Course Project Milestone 2- Data Selection and Project Proposal- Air Quality Prediction

Guruprasad Velikadu Krishnamoorthy and Hemalatha Subbiah

College of Science and Technology, Bellevue University

DSC630-T302: Predictive Analysis

Professor. Fadi Alsaleem

December 10, 2023

Course Project Milestone 2- Data Selection and Project Proposal- Air Quality Prediction

Introduction

Air pollution is a major concern in the modern world. With the rise of the Industrial Revolution in the 19th century and with the invention of new machinery and technologies, economies around the world saw explosive growth in the following decades accompanied by the migration of farmers from the countryside to fast-growing crowded cities in search of factory jobs. (Kiger, 2021) While the industrialists were amazed by the profits being made, the governments were astonished by the progress made in the economy, little did they all know the side effects that changed the landscape of the earth forever, for the worse. The emission of gases from the factories and the pollutants from the automobiles had left an everlasting impact on the face of the earth. In this project, the effects of air pollution and the ways to combat it by being able to predict it using machine learning algorithms will be discussed in detail.

Problem statement:

On average, it is estimated that a person inhales about 2000 gallons of air daily. Hence it is essential that the we breathe must be of good quality. The polluted air when inhaled, gets straight into our lungs, then enters the bloodstream, and can cause more damage to internal organs such as the brain, heart, etc. and young children are the most affected ones.

This project aims to develop a machine-learning model capable of predicting the Air Quality Index(AQI) based on various environmental parameters and pollutant concentrations. The model should be capable of providing accurate AQI predictions for future time points, allowing for early detection of potential air quality issues.

Why the problem is important:

According to a recent study, air pollution-related ailments are the fourth largest contributing factor to premature deaths, and about 4.5 million deaths around the world are reported to be related to air pollution. Some air pollutants such as mercury, lead, and benzene can cause several health issues, in

some cases even death. (10 Things You Never Knew Could Cause Lung Cancer, n.d.) In New Delhi, India where the air quality is ranked among the worst, people reported that it felt like breathing poison during extreme smog conditions. The level of PM2.5 pollutant which is small enough to enter the bloodstream was reported to be 25 times more than the limits recommended by the World Health Organization in the city. ("Like Breathing Poison": Children in India's Delhi Hit Hard by Smog, n.d.)

Though some measures were taken in the last decade to bring awareness to people about the impacts of air pollution, it was too little, and it will be too late if all the countries around the globe don't work together to solve this looming problem. If the current trend continues, in the future people may be forced to pay for clean breathable air, making it an absolute priority to take measures to clean up the air and provide a sustainable and healthy environment for our future generations.

Who would be interested in solving the problem?

Pretty much everyone is affected by the rising air pollution-related problems around the world. Children, older adults, and people living with asthma and other breathing-related disorders are the most affected. People who have lower incomes and those who live near the sources of pollution such as factory workers may also face greater harm. Though it is believed that people in bigger cities are the most affected due to vehicle and industrial emissions, small cities and towns are not spared either. Agricultural activities, Fertilizer manufacturing, and livestock production that release Methane can all contribute to air pollution in rural areas. Hence this is a universal problem that affects most of the population around the globe.

The availability of comprehensive air quality datasets provides an opportunity to develop predictive models that can forecast the Air Quality Index (AQI) with high accuracy. Being able to predict the air quality based on the pollutant concentration can not only alert the public to take precautions but can also help take measures in advance to be prepared or to avoid the situation altogether.

Source of the Data:

The data is extracted from the Environmental Protection Agency website. (Download Files | AirData | US EPA, 2015). A total of 13 datasets will be used to predict the air quality in this project, such as Ozone concentration datasets, SO2, NO2, and Carbon Monoxide concentration datasets; datasets of pollutants such as PM2.5, and PM10; datasets containing meteorological data such as Temperature, Pressure, Humidity, etc. All these datasets are for the year 2022.

