Assignment - 1

6/11/2024

Name: Sahil Narang
Course: B.Tech. C.S.E. **Batch**: 12
Sap.id.: 500119480
Enrollment no.: R2142230356
Instructor: Prof. C.M. Sherma

Topic

SDG13. Climate Action

Project Idea: Air Quality Prediction

Objective: The objective of this project is to predict the Air Quality Index (AQI) using data on various pollutants and time-based factors. The goal is to identify how different pollutants and seasonal patterns affect air quality and create a model that can forecast AQI values.

Detailed Methodology

1. **Data Exploration**:
    - Examine the data to understand pollutant levels and patterns in AQI over time.
    - Generate basic statistics (average, minimum, maximum values) for each pollutant and visualize trends, such as line charts of pollutant levels over time.
2. **Data Cleaning**:
    - Handle missing values by filling them with appropriate estimates, like the mean of the column, or remove rows with excessive missing values.
    - For instance, missing values in PM2.5 can be filled with the column's average PM2.5 value to ensure completeness.
3. **Feature Engineering**:
    - Extract meaningful time-based features from the Date column to capture seasonal and temporal trends in air quality.
    - This includes adding new columns for month and day of the week, allowing the model to account for seasonal variations in AQI.
4. **Model Selection**:
    - Use two machine learning models:
        - **Linear Regression**: A simple model to predict continuous values, useful as a baseline for AQI prediction.
        - **Random Forest Regressor**: A more advanced model that often yields better accuracy and highlights which pollutants are most important in predicting AQI.
5. **Model Validation**:
    - To evaluate model reliability, use **K-Fold Cross Validation**. This method splits the dataset into multiple parts (folds), trains the model on each part, and averages the results to provide a more accurate performance estimate.
6. **Performance Metrics**:
    - **Mean Absolute Error (MAE)**: Measures the average prediction error to assess how close predictions are to the actual AQI values.
    - **R^2 Score**: Evaluates the model's ability to explain the variation in AQI data, with a higher score indicating better model fit.

Summary of Objectives and Methodology

- **Objective**: Predict AQI based on pollutant levels and time-based features.
- **Methodology Steps**:
    - **Data Exploration** to understand pollutant patterns.
    - **Data Cleaning** to handle missing values.
    - **Feature Engineering** to add time-based information.
    - **Model Selection** with Linear Regression and Random Forest.
    - **Model Validation** using K-Fold Cross Validation.
    - **Performance Evaluation** with MAE and R^2.