**COGNIZANCE 2nd YEAR PRELIMINARY TASK**

**AI DOMAIN**

**Task: Data Preprocessing with the Iris Dataset**

**Objective**: The goal of this task is to understand and apply basic data preprocessing techniques using the Iris dataset. This task will help you understand the basics of data preprocessing, including handling missing data, feature scaling, encoding categorical variables, and performing exploratory data analysis.

Data preprocessing is a crucial step in the data analysis and machine learning pipeline, where raw data is transformed into a clean and usable format. This process involves several steps

**Data Cleaning:** Handling missing values, correcting errors, and removing duplicates. Techniques include imputation, where missing data is filled in, or outlier removal to eliminate anomalies.

**Data Transformation:** Scaling features to ensure they are on the same scale (e.g., normalization or standardization), encoding categorical variables into numerical values (e.g., one-hot encoding), and transforming data to meet the assumptions of a model (e.g., log transformation).

**Data Reduction:** Reducing the dimensionality of the data through techniques like Principal Component Analysis (PCA) or feature selection, which helps improve model performance and reduce computation time.

**Data Splitting:** Dividing the dataset into training, validation, and test sets to evaluate model performance and prevent overfitting.

The Iris dataset is a classic dataset in machine learning, consisting of 150 samples of iris flowers, with four features (sepal length, sepal width, petal length, petal width) and a target variable (species of iris).

**Tasks**

**Load the Dataset:**

* Load the Iris dataset from scikit-learn’s datasets module
* Display the first 5 rows of the dataset to understand its structure.

**Data Cleaning:**

**Exploratory Data Analysis (EDA):**

* Print the summary statistics of the dataset (mean, median, mode, standard deviation, etc.).
* Check for any missing values in the dataset and handle them appropriately.
* Plot the distribution of each feature using histograms.
* Visualize the pairwise relationships between features using a pair plot (or scatter plot matrix).

**Data Transformation:**

**Feature Scaling:**

* Standardize the features by removing the mean and scaling to unit variance using StandardScaler from scikit-learn.
* Alternatively, perform Min-Max scaling on the features using MinMaxScaler.

**Encoding the Target Variable:**

* Encode the categorical target variable (species) into numeric values using label encoding.

**Data Splitting:**

**Splitting the Dataset:**

* Split the dataset into training and testing sets using an 80-20 split. Use train\_test\_split from scikit-learn.

**Deliverables:**

* + A Jupyter Notebook or Google Colab Notebook containing:
  + Code for each of the tasks.
  + Comments explaining each step.
  + Plots and visualizations generated during EDA.

**References:**

Exploratory data analysis: <https://www.kaggle.com/code/imoore/intro-to-exploratory-data-analysis-eda-in-python>

Data preprocessing (Importing the dataset, identifying and handling the missing values, encoding the categorical data, splitting the dataset): <https://www.kaggle.com/code/alirezahasannejad/data-preprocessing-in-machine-learning>

Data loading: <https://youtu.be/h_NWeliQNOQ?si=-r7ZOcMmrGIqveHB>

Handling missing values: <https://youtu.be/J-KfMnhUrdA?si=1CRhXKqzL_lefQmf>

Data encoding: <https://youtu.be/r3pvRpCtaLQ?si=JAc5IQdZyIp8B7H1>

Splitting data into train and test: <https://youtu.be/Y2YoiAgG-Bk?si=ejqlmWjDTQgfl6sO>

**Bonus Task**

**Data Reduction:**

**Principal Component Analysis (PCA):**

* + Perform PCA on the Iris dataset to reduce the dimensionality to 2 components.
  + Plot the data points in the new 2D space with different colors for each species.