COMP-5411 FA & FB
Fall 2019
Assignment 1

This is NOT a group assignment

Due Date: Wednesday, October 9, 11:59 pm
Late submission is NOT allowed.

1) Write a program that will open a data file ("data.csv") containing multiple rows and columns (no missing value). Assume both your program and the data file are in the same directory so that explicitly mentioning the path is not required while loading the file. Each column represents a feature and each row is an instance. The dataset should have at least 2 continuous features, 1 categorical feature and 100 instances. Your program then should randomly create missingness in each feature. For each feature implement the following:
    a. Randomly remove 5% of values and mark the removed values as missing
    b. Employ 3 nearest neighbor imputation methods (1-NN, k-NN and weighted k-NN) [hint: distance from complete cases]. Choose any k>=5. For continuous features, use 2 different distance measures and for categorical, use 1 distance measure. ***You cannot use any existing library to calculate distance or to implement imputation.***
    c. For each feature, calculate the accuracy of imputation. You will define your own accuracy measures for both continuous and categorical features [hint: how different the imputed values are from the original values].
    d. For each feature, conduct above steps (a-c) on the original data.csv file.

    Repeat above three steps (a-d) where you remove 10% and 20% of values from each feature for creating missingness. The output (imputation accuracy) of the program should be written on "results.csv". the output file will list the features and corresponding imputation accuracy for all possible combinations of methods, such as: 1-NN & distance measure 1, 1-NN & distance measure 2, etc.

2) Repeat the experiments in Step 1 where distance is calculated after scaling the continuous features using 2 different feature scaling methods of your choice. You are allowed to use existing libraries for feature scaling. Note that, imputation should be done using original values, not the scaled ones. Add imputation accuracies from the experiments using scaled values to the file "results.csv".

3) Write a report which will have the following 3 sections:

    a. Data: Describe features, instances and the source.
    b. Methods: Explain the 3 imputation methods, 3 distance methods, 2 feature scaling methods and imputation accuracy measures used in your experiments.
    c. Tools: Mention which language (R or Python), version and IDE you used for implementation. Mention if any library needs to be installed for running your code. Provide installation instructions.
    d. Results: Present imputation accuracies for each feature for all 3 different missing percentages (5%, 10%, and 20%) and different imputation methods. Use tables and graphs for presenting the results. Also, present comparative analysis of imputation

performance when original values are used for distance calculation vs. when scaled values are used. All of these should be reported for two different distance measures.

Submit following files:

1) Source code: Add enough comments in the code explaining your program. Commands to load necessary libraries must be included in your code.
2) data.csv
3) results.csv
4) Report (both word and pdf document)