Project Requirements - MLDL

Dataset Requirements:

Acquire one dataset suitable for exploratory data analysis and machine learning modeling. You can select a dataset for either a regression or classification goal. Ensure the dataset contains at least 1000 records. The dataset must include more than 5 variables.

Required Steps:

1. Import the dataset: Load the data into your analysis environment.
2. Display records: Show the first 5 and last 5 records of the dataset.
3. Identify data types: Check and note the data types for each variable.
4. Missing entries: Determine the number of missing entries per variable.
5. Duplicate records: Identify and count any duplicate records.
6. Univariate analysis: Conduct this analysis on all variables, creating appropriate visualizations.
7. Outlier detection: Use the Local Outlier Factor (LoF) method to identify outliers. Refer this link for LoF method - https://dataheroes.ai/blog/outlier-detection-methods-every-data-enthusiast-must-know/
8. Bivariate analysis: Perform at least one analysis for each of the following hypothesis tests:
    - Chi-square test to assess independence between two categorical variables.
    - Correlation analysis to examine relationships between two numeric variables.
9. Check for presence of collinearity and multi-collinearity and address it appropriately.
10. Encode the data if required (if there are categorical independent variables).
11. Split the dataset into training and testing subsets.
12. Scale the training data and use the same scaler to also scale the test data. (use scaled data for algorithms requiring scaling)
13. Perform PCA. Based on outcome, recommend if Principal Components would be useful for data preparation or not.
14. Depending on prediction goal, refer to appropriate section below:
    - **Classification**: Build models based on atleast 3 different algorithms
        i. Logistic Regression / DecisionTreeClassifier / LDA: Choose any one of these 3.
        ii. KNN, SVM, RandomForest, AdaBoost, XGBoost: For models other than KNN, tune atleast 3 hyperparameters using GridSearchCV.

    - **Regression**: Build models based on atleast 3 different algorithms
        i. Linear Regression / DecisionTreeRegressor: Choose any one of these 2. Check for validity of assumptions (LINE) if using Linear Regression.
        ii. KNN, SVM, RandomForest, AdaBoost, XGBoost: For models other than KNN, tune atleast 3 hyperparameters using GridSearchCV.

15. Check for overfitting and take steps to address it

Report:

- Dataset description: Provide a comprehensive description of the dataset.

- Code demonstration: Showcase the analytical code used for the analysis and modeling.
- Findings presentation: Present the results from the analysis and modeling.
  1. Do K-fold cross-validation for both
  2. For regression show: R^2, Adjusted R^2, RMSE, correlation matrix, and p-values of independent variables
  3. For classification show: Accuracy, confusion matrix, (Macro recall and precision)
- Conclusions: Offer insights and interpretations based on the findings.

Presentation:

- *Dataset overview: Introduce the dataset to provide context and understanding.*
- *Univariate analysis presentation:*
- *- Show univariate analysis for at least one numeric variable with appropriate visualization.*
- - Present univariate analysis for at least one categorical variable with appropriate visualization.
- Bivariate analysis presentation:
  - Exhibit at least one bivariate analysis
- Comparing the performance of your classification models.
- Comparing the performance of your Regression models.