabulate True Positive Rate (TPR) and False Positive Rate (FPR) and the number of misclassifications (from confusion matrix).

9. Repeat step 8 for percentage split of 70%.

10. Repeat step 8 for 1 other seed.

11. Take one instance as a test instance and show the calculations in Excel to find the class of that instance by applying 5NN. You need to explain the process and include the screenshot of the excel file (of the final stage where you make the prediction of the class of the test instance) in the report. Also, include the excel file in the zipped folder. This step has 5 marks. So make sure that you include every detail in the report (If you forgot the process, refer to the video where I showed this process for Iris file).

Submission Details:

This is a partner group assignment. Assignment should have a **cover page** with the names (Last name, first name) and student numbers. Create a zipped folder with the name,

LastNameFirstStudent_FirstNameFirstStudent_LastnameSecondStudent_FirstNameSecondStudent.zip. Include your **assignment, Weka files and models and the excel file** in the zipped folder. Upload the zipped folder to Brightspace. There will be mark deduction if you are not following the submission requirements.

Marks: (25 marks in total)

Step 1-2: 4 marks

Step 5: 2 marks

Step 6: 3 marks

Step 7: 3 marks

Step 8: 2 marks

Step 9: 2 marks

Step 10: 2 marks

Step 11: 5 marks

Submission (correct name, zipped folder with all required contents): 2 marks

This assignment is worth 10% of your term mark. So, each step should be explained in detail. Just tables and number are not enough. Prepare your assignment in a professional report style.