Some of the commonly used fields in the dataset are listed below:

- Columns to identify the site location where the readings were measured:
    - State code, County Code, Site Number, Latitude, Longitude, Local Site name, Address, State Name, County Name, City Name.
- Columns to identify the Pollutant/ Gases/ the Metrological quantity:
    - Parameter Name: Represents the Parameter of the Pollutant /Particulate/Toxin/Meteorological measure.
    - Parameter Code: Unique code assigned to the parameter describing the Pollutant /Particulate / Toxin /Meteorological measure.
    - Units of Measure: The unit in which the parameter is measured.
    - Observation Count: Number of observations captured on the specified date in the site location.
    - Arithmetic Mean: The Mean value of the quantity of the parameter captured on the given date at the given site.
- Column to identify the Air Quality:
    - AQI: Represents Air Quality Index value measured on the specified date at the specified site location.
- Date fields:

    - Date Local: Date when the parameter was measured and recorded.

- Columns for Pollutants:
    - Ozone (O3): Concentration of ozone in the air
    - Carbon Monoxide (CO): Concentration of carbon monoxide in the air.
    - Nitrogen Dioxide (NO2): Concentration of nitrogen dioxide in the air.
    - Sulfur Dioxide (SO2): Concentration of sulfur dioxide in the air.

Why is the data useful to the Problem:

The datasets contain information about the concentration of Gases such as Ozone, Sulfur Dioxide (SO2), and Nitrogen Dioxide (NO2) measured at different sites around the US for the year 2022. This along with the concentration of particulates such as PM2.5, and PM10.0, and other parameters such as Temperature, pressure, and humidity can provide valuable information to predict the Air Quality index. As all these metrics are available in individual datasets, they must be combined to extract useful insights from them. There are about 250,000 rows on average in these datasets thus providing sufficient data required for the research.

Type of Models to be used:

The target variable that will be predicted in this machine learning project is the Air Quality Index (AQI). As it is a continuous numeric quantity, Regression algorithms will be used to predict the air quality index. Various regression algorithms such as Linear Regression and Regularization techniques such as Lasso and Ridge algorithms will be used in this project.

Also, the project can be expanded to include classification models like Random Forest, Gradient Boosting, K-Neighbor, and Decision Tree to predict the Levels of Health concern based on the range of AQI values.

The details are included in Table 1 below (US EPA, 2016)

Air Quality Index (AQI) measured	Health concern Level	Colors
0-50	Good	Green
51-100	Moderate	Yellow
101-150	Unhealthy for Sensitive Groups	Orange
151 to 200	Unhealthy	Red
201 to 300	Very Unhealthy	Purple
301 to 500	Hazardous	Maroon

(Table 1)

Plan to evaluate the results:

1. The 13 datasets used in this project will be loaded into their respective Data Frames and then combined based on the common keys.
2. Exploratory Data Analysis will be performed on the combined data frame to handle null values and treat the outliers. In this step correlation between the variables in the dataset will also be measured.
3. Visualize the dataset to gain insights, identify patterns, and make informed decisions about preprocessing and modeling. More Visualization types like Box plots, Scatter plots, Correlation heat maps, and Confusion Matrix will be incorporated.
4. Features and Target variables will be declared, and the dataset will then be split into train and test sets. During this step, feature selection and extraction techniques will be deployed to reduce the number of features and extract useful information from the features.
5. The regression model algorithms will then be used to build the prediction model. The accuracy of the model along with the confusion matrix to demonstrate the performance of the model with predicted vs actual values will be computed.
6. Root Mean Square Error (RMSE) which measures the difference between the actual and predicted AQI values will be used as an evaluation metric, while R-squared which explains the proportion of the variance in the AQI, will be used to evaluate the results.

7. Precision, recall, and F1 scores can offer insights into the model's behavior.
8. Also, cross-validation techniques such as K-fold validation will be deployed to reduce overfitting and to improve the performance of the model on unseen data.

