IST 652: Scripting for Data Analysis

Project Proposal

Table of Contents

1. Project Title

Air Quality Analysis and Forecasting

2. Team Members

- Tanu Rana
- Rijul Chaturvedi
- Ninad Bhikaje
- Rushikesh Shinde

3. Topic of Investigation

- The project will investigate air quality patterns using the dataset from data.gov("https://catalog.data.gov/dataset/air-quality"). The analysis will focus on understanding the factors influencing air quality, identifying patterns over time, and forecasting future air quality levels.
- The project aims to answer key questions such as:
    - What are the major pollutants affecting air quality?
    - How does air quality vary across different regions or time periods?
    - Can future air quality levels be accurately predicted based on historical data?

4. Data Sets

- **Primary Dataset:** The main dataset for this project is "https://catalog.data.gov/dataset/air-quality" which contains information about air quality metrics such as levels of pollutants (e.g., PM2.5, NO2), geographic locations, and timestamps.

5. Methods of Data Acquisition and Analysis

- **Data Cleaning and Preprocessing:**
    - Handle missing values and outliers.
    - Convert timestamp columns to proper date-time formats.
- **Exploratory Data Analysis (EDA):**
    - Visualize the distribution of pollutants over time.
    - Identify trends and seasonality in air quality data.
- **Correlation and Statistical Analysis:**
    - Calculate correlation coefficients to understand relationships between pollutants and other variables (e.g., weather conditions).
- **Forecasting Models:**
    - Implement time series forecasting using models such as ARIMA, SARIMA, or Prophet to predict future air quality levels.

    - Consider machine learning techniques like Random Forest or LSTM for advanced forecasting.

6. Development Tasks and Guidance

- **Data Cleaning:**
    - Cleaning the given air quality data with various methods.
- **Feature Engineering:**
    - Create new features (e.g., daily average pollutant levels) to improve model performance.
- **Model Development:**
    - Build and evaluate time series forecasting models.
- **Model Evaluation and Tuning:**
    - Fine-tune models and compare their performance using metrics such as MAE and RMSE.

7. Tools and Libraries

- **Python Libraries:**
    - `pandas` and `numpy` for data manipulation.
    - `matplotlib` and `seaborn` for data visualization.
    - `statsmodels`, `prophet`, and `scikit-learn` for modeling and forecasting.
    - `tensorflow` or `pytorch` for implementing deep learning models (if needed).
- **Data Sources:**
    - The "https://catalog.data.gov/dataset/air-quality" dataset and any additional data sources.

8. Online Services / Tools required

No additional online services required besides the open-source Python libraries mentioned in the previous section.

9. Scalability Considerations

While we expect our laptops to be sufficient for initial analysis, if the dataset proves to be too large or complex for local computation, we may need to utilize cloud-based virtual machines (VMs) with higher computational power. A more powerful VM on platforms such as AWS, Google Cloud, or Azure could provide the necessary resources for scaling our analysis and models. However, this requirement will be better assessed after the initial exploration of the dataset.

10. Expected Outcomes

- A comprehensive analysis of air quality patterns, including trends over time.
- Forecasting future air quality levels to help identify potential risk periods.
- Insights into the factors influencing air quality, with actionable recommendations.

11. Challenges and Considerations

- Handling missing or incomplete data.
- Identifying the best modeling approach for accurate forecasting.
- Integrating additional datasets to enrich the analysis.