Project Demo: Air Quality Prediction in Asia using Decision Tree (ML model)

Prepared by Ritika Deshmukh, Data Analyst | MacroEdtech

Objective:

Air pollution has become one of the most serious environmental and public health challenges across Asia due to rapid industrialization, urbanization, increasing vehicle usage, and population growth. Many major Asian cities frequently experience poor air quality levels that can lead to respiratory diseases, cardiovascular problems, reduced life expectancy, and overall decline in quality of life. Monitoring and predicting air quality is therefore essential for governments, environmental agencies, and city planners to take timely preventive actions and develop effective pollution control policies.

Air quality is typically measured using the Air Quality Index (AQI), which is calculated based on the concentration of several harmful pollutants. Among these, PM2.5 (fine particulate matter with a diameter less than 2.5 micrometers) is especially dangerous because it can penetrate deep into the lungs and bloodstream. PM10 refers to larger particulate matter (diameter less than 10 micrometers) that can cause respiratory irritation and health issues. Other important pollutants include NO_2 (Nitrogen Dioxide) produced mainly from vehicle emissions, SO_2 (Sulfur Dioxide) from industrial processes and fuel combustion, CO (Carbon Monoxide) from incomplete burning of fuels, and O_3 (Ground-level Ozone) formed through chemical reactions in the atmosphere. Weather conditions such as temperature, humidity, and wind speed also influence pollutant dispersion and concentration.

The objective of this project is to analyze air quality patterns across multiple Asian countries and develop a machine learning model using the Decision Tree algorithm to predict the AQI category based on environmental, industrial, traffic, and meteorological factors. Decision Trees are particularly useful because they create clear and interpretable decision rules, allowing stakeholders to understand how different factors contribute to air pollution levels. By identifying the most influential pollutants and conditions, the model can assist environmental authorities in data-driven decision making, early warning systems, pollution control planning, and resource allocation. Ultimately, this project demonstrates how machine learning can support smarter environmental monitoring and policy decisions to address the growing air quality crisis in Asia.

Dataset Description

The dataset used in this project is a synthetic but realistic Asia Air Quality dataset designed to represent environmental conditions across major cities in Asia. It contains 30,000 observations collected for multiple countries and cities, along with pollution indicators, weather conditions, and human activity factors such as industrial and traffic intensity. The dataset is structured to reflect real-world air quality patterns, including seasonal variation, the impact of urban activities, and meteorological influences on pollution levels.

This dataset includes observations from major Asian countries such as India, China, Japan, South Korea, Indonesia, Thailand, Pakistan, and the United Arab Emirates, covering large metropolitan cities where air pollution is a significant concern. Each record represents air quality measurements for a specific city and date, along with the corresponding environmental and activity-related factors affecting pollution.

Dataset Size and Structure

- **Number of records:** 30,000
- **Number of features:** 17
- **Data type:** Mixed (Numerical, Categorical, Date)
- **Target variable:** AQI_Category (multi-class classification)

Feature Description

1. Location Features

Country

- Type: Categorical
- Description: Name of the Asian country where the observation was recorded.
- Purpose: Helps analyze regional pollution patterns and compare countries.

City

- Type: Categorical
- Description: Major metropolitan city within the country.
- Purpose: Allows city-level analysis of pollution trends.

2. Time Feature

Date

- Type: Date
- Description: Observation date within a one-year period.
- Purpose: Useful for time-based or seasonal analysis.

Season

- Type: Categorical
- Categories: Winter, Summer, Monsoon, Autumn
- Description: Derived from the date to represent seasonal environmental conditions.
- Importance: Air pollution often increases during winter due to temperature inversion and decreases during monsoon due to rainfall.

3. Human Activity Indicators

Industrial Index

- Type: Numerical (0–100)
- Description: Represents the level of industrial activity in the city.
- Impact: Higher industrial activity leads to increased emissions of PM, SO_2, and other pollutants.

Traffic Index

- Type: Numerical (0–100)
- Description: Represents traffic density and vehicle usage.
- Impact: High traffic increases NO_2, CO, and particulate matter.

These two variables simulate the major human contributors to urban air pollution

4. Pollutant Concentration Features

PM2.5

- Fine particulate matter ($\mu g/m^3$)
- Highly harmful as it penetrates deep into the lungs and bloodstream.

PM10

- Coarse particulate matter ($\mu g/m^3$)
- Causes respiratory irritation and breathing problems.

NO2 (Nitrogen Dioxide)

- Mainly from vehicle emissions and fuel combustion.

SO2 (Sulfur Dioxide)

- Produced by industrial processes and burning fossil fuels.

CO (Carbon Monoxide)

- Generated from incomplete fuel combustion, especially vehicles.

O3 (Ground-level Ozone)

- Formed through chemical reactions between pollutants under sunlight.

These variables are the primary inputs used to calculate the Air Quality Index.

5. Meteorological Features

Temperature (°C)

- Influences chemical reactions and pollutant formation.

Humidity (%)

- Affects particulate matter concentration and atmospheric conditions.

Wind_Speed (km/h)

- Higher wind speeds help disperse pollutants, reducing AQI levels.

Weather conditions play a critical role in pollution accumulation and dispersion.

6. Air Quality Indicators

AQI (Air Quality Index)

- Type: Numerical
- Description: A weighted index calculated from pollutant concentrations.
- Represents overall air pollution level.

AQI_Category (Target Variable)

- Type: Categorical
- Classes:
    - Good (0–50)
    - Moderate (51–100)
    - Poor (101–200)
    - Very Poor (201–300)
    - Severe (>300)

This is the target variable for the Decision Tree model, making the problem a multi-class classification task.

Dataset Characteristics

- Contains both categorical and numerical features
- Shows non-linear relationships between variables
- Includes seasonal and regional variation
- Simulates real-world dependencies:
    - Winter increases pollution
    - Wind reduces pollutant concentration
    - Traffic affects NO_2 and CO
    - Industrial activity increases PM and SO_2

Importance for Machine Learning

This dataset is suitable for:

- Decision Tree classification
- Feature importance analysis
- Exploratory Data Analysis (EDA)
- Hyperparameter tuning
- Model interpretation
- Environmental decision support

Steps (Till Model Saving)

Steps:

1. Data Loading
2. Preprocessing
3. Train/Test Split
4. Model Training
5. Hyperparameter Tuning
6. Evaluation
7. Model Saving

Final Outcome

- Predict AQI category for any Asian city
- Identify key pollution drivers
- Optimized Decision Tree using hyperparameter tuning
- Full EDA completed
- Model saved for future deployment

Complete Analysis using Machine Learning Model – Decision Tree

Link :
https://colab.research.google.com/drive/1rZk8b7xHvoVCTBbxfs1VrnShnkwQi3aS?usp=sharing