What do I hope to learn?

As the project involves combining the data from various datasets with approximately 250,000 records in each, handling huge volumes of data and combining them without causing performance issues is something that we hope to learn from this project. As the dataset resembles a real-life dataset, this project can help us learn to overcome the challenges of handling large volumes of data at different stages of the project such as Data Preparation, Model Building, Cross-validation, etc. Moreover, from the dataset, I want to understand how pollutant concentrations detected by the sensors are highly correlated with one another and want to analyze the real effect of it.

Risks:

One potential risk of this project is the quality and availability of the data available across the 13 datasets. During the EDA and Data Preparation phase, the data must be carefully studied and joined using common fields such as Site Location and Dates. There is a risk that the data may not be available for all site locations across the different datasets, thus resulting in a much smaller dataset than intended after all joins are performed. Also, the presence of nulls in the essential columns in these datasets can pose an additional risk resulting in data reduction.

Ethical Implications:

While studying the impacts of Air pollution and its relation to global warming and climate change, the ethical implications should be carefully assessed. Some of them are listed below:

- Though gas-powered vehicle emissions and industrial smoke have played a significant role in air pollution, they cannot be entirely replaced by sustainable solutions, as they can lead to many

job losses affecting many families employed by the manufacturing industries. Care must be taken while publishing the results keeping in mind the impact it can have on the families.

- While determining the acceptable levels of greenhouse gases and pollutants for humans, careful assessment should be made while determining the values, as the acceptable levels for humans may cause significant damage to other ecosystems and species. (Brown, 2001)
- The acceptable levels should also be carefully assessed with international considerations in mind, as the gas emissions and pollutants from the developed countries are no longer a local issue. These impacts are already seen on the other side of the world, with extreme floods and drought conditions that were not seen in the past.

Contingency plan:

If we run into problems while combining the datasets or if the performance of the model lags due to a large volume of data, the contingency plan is to reduce the number of datasets. However, the downside of this approach is that some useful information presented by the excluded datasets will be lost and can affect the model performance. If we still face issues, a simplified version of the Air quality dataset with about 9000 rows from Kaggle will be used to continue with the project. (Air Quality Dataset, n.d.) or from the UCI Machine Learning Library. (UCI Machine Learning Repository: Air Quality Data Set, n.d.).

References:

Kiger, P. (2021, November 9). 7 negative effects of the industrial revolution. History. https://www.history.com/news/industrial-revolution-negative-effects

Air quality. (n.d.). Transport & Environment. https://www.transportenvironment.org/challenges/air-quality/

10 Things You Never Knew Could Cause Lung Cancer. (n.d.). WebMD. https://www.webmd.com/cancer/ss/slideshow-lung-cancer-surprising-causes

"Like breathing poison": Children in India's Delhi hit hard by smog. (n.d.). Al Jazeera. Retrieved December 8, 2023, from https://www.aljazeera.com/gallery/2023/11/10/photos-like-breathing-poison-delhi-children-hardest-hit-by-smog#:~:text=Many%20have%20asthma%20or%20pneumonia

Download Files | AirData | US EPA. (2015). Epa.gov. https://aqs.epa.gov/aqsweb/airdata/download_files.html

Air Quality Dataset. (n.d.). Www.kaggle.com. https://www.kaggle.com/datasets/fedesoriano/air-quality-data-set

UCI Machine Learning Repository: Air Quality Data Set. (n.d.). Archive.ics.uci.edu. https://archive.ics.uci.edu/ml/datasets/Air+Quality

Brown, D. (2001). The Ethical Dimensions of Global Environmental Issues. American Academy of Arts & Sciences. https://www.amacad.org/publication/ethical-dimensions-global-environmental-issues

US EPA, O. (2016, August 30). Air Data Basic Information. US EPA. https://www.epa.gov/outdoor-air-quality-data/air-data-basic-information