MCEN 791 Data Mining Lab | Assignment I

Basic Data Mining and Analysis

Instructions for Students

In this project, you will apply your knowledge of Python, NumPy, pandas, and matplotlib to analyze a sustainability-related dataset and extract insights. Follow the instructions and complete each section. We are using a sample dataset on CO2 emissions by country and year. Note: These instructions are only suggestive. You may modify your code or use a different method to achieve the same results.

1. **Import Required Libraries**
    a. numpy as np
    b. pandas as pd
    c. matplotlib.pyplot as plt
    d. seaborn as sns
    e. sklearn.linear_model, LinearRegression
    f. sklearn.model_selection, train_test_split
    g. sklearn.metrics, mean_squared_error, r2_score

2. **Load the Dataset**
    a. We will use a sample dataset on CO2 emissions. https://datahub.io/core/co2-fossil-global
    b. https://raw.githubusercontent.com/owid/co2-data/master/owid-co2-data.csv

3. **Filter** the dataset for a manageable subset, years 2000 onward and selected countries, 'India', 'United States', 'China', 'Germany', 'Brazil'

4. **Explore the Data**
    a. First 5 rows of the dataset
    b. Print the head of the data
    c. Print summary statistics

5. **Clean the data**
    a. Check for missing values in key columns
    b. Drop rows with missing CO2 data

6. **Data Mining and Analysis**
    a. **Compute** Mean CO2 emissions per country
    b. Compute maximum CO2 emissions per year

7. **Visualize the data**
    a. Line plot of CO2 emissions over time
    b. Bar plot of average CO2 per capita
    c. Heatmap of correlation

8. **Model the data**
    a. Linear Regression Model
    b. Predict CO2 emissions based on GDP and population
    c. Split the data into train and test sets
    d. Train the model
    e. Make predictions
    f. Evaluate the model
        i. Mean Squared Error
        ii. R^2 Score
        iii. Model Coefficients
        iv. Model Intercept

9. **Interpret your model results**
    a. Based on the above analysis and visualizations, write your observations here: (e.g., Which country has the highest CO2 per capita? How do emissions relate to GDP or population?)

10. **Conclusion**
    Write a short summary of your findings and what you learned during this project.

11. **Save your work for submission**: Save your completed notebook and submit it as instructed by your course instructor.