ML TEAM BETA

Title: Air Pollution - Analysis on Air Quality Index (AQI)

Project Overview

This project aims to analyze air pollution levels by studying the Air Quality Index (AQI) data. Using data science techniques, we will perform data collection, cleaning, exploration, visualization, and statistical analysis to understand the factors contributing to air pollution. The insights derived from this project will help policymakers, environmentalists, and the public make informed decisions to improve air quality.

Objectives

1. **Data Collection**: Gather AQI data from reliable sources.
2. **Data Cleaning and Preprocessing**: Prepare the data for analysis by handling missing values, outliers, and ensuring consistency.
3. **Exploratory Data Analysis (EDA)**: Explore the dataset to understand the distribution, relationships, and key statistics.
4. **Visualization**: Create visualizations to illustrate trends and patterns in the data.
5. **Statistical Analysis**: Perform statistical tests and build models to identify factors influencing air quality.
6. **Reporting**: Compile findings into a comprehensive report with actionable insights.

Methodology

1. **Data Collection**:
    - Source AQI data from government databases, environmental agencies, or open data platforms.
    - Key attributes to collect: Location, Date, AQI, Main Pollutants (PM2.5, PM10, O3, NO2, SO2, CO), Weather conditions (temperature, humidity, wind speed).
2. **Data Cleaning and Preprocessing**:
    - Handle missing values: Impute or remove missing data.
    - Remove duplicates and outliers.
    - Standardize categorical data and normalize numerical data.
    - Convert date and time data into appropriate formats for time series analysis.

3. **Exploratory Data Analysis (EDA)**:
    - Descriptive statistics: Mean, median, mode, standard deviation.
    - Distribution analysis: Histograms, box plots.
    - Correlation analysis: Heatmaps to identify relationships between variables.
    - Time series analysis: Trend and seasonality detection.
4. **Visualization**:
    - Line graphs to show trends over time.
    - Bar charts and pie charts to show the distribution of AQI levels across different locations.
    - Scatter plots to visualize relationships between pollutants and AQI.
    - Geographic maps to display spatial distribution of AQI.
5. **Statistical Analysis**:
    - Hypothesis testing: T-tests, chi-square tests to determine the significance of findings.
    - Regression analysis: Linear and multiple regression to identify factors that predict AQI levels.
    - Clustering: Group locations with similar pollution characteristics using k-means clustering.
    - Machine learning models: Predictive modeling using techniques like decision trees, random forests, and support vector machines.
6. **Reporting**:
    - Summarize findings in a detailed report.
    - Include visualizations and statistical evidence.
    - Provide actionable insights for policymakers, environmental agencies, and the public.

Tools and Technologies

- **Programming Language**: Python
- **Libraries**: Pandas, NumPy, Matplotlib, Seaborn, Scikit-learn, Statsmodels, Geopandas
- **IDE**: Jupyter Notebook or any Python IDE
- **Version Control**: Git/GitHub
- **Data Sources**: Government and environmental agencies' databases, open data platforms like Kaggle or data.gov.

Expected Outcomes

- A cleaned and well-documented dataset of AQI.
- Comprehensive EDA with visualizations.
- Insights into the factors affecting air quality.
- Predictive models for AQI levels.
- A detailed report with actionable recommendations.

Conclusion

This project will provide valuable insights into the factors that contribute to air pollution by analyzing AQI data. By leveraging Python for data analysis, we aim to uncover patterns and trends that can inform policymakers, environmentalists, and the public, ultimately contributing to better air quality management and pollution control strategies.