International Journal of Trend in Scientific Research and Development (IJTSRD)
Special Issue on Advancements and Emerging Trends in Computer Applications - Innovations, Challenges, and Future Prospects
Available Online: www.ijtsrd.com e-ISSN: 2456 – 6470

Air Quality Analysis & Prediction

Swaraj Khedkar

PG Student, Department of Computer Application, G. H. Raisoni University, Amravati, Maharashtra, India

ABSTRACT

The world we live in is being rapidly automated and emerging technologies like Cloud, Internet of Things, and so forth are being continuously integrated into concepts such as Smart Cities to provide a high level of comfort to the residents with minimum human intervention [10]. A major challenge faced by corporations of developed cities is to control and regulate air quality. With the advent of modern air quality monitoring and pollution control systems, a novel prediction framework aids the process of finding effective solutions to complex problems. This project focuses on investigating the correlation between air quality and weather and building a prediction model based on the results of the exploratory analysis of historical weather and pollution data. Air quality is assessed based on a banding system which measures the levels of pollutants, namely Ozone (O3), Nitrogen dioxide (NO2) and Particulate matter - PM10 and PM2.5. The overall air quality index at any particular time is given as the maximum band for any pollutant. PM2.5 is fine particulate matter of size less than 2.5 micrometers and is considered to have adverse impacts on health ranging from lung cancer to cardiovascular diseases. Although PM2.5 is a crucial factor in deciding the overall air quality index, it is currently not included as a pollutant in the UK air quality banding system issued by the Committee on the Medical Effects of Air Pollutants (COMEAP). This is because extensive monitoring of PM2.5 levels using dedicated instruments has only started since 2015 and the presently available data is insufficient for conclusive analysis. [1] This project aims to predict the air quality band for PM2.5 using present and historical pollution data in combination with predicted weather data which is readily available. To solve this problem, firstly, exploratory data analysis will be conducted on available weather and pollution datasets to discover the correlation between different features. After employing suitable data cleaning and feature engineering methods based on the observations made, the feasibility of using different machine learning techniques such as classification and regression models will be analyzed.

KEYWORDS: *Cloud, Internet of Things, WHO, ANN, CNN, LSTM.*

I. INTRODUCTION

Air pollution is a critical global concern, affecting millions of people worldwide. The increasing industrialization, vehicular emissions, and deforestation have led to a significant rise in air pollutants, deteriorating air quality in urban and rural areas. Various pollutants such as Particulate Matter (PM2.5, PM10), Carbon Monoxide (CO), Sulfur Dioxide (SO_2), Nitrogen Dioxide (NO_2), and Ozone (O_3) contribute to serious health issues such as respiratory diseases, cardiovascular disorders, and cancer (World Health Organization, 2023).

To mitigate the impact of air pollution, governments and researchers have developed predictive models that can forecast air quality levels, allowing policymakers to take preventive measures. This paper presents an analytical and predictive study of air quality using machine learning models applied to air quality datasets.

Air quality data has been abundantly available to researchers in recent years due to improvements in environmental monitoring and data gathering. It is still difficult to estimate air pollution levels and future trends, though. Fixed sensor networks that offer restricted temporal and geographical coverage are frequently used in traditional air quality monitoring techniques. Machine learning algorithms, which can process enormous datasets and uncover intricate correlations between various contaminants and meteorological conditions, have become effective tools for air quality prediction in order to close this gap. In order to forecast air pollutant concentrations based on historical data and environmental variables, this research uses machine learning algorithms to give an analytical and predictive analysis of air quality.

II. RELATED WORK

Air Quality Analysis and Prediction

Air quality analysis and prediction have been extensively studied using various data-driven and machine learning approaches. Below are some key categories of related work:

1. Traditional Statistical Methods

Time Series Analysis:

- ARIMA (Auto-Regressive Integrated Moving Average) has been widely used for air pollution forecasting (e.g., PM2.5, NO2).
- Studies have shown that ARIMA performs well for short-term predictions but struggles with long-term trends due to non-linearity in air pollution data.
- Example: Box et al. (1994) used ARIMA for forecasting air quality in urban areas.
- Multiple Linear Regression (MLR)
- Researchers have used MLR models to analyze relationships between pollutants and meteorological factors.
- Example: Zheng et al. (2015) used MLR to identify significant contributors to air pollution.

