Lab-1

Demonstrate various data pre-processing techniques for a given dataset

Demonstrate various data pre-processing techniques for a given dataset

In this lab program, you will work on the following Data Preprocessing techniques for text or nurmerical data present in Dataframe or .csv file.

Data Preprocessing techniques:

1. Data Cleaning: Handling Missing Values, Handling categorical data, handling Outliers

2. Data Transformations: Min-max Scaler/Normalization , Standard Scaler

To write in Observation Book

Before switching on the Desktop systems in the lab, at the start of the lab, you should write following in your observation book

Write python code , consider filename as “housing.csv”

i. To load .csv file into the data frame

ii. To display information of all columns

iii. To display statistical information of all numerical

iv. To display the count of unique labels for “Ocean Proximity” column

v. To display which attributes (columns) in a dataset have missing values count greater than zero

Implementation

In your Google Co-lab account, execute sample notebook file “Lab-1-ML-DataPreprocessing.ipynb”

To Do: Implementation

Write Python code to implement the following data preprocessing techniques for Diabetes and Adult income data sets

Data Preprocessing techniques:

1. Data Cleaning: Handling Missing Values, Handling categorical data, Handling Outliers
2. Data Transformations: Min-max Scaler/Normalization , Standard Scaler

Download the following dataset files and upload in your Google Colab folder

I. Diabetes datasets

https://data.mendeley.com/datasets/wj9rwkp9c2/1

II. Adult income dataset

https://www.kaggle.com/datasets/wenruliu/adult-income-dataset

To Upload to Github

- In your github account create folder “6thSem-ML-Lab” and upload your python notebook worked on data processing techniques for Diabetes and Adult income data sets with file name as “yourUSN_Lab-1-DataProcessing. ipynb”

To write in the Observation Book

After applying data processing techniques to Diabetes and Adult income data sets, write the answer for the following questions in your observation book

For both the datasets Diabetes and Adult income

1. Which columns in the dataset had missing values? How did you handle them ?

2. Which categorical columns did you identify in the dataset? How did you encode them ?

3. What is the difference between Min-Max Scaling and Standardization? When would you use one over the other?