**DATA ANALYSIS OF AIR QUALITY IN INDIA**

Group members:

* ALIYA P SUBEER
* HARITHA C M
* IRFANA ISMATH
* PRINCY P
* SETHULEKSHMI S
* SREEJAYAN S

Mentor:

* BILAL B N

1. **ABSTRACT**

Air, an essential natural resource, has been compromised in terms of quality by

economic activities. Considerable research has been devoted to predicting instances of

poor air quality, but most studies are limited by insufficient longitudinal data, making it

difficult to account for seasonal and other factors. Several prediction models have been

developed using an 11-year dataset collected by Taiwan’s Environmental Protection

Administration (EPA). Machine learning methods

Air, an essential natural resource, has been compromised in terms of quality by

economic activities. Considerable research has been devoted to predicting instances of

poor air quality, but most studies are limited by insufficient longitudinal data, making it

difficult to account for seasonal and other factors. Several prediction models have been

developed using an 11-year dataset collected by Taiwan’s Environmental Protection

Administration (EPA). Machine learning methods

Air, an essential natural resource, has been compromised in terms of quality by

economic activities. Considerable research has been devoted to predicting instances of

poor air quality, but most studies are limited by insufficient longitudinal data, making it

difficult to account for seasonal and other factors. Several prediction models have been

developed using an 11-year dataset collected by Taiwan’s Environmental Protection

Administration (EPA). Machine learning methods

The survival of mankind cannot be imagined without air. Consistent developments in almost all realms of modern human society have adversely affected the health of the air. Daily industrial, transport, and domestic activities are stirring hazardous pollutants in our environment. Monitoring and predicting air quality have become essentially important in this era, especially in developing countries like India. In contrast to the traditional methods, the prediction technologies based on machine learning techniques are proved to be the most efficient tools to study such modern hazards. The present work investigates **Air Quality Data in India (2015 - 2020) and (2020-2023) Dataset** taken from Kaggle consisting of AQI data published by CPCB (Central Pollution Control Board).

The main focus of this project is learning about EXPLORATORY DATA ANALYSIS, DATA PRE-PROCESSING, FEATURE ENGINEERING, MODELING AND PREDICTION of data by supervised algorithms i.e. (Linear Regression (regression), Logistic Regression, Random Forest Classifier, Decision Tree Classifier (classification)). The main focus of this particular project is AQI (Air Quality Index) prediction, and factors that affects AQI i.e. (SO2, NO2, SPM, RSPM). In this project, a basic data analysis is done on India's Air Quality data, and the value of the air quality index is predicted based on the given features of concentration of sulphur dioxide, nitrogen dioxide, Respirable suspended particulate matter, and suspended particulate matter. The air quality is classified as good, satisfactory, moderately polluted, poor, very poor, and severe.

1. **INTRODUCTION**

Air pollution occurs when harmful or excessive quantities of substances including gases, particles, and biological molecules are introduced into the Earth's atmosphere. Air pollution in India is a serious issue, ranking higher than smoking, high blood pressure, child and maternal malnutrition, and risk factors for diabetes. At least 140 million people breathe air 10 times or more over the WHO safe limit and 13 of the world's 20 cities with the highest annual levels of air pollution are in India. Air pollution contributes to the premature deaths of 2 million Indians every year. In urban areas, most emissions come from vehicles and industry, whereas in rural areas, much of the pollution stems from biomass burning for cooking and keeping warm. In autumn and winter months, large scale crop residue burning in agriculture fields – a low cost alternative to mechanical tilling - is a major source of smoke, smog and particulate pollution.

1. **PROBLEM STATEMENT**

The prediction of air quality index is a complex problem that requires the consideration of multiple factors. Industrial, transport, and domestic activities release hazardous pollutants into the environment, affecting air quality. Additionally, atmospheric patterns such as rain, air pressure, and temperature can influence the volume of each pollutant in the air. To accurately predict air quality index, it is necessary to develop a model that takes into account these factors and their interactions. Machine learning techniques have shown promise in this regard, allowing for the development of predictive models that can accurately forecast air quality.

1. **MAIN CHALLENGES**

1. Difficulty in obtaining or modeling in advance the many factors that affect air quality, such as traffic, weather condition, etc.

2. High uncertainty in the time dimension of air quality.

3. The need to reduce uncertainties in emissions from diffuse sources and to consider both indoor and outdoor sources.

4. The small spatial and temporal scales and nonlinearity of climate effects.

5. The need for integrated air pollution observations from both ground-based and remote sensing instruments, including those on satellites, while ensuring the quality of the measurements.

6. Data Cleaning: The dataset may contain missing values, inconsistencies, or errors that need to be addressed before analysis. Cleaning and preprocessing the data to ensure its accuracy and completeness can be a challenge.

7. Data Exploration: Understanding the characteristics, distributions, and relationships within the dataset can be challenging. Exploratory data analysis techniques may be required to uncover patterns, trends, or outliers in the data.

8. Feature Selection and Engineering: Identifying relevant features for analysis and potentially creating new features from existing ones can be a challenge. Understanding the domain and selecting informative features is important for accurate modeling.

1. **OBJECTIVES**

The main objectives of the project are:

**WHAT WE DO**

1. To identify the patterns and relationships in the data.
2. To study the various features depending on the Air Quality Index.
3. To develop accurate models for the various components of air, such as particulate matter, nitrogen dioxide, sulphur dioxide, carbon monoxide, and PM2.5.

**WHY WE DO**

1. To predict the value of Air Quality Index based on given features of concentration of Sulphur dioxide, nitrogen dioxide, Respirable suspended particulate matter.
2. To classify the Air Quality as good, moderate, poor, unhealthy and healthy.
3. To create a web hosting platform where inputs are given by anyone and to predict the AQI.

**WHERE IT USE**

1. The AQI prediction is useful for a wide range of individuals and organizations. It is commonly used by environmental agencies, government bodies, and health organizations to track air quality levels and issue warnings or advisories when necessary.

**TO WHOM THIS USEFUL**

1. Individuals with respiratory conditions such as asthma or COPD may use AQI predictions to plan their activities and mitigate their exposure to poor air quality. Many industries and businesses also rely on AQI predictions to make decisions related to operations, transportation, and resource management.
2. **DATASET**

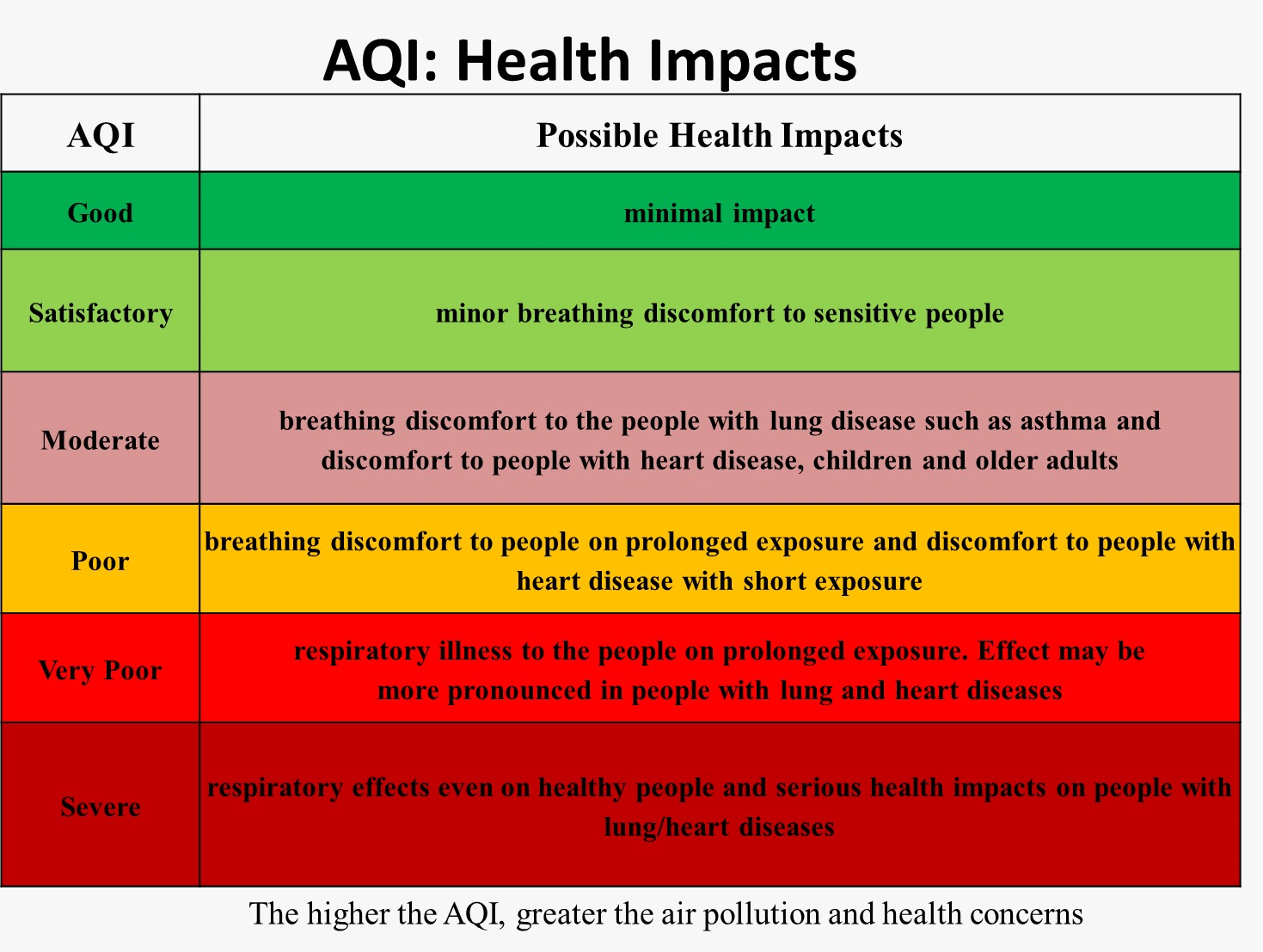
In the dataset there are (53036 rows, 17 columns). The analysis combines different datasets of for the year (2015-2020) and (2020-2023) to provide a comprehensive understanding of the trends and patterns in various locations over time. We are attaching the dataset.

**FEATURES DESCRIPTION**

1. **City:** Ahmedabad, Aizawl, Amaravati, Amritsar, Bengaluru, Bhopal, Brajrajnagar,

Chandigarh, Chennai, Coimbatore, Delhi, Ernakulam, Gurugram, Guwahati, Hyderabad, Jaipur, Jorapokhar, Kochi, Kolkata, Lucknow, Mumbai, Patna, Shillong, Talcher, Thiruvananthapuram, Visakhapatnam.

1. **Date:** Date in which the samples are taken.
2. **PM2.5**: It refers to particles that have diameter less than 2.5 micrometers (more than 100 times thinner than human air) and remain suspended for longer. These particles are formed as a result of burning fuel and chemical reactions that take place in the atmosphere.
3. **PM10**: PM10 are the particles with a diameter of 10 micrometers and they are also called fine particles. An environmental expert says that PM10 is also known as Respirable particulate matter. Particulate matter is a complex mixture of smoke, metals, nitrates, sulphates, dust water and rubber etc.
4. **NO**: Nitric oxide is not considered to be hazardous to health at typical ambient concentrations, but nitrogen dioxide can be. ... NOx gases react to form smong and acid rain as well as being central to the formation of fine particles (PM) and ground level ozone.
5. **NO2**: Nitrogen dioxide is part of a group of gaseous air pollutants produced as a result of road traffic and other fossil fuel combustion processes. Its presence in air pollutants.
6. **NOx**: NOx pollution refers to the release of nitrogen oxides (NOx) into the atmosphere. NOx is a group of toxic gases formed during high-temperature combustion processes, especially in vehicles, power plants, and industrial facilities. It is a major contributor to air pollution, which can have harmful effects on human health, the environment, and wildlife. The adverse effects of NOx pollution include respiratory problems, reduced visibility, acid rain, and eutrophication (excess nutrients leading to algal blooms and loss of aquatic life). Governments and industries have implemented various measures to reduce NOx emissions, such as stricter emissions standards, cleaner technologies, and improved transportation systems.
7. **NH3**: Gaseous ammonia (NH3) is the most abundant alkaline gas in the atmosphere. NH3 plays a significant role in the formation of atmospheric particulate matter, visibility degradation and atmospheric deposition of nitrogen to sensitive ecosystem.
8. **CO**: It is produced in the incomplete combustion of carbon-containing fuels such as gasoline, natural gas, oil, coal and wood.
9. **SO2**: It is formed when fuel containing sulphur, such as coal and oil, is burned, creating air pollution. It affects the environment when it reacts with substances in the atmosphere to form acid rain.
10. **O3**: It is harmful to air quality outside of the ozone layer. Ground level ozone is a colorless and highly irritating gas that forms just above the earth's surface.
11. **Benzene**: The benzene in indoor comes from product that contain benzene such as glues, paints, furniture wax, and detergents. The air around hazardous waste sites or gas stations can contain higher level of benzene than in other areas.
12. **Toluene**: Motor vehicle and industrial emissions are the major sources of pollutants.
13. **Xylene**: Motor vehicle emissions are the predominant source of xylene in the urban air environment. Evaporation from petroleum fuel storage facilities and service stations.
14. **AQI**: The air quality index is an index for reporting air quality on a daily basis. It is a measure of how air pollution affects one's health with in short time period. The purpose of the AQI is to help know how the local air quality impacts their health.



1. **AQI\_Bucket:** it is variable of AQI.