Assignment 01 – Weightage 5%

Requirements:

- This activity is team-based (size=2 obligatory)
- Each group maintains a GitHub profile and uploads all results.
- Upload:
  - **Executed** Jupyter Notebook (content given below as "JN")
- Only one member should submit. Submit as:
  - **<Name1>(<ERP1>)_<Name2>(<ERP2>).ipynb**
- **Deadline: Tuesday**, 13th February 2024 @11.55pm

Task Specification

Create and implement an ML pipeline for ML execution (K-NN Classifier) on 5 datasets.

General Details:

- Pipeline to be used later for all hands-on activities (classification and regression).
- Pipeline will consist of a series of functions executed sequentially.
- Ideally, you need a Master function to control this workflow/pipeline execution.
- All functions to be commented.
- Use of ChatGPT allowed – but adapt its content.

Specific Details:

- Select 5 classification datasets from UCI ML Repository
  - Use filters to select the 5 datasets (*e.g., domain, # instances, # features*)
  - [JN]: Mention details of 5 datasets (URLs, Business Domain, Size)
- [JN] Create Commented Python functions for:
  - Data collection/connection
    - *Connect to data source, store in Pandas etc.*
  - Data cleaning
    - *Remove missing values, data entry errors, unnecessary columns, and rows etc.*

- Data transformation
  - *Change feature names, categorical encoding, standardization of numerical (z-score) etc.*
- Exploratory Data Analysis
  - *5-number summary (mean, mode, median, quartiles etc.)*
  - Histograms and Boxplots of important numerical variables
- Detecting outliers and anomalies
  - *Trend lines, regression, clustering – this is a bit difficult so not a stringent requirement – but doing is better than not doing it – outliers and anomalies disrupt ML.*
- Feature Engineering
  - S*election of relevant features through scoring and other methods*
- Dimensionality reduction
  - *Map dataset into a new feature space, e.g., map a 25-feature dataset into a 2-feature dataset through Principal Component Analysis*
- Manual Data Splitting and Cross-Validation
  - *Manual Train-validation-test split - Decide the percentages yourself.*
  - *CV – to determine the benefit of CV.*
- Model Selection
  - *From a given list – you can use LazyPredict to determine the possible available algorithms if you want.*
- Model Training
  - *Fitting the model to the data and tuning*
- Model Evaluation
  - Precision, recall, and F1 of each class, and overall accuracy, AUC, ROC Curve (Classification)
  - RMSE, MSE, MAE, R^2, Adjusted R^2 (Regression)
- A master function called *Master* to execute the workflow (sequence of functions) with defined input parameters.

- For this submission, use Master to execute KNN-Classifier on the workflow.
  - Use default setting of KNN hyperparameters.
  - Ftr Sel and Dim Red *not to be done in this submission.*
  - Execute Manual Splitting
  - Execute CV – to determine the benefit/disadvantage of CV vs Manual
  - Show ML Results in a structured way (Excel, Dashboard, e.g., using Flash if you want)

- Add interpretation of EDA and ML Results in notebook in separate cells
  - *This carries the most marks.*