Total Marks: 20

Assignment-1

Data Mining

Instructions

1. Do the code using python; it can be used in 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

April 25, 2021, by 6 PM.

Please follow the steps below to complete your assignment:

1. You need to download 'breast cancer wisconsin' data using the library Scikit learn; ref is given below. [2]
2. Remove the missing/infinite values using the mean strategy if required. [3]
3. Visualize the data in 2-D scatter plot and write the inferences, How the data look like. [5]
4. Make a boxplot for each feature and highlight the outlier, if any, then remove the outlier, make again box plot to show the outlier effect and write the inferences. [5]
5. Normalized the data if required, and write a note for what, why and how you performed normalization.[5]

Ref:

1. https://scikit-learn.org/stable/modules/generated/sklearn.datasets.load_breast_cancer.html#sklearn.datasets.load_breast_cancer

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