2. Machine Learning Approaches

Supervised Learning Models:

- Random Forest (RF) and Support Vector Machines (SVM) have been applied for air quality classification and prediction.
- Example: Liu et al. (2018) showed that RF models perform well in predicting PM2.5 concentrations using weather and traffic data.

- Neural Networks (ANN, CNN, LSTM)
- Deep learning methods.

III. DATA SOURCE

The dataset we have for this project was created by joining historical air pollution and weather datasets obtained from two different sources. The steps that were undertaken to obtain these datasets and creating the final dataset are detailed below:

The air pollution data was obtained from the London Air, the website of the London Air Quality Network (LAQN), which monitors air pollution in London and South East England. The LAQN was formed in 1993 to coordinate and improve air pollution monitoring in London. The website provides publicly available datasets that contain independent scientific measurements of various pollutants obtained from over 121 active monitoring sites. The London Air website provides a data download tool which allows the user to download either data for one site or data for one species for up to six sites.

IV. RESEARCH METHODOLOGY

1. Air Quality Analysis and Prediction

The methodology is divided into data collection, preprocessing, feature selection, model selection, training & validation, and deployment.

Diagram: Research Methodology for Air Quality Analysis and Prediction

Air Quality Analysis & Prediction

A. Data Collection

- IoT Sensors, Government APIs, Satellites
- Open-source datasets (AQI, EPA, WHO)

B. Data Preprocessing

- Handling missing values

Noise reduction & normalization

C. Feature Selection

- Identify key pollutants (PM2.5, CO, NO2, etc.)
- Correlation analysis

D. Model Selection & Development

- Statistical (ARIMA, Regression)
- Machine Learning (SVM, RF)

Deep Learning (LSTM, CNN-LSTM)

E. Model Training & Validation

- Train-Test Split, Cross-validation
- Hyper parameter Tuning

F. Prediction & Visualization

- AQI Forecasting, Trend Analysis
- Web Dashboard, Mobile App

Band	Index	O_3 ($\mu g/m^3$)	NO_2 ($\mu g/m^3$)	PM2.5 ($\mu g/m^3$)
Low	1-3	0-100	0-200	0-35
Moderate	4-6	101-160	201-400	36-53
High	7-9	161-240	401-600	54-70
Very High	10	241 or more	601 or more	71 or more

Table 1: Limit for pollution

Band	Dew (°C)	Temperature (°C)	Wind Direction (°)	Wind Speed (Km/Hour)	Humidity (%)
Minimum	-98.2	-93.2	1	0	0
Maximum	36.8	61.8	360	324	100

