Total Marks: 50

Assignment-3

Data Mining

PGDDS&AI

Instructions

1. Do the code using python; it can be used in a Jupyter notebook.
2. Create a Github private repo where all your assignments and projects would be stored. At an opportune time, you would be asked to share your repo with our evaluation team.
3. Your submission should be sufficiently original to be considered for evaluation.
4. Submissions would include source code, data availability at Github, and Google classroom submission of the markdown pdf (do not submit any other format on Google classroom).
5. You can use any library, no need to code from scratch.

Submission date and time

May 31, 2021, by 6 PM.

Please follow the steps below to complete your assignment:

1. You need to download 'Stroke Prediction Dataset' data using the library Scikit learn; ref is given below. [5]
2. Divide the data randomly in training and testing with a 7:3 ratio 100 times, perform the following tasks with training data and test the performance on testing data. Testing data should remain unseen for all steps.
    a. Apply one of the best-known imputation methods to handle the missing/infinite values and state the significance of the used method if required. [5]
    b. Visualize the data in 3-D scatter plot and write the inferences, How the data look like. [5]
    c. Make a boxplot for each feature and highlight the outlier, if any, then remove the outlier, again visualize the data in 3-D scatter plot to show the outlier effect and write the inferences. [5]
    d. Normalized the data if required, and write a note for what, why and how you performed normalization.[5]
    e. Balance the data if required; you may increase the sample using upsampling if needed.[5]
    f. Perform at least three clustering methods with varying cluster sizes. Perform any three best-known methods to find out correct cluster numbers for each method; how you finalized this cluster number.[10]
    g. Perform at least three supervised methods for classification, and report at least three performance metrics out of (accuracy, precision, Cohen's kappa, F1-score, MCC, sensitivity and specificity) with proper reason. [10]

Ref:

1. https://www.kaggle.com/fedesoriano/stroke-prediction-dataset

Note: References can be used only for learning purposes, not like copy-paste.