**ASSESSMENT OF MARGINAL WORKERS IN TAMIL NADU**

**PHASE-3**

**Name :**Hariprasath

**E mail id :**harikappini@gmail.com

**Description:**

This Python script serves the purpose of data analysis and preprocessing for a dataset stored in the CSV format. It utilizes various libraries, such as Pandas, NumPy, Seaborn, and Matplotlib, for data manipulation, visualization, and analysis.

**Required Libraries:**

- os: For interacting with the operating system.

- gc: For garbage collection purposes.

- numpy (np): For handling arrays and numerical operations.

- pandas (pd): For data manipulation and analysis.

- seaborn: For data visualization.

- matplotlib.pyplot (plt): For creating visualizations using pyplot.

- sklearn.model\_selection: For data splitting into training and testing sets.

- sklearn.impute: For handling missing values using the K-Nearest Neighbors (KNN) algorithm.

- sklearn.preprocessing: For label encoding and one-hot encoding categorical variables.

**Code:**

import os

import gc

import numpy as np

import pandas as pd

import seaborn as sns

import matplotlib.pyplot as plt

from sklearn.model\_selection import train\_test\_split

from sklearn.impute import KNNImputer

from sklearn.preprocessing import LabelEncoder

from sklearn.preprocessing import OneHotEncoder

DATA\_PATH = ' /kaggle/input/marginal-workers-in-tamilnadu-dataset.csv’

OUT\_PATH = ' /kaggle/input/marginal-workers-in-tamilnadu-dataset.csv'

Reading the CSV file into a pandas DataFrame

df = pd.read\_csv("C:\\Users\\HP\\Desktop\\cpcb\_dly\_aq\_tamil\_nadu-2014.csv")

Printing the DataFrame

print(df)

Printing information about the DataFrame

print(df.info)

Converting DataFrame to a NumPy array

ndarray = df.to\_numpy()

print(ndarray)

Iterating through the columns of the DataFrame

for x in df.columns:

Printing unique values for each column

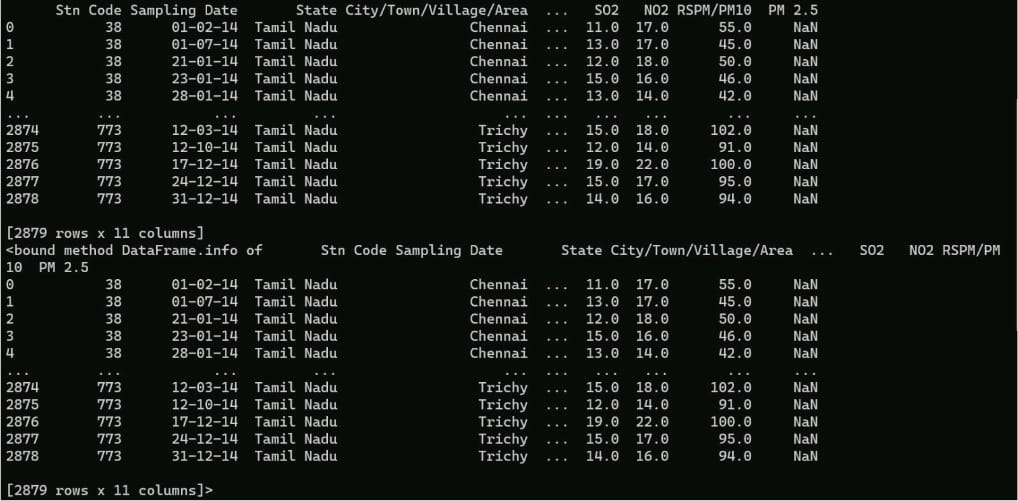
print(x ,':', len(df[x].unique()))

Using the train-test split function

X\_train, X\_test, y\_train, y\_test = train\_test\_split(

X, y, random\_state=104, test\_size=0.25, shuffle=True)

**OUTPUT:**



**Functionality:**

1. Data Reading and Information Display : The script first imports necessary libraries, reads a CSV file into a pandas DataFrame, and displays the DataFrame and its information.

2. Conversion to NumPy Array : The DataFrame is converted to a NumPy array for further analysis.

3. Column Information Display : Unique values for each column in the DataFrame are printed.

4. Train-Test Split : Utilizes the `train\_test\_split` function to split the data into training and testing sets for machine learning tasks.

**Usage:**

To utilize this script, ensure that the required libraries are installed. Adjust the `DATA\_PATH` and `OUT\_PATH` variables according to the file paths of the input and output data. Run the script in a Python environment to execute the data analysis and preprocessing operations.