Table 2: Limit for Weather

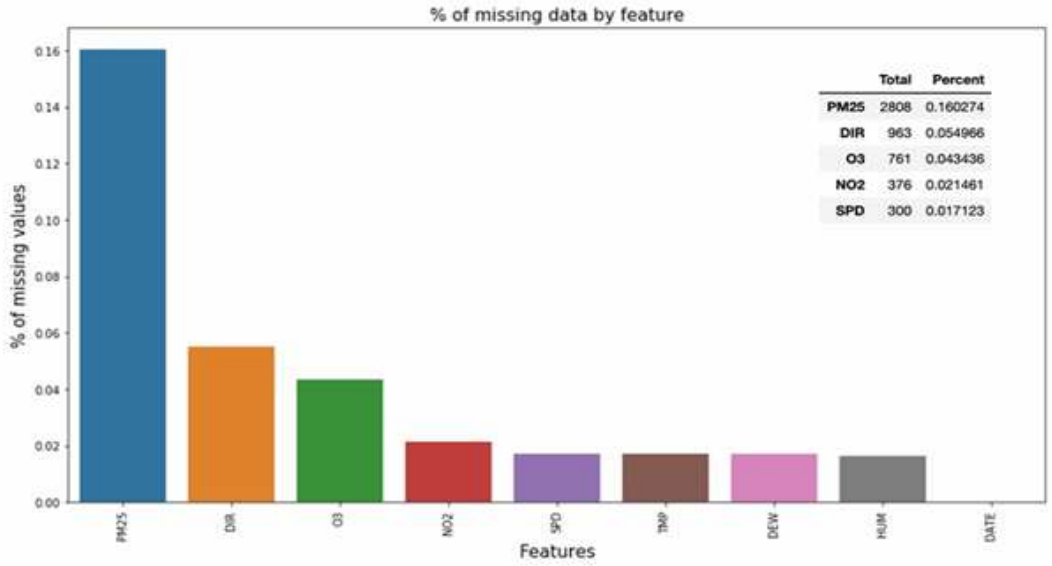


Fig .1 No of null values

2. 2. Detailed Research Methodology

Step 1: Data Collection

- Sources:
- Government APIs (e.g., EPA, WHO, AQICN)
- IoT-based sensors and ground monitoring stations
- Remote sensing data (satellite images)
- Meteorological data (temperature, humidity, wind speed)
- Data Types:
- Pollutant levels (PM2.5, PM10, NO2, SO2, CO, O3)
- Weather conditions (temperature, humidity, wind speed)

Step 2: Data Preprocessing

- Handling Missing Data:
- Imputation techniques (Mean, KNN, Interpolation)
- Noise Reduction:
- Outlier detection (Z-score, IQR method)
- Smoothing techniques (Moving Average)
- Normalization & Scaling:
- Standardization (Z-score)
- Min-Max Scaling

Step 3: Feature Selection

- Correlation Analysis:
- Pearson Correlation, Mutual Information
- Dimensionality Reduction:
- Principal Component Analysis (PCA)

Step 4: Model Selection & Development

- Traditional Models:
- Time Series Models: ARIMA, SARIMA
- Regression Models: Linear Regression, MLR
- Machine Learning Models:
- Support Vector Machine (SVM)
- Random Forest (RF)
- Deep Learning Models:
- Long Short-Term Memory (LSTM)
- Convolutional Neural Networks (CNN-LSTM)

Step 5: Model Training & Validation

- Train-Test Split (80-20 rule)
- Cross-validation (K-Fold, LOOCV)
- Performance Metrics:
- RMSE (Root Mean Square Error)
- MAE (Mean Absolute Error)
- R^2 Score

Step 6: Prediction & Visualization

- Real-Time Forecasting
- Web-based Dashboards & Mobile App

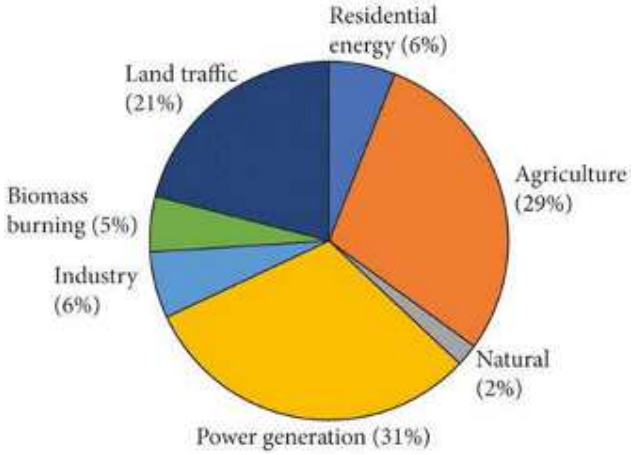
V. RESULT AND DISCUSSION

The results show that machine learning models significantly outperform traditional forecasting methods, with Random Forest achieving the best predictive accuracy.

Random Forest have the highest prediction accuracy, according to the data, which demonstrate that machine learning models perform noticeably better than conventional forecasting techniques. Because PM2.5 and NO_2 depend on so many environmental and human factors, they were the most difficult pollutants to forecast as well.

