Airnode AI

Advancing Air Quality Monitoring: Predictive Evaluation, Anomaly Detection & Classification for Enhanced Sensor Performance

Contents

1. Executive Summary

Ensuring air quality is crucial for public health and safeguarding the environment in today's world. Poor air quality can lead to serious health problems, such as heart disease, lung cancer, type 2 diabetes, and adverse effects on ecosystems. As urbanization and industrial activities continue to increase, managing air pollution has become increasingly challenging.

Monitoring air quality has become critical to mitigate these health issues and environmental impacts. Our project with AirNode addresses this issue by leveraging advanced machine learning techniques to monitor air quality and predict future trends. The dataset comprises different air quality parameters for one location and relevant features that enhance prediction accuracy.

The aim is to develop an air quality monitoring system that involves predictive algorithms to understand historical air quality data trends, forecast future trends, and identify sensