CST8390 Assignment 1

Due: Jan 28, 2023 at 11:59 PM Sharp!!!

(Late submissions will not be accepted)

Goal: The goal of this lab is to explore and analyze **both 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. **Follow CRISP-DM steps in this assignment.**

Steps

1. https://archive.ics.uci.edu/ml/datasets/Yeast
2. https://archive.ics.uci.edu/ml/datasets/ILPD+%28Indian+Liver+Patient+Dataset%29

Introduction

Explore and analyse data and provide a brief description (10 lines each) about each of the datasets. Then **select one** of them and specify the reason for your selection (10 lines). From the description and the papers given in the UCI website, you may be able to find the performance of some classification methods applied on those datasets. You can include those results too in your description. Altogether, you need to write 3 paragraphs (one paragraph each for each dataset, third paragraph with reason for the final selection).

Data Understanding

Thoroughly analyze your data to have a clear understanding of your data and their attributes and types. Along with the description of the dataset, tabulate all attributes, its meaning and the expected data type.

Data Preparation

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. Tabulate statistics and counts (whichever apply) for each attribute. Provide that information in **one** table (like the summarized table for Iris in Lab 1).
Perform preprocessing, data cleaning, remove duplicates, handle missing information etc. Specify which all filters you applied and the corresponding reason. Every step of Data Preparation phase of CRISP-DM should be done and reported. 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)

Modeling & Evaluation

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.

Only for the worst and best k's (in terms of accuracy), tabulate True Positive Rate (TPR) and False Positive Rate (FPR) and the number of misclassifications (from confusion matrix). You are applying distance-based approach here. So, make sure that your dataset is prepared for that.

Repeat the classification with a percentage split of 70% and tabulate its results. Now, repeat 10-fold cross validation with a different seed and tabulate the results. Make sure to include the seed value in your description.

Manual Testing

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 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. This assignment should follow a professional writing style. The report should have table of contents, images, etc, sections titled introduction, Data Understanding, Data Preparation, Modeling and Evaluation, Manual Testing, Conclusion, References etc.

Marks: (25 marks in total)

Introduction: 3 marks

Data Understanding: 3 marks

Data Preparation: 6 marks

Modeling & Evaluation: 6 marks

Manual testing: 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.