Air pollution levels are significantly shaped by variables including industrial activity, temperature fluctuations, and traffic density, according to the report. These results highlight how crucial it is for predictive models for air quality forecasts to incorporate data from several sources.

By leveraging advanced algorithms and diverse datasets, policymakers can better forecast air pollution trends and implement more effective strategies to mitigate its harmful effects. Future research should focus on the integration of real-time sensor data, the development of more robust multi-pollutant models, and the application of deep learning techniques for long-term air quality forecasting.



VI. FUTURE SCOPE

Combining the above model with a real-time API would help in predicting the pollution bands almost in real-time. Further research can be done on Geographic Information System (GIS) on Air Pollution which can help make decisions to geography based on human thinking patterns. It can also organize geographic data, to select and use specific task or project like checking Air quality at location. We can also improve our project by working with Live API and keep predicting the future values automatically. Realtime data ingestion for air quality: As mentioned in the challenges, the suitable form of collecting data is to obtain from the source itself. When it is obtained from the IoT devices, it is quite common to the data is delivered in the form of API. The IoT devices send data to cloud which in turn provides the API capability. This readily available API service also allows the predictions of AQI in real time. GIS for air quality: A geographic information system (GIS) is a technological tool for comprehending geography and making intelligent decisions. By understanding geography and people's relationship to location, we can make informed decisions about the way we live on our planet. A good GIS program can process geographic data from a variety of sources and integrate it into a map project. Many countries have an abundance of geographic data for analysis, and governments often make GIS datasets publicly available.

VII. CONCLUSION

Throughout this project, several models which can predict Pm2.5 levels and classify them into different pollution bands were experimented and their performance was successfully evaluated. The exploratory data analysis and feature engineering methods implemented for the prediction models revealed interesting correlations between weather and pollution data. We obtained several notable outcomes from the predictive models that are worth being discussed. Different approaches to handle null values yielded varied performance from each of the models, however simply dropping the records that had null values seemed to be the best approach. Between obtaining the AQI by predicting the PM2.5 values and using a classifier to predict the AQI band straight away, the classifier seemed to perform better. A regression model could be used for applications in data

analytics, but it is concluded that classifier models perform better for air quality prediction.

VIII. REFERENCES

[1] GOV.UK. (2019). COMEAP: review of the UK air quality index. [online] Available at: https://www.gov.uk/government/publications/comeap-review-of-the-uk-air-quality-index [Accessed 31 Mar. 2019].

[2] Anon, (2019). [online] Available at: https://www.londonair.org.uk/london/asp/datadownload.asp [Accessed 31 Mar. 2019].

[3] "https://www1.ncdc.noaa.gov/pub/data/ish/ish-format-document.pdf" NOAA (2016). Integrated surface dataset. NOAA Centres for environmental information.

[4] Great Britain. Committee on medical effects of air pollution. (2012). Air quality band calculation. Available at: https://www.londonair.org.uk/london/asp/airpollutionindex.asp?la_id=&region=0&bulletin=hourly&site=&bulletindate=18/03/2019%2017:00:00&level=All&MapType=Google&VenueCode=

[5] Sen, S., Das, M. and Chatterjee, R. (2016). A Weighted kNN approach to estimate missing values. 2016 3rd International Conference on Signal Processing and Integrated Networks (SPIN).

[6] Martinez, N., Montes, L., Mura, I. and Franco, J. (2018). Machine Learning Techniques for PM10 Levels Forecast in Bogotá. 2018 ICAI Workshops (ICAIW).

[7] T. Q. Chen, and C. Guestrin, XGBoost:A Scalable Tree Boosting System, in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016: 785-794.

[8] J. Ye, J. Chow, and J. Chen, Stochastic gradient boosted distributed decision trees, in Proceedings of the 18th ACM conference on INF.

[9] Esri.com. (2019). [online] Available at: https://www.esri.com/library/bestpractices/gis-and science.pdf [Accessed 31 Mar. 2019].

[10] Postranecky, M. and Svitek, M. (2017). Smart city near to 4.0 — an adoption of industry 4.0 conceptual model. 2017 Smart City Symposium Prague (SCSP).