Introduction to Data Science

Final Term Project - Summer 22-23

Project Description

The goal of this project is to apply K-Nearest Neighbor (KNN) classification algorithm and its relevant tasks to a supervised dataset. You need to select an appropriate dataset where it is possible to apply all the required tasks. The required tasks are defined at the end of this document. However, you are not allowed to select Iris and USArrest datasets. You may explore the following repositories to find your dataset. **You are required to select a unique dataset that means your dataset will not match with any of your course mates.** For this purpose, please entry your dataset name in the record sheet shared in MS Teams and before your entry, check the entry of others.

- https://archive.ics.uci.edu/datasets
- https://www.kaggle.com/datasets

Project Deliverables

- Submit the implemented R program (R file or Text file) in the MS Teams. During the VIVA session, you will bring this implemented program and we may ask you to execute the program.
- Submit the report in the MS Teams. See the instruction section below for the report details. **Please bring the printed copy of the submitted report during the VIVA session.**

Instructions

- The submission deadline for all deliverables is **August 15, 2023** (you must submit the assignment within **11:59 PM**). There will be a penalty for the late submission.
- At the beginning of the report (after the cover page), write a short note about the dataset, i.e., the source of the dataset, details of the attributes, etc.
- For each implemented code segment in the R program, provide the code and its output along with their description in the report. In the description part, only write the content (do not write unnecessary content) that is sufficient to understand the code and its output.
- **Comments are not allowed in the R program.**
- The following tasks need to be completed for this project.
    - Select a dataset where it is possible to apply KNN classification.
    - Depending on the selected dataset, perform the data preparation steps that are required to apply KNN such as convert all data to numeric, handle missing values and normalization. Do only the tasks that are required for your dataset.
    - Using the correlation technique (pearson's correlation coefficient), find the important attributes that will be employed for the classification task while applying KNN algorithm. Remove only those attributes that do not have any relation at all with the class/label (attribute).
    - Apply KNN classification algorithm to the dataset that contains only the important attributes selected using the correlation technique.
    - Find the predictive accuracy of the KNN classifier. Report the accuracy using two approaches:
        - Dividing the data into training and test set
        - 10-fold cross validation
    - Generate the confusion matrix for your dataset using the KNN classifier that you built. Also report the Recall and Precision value of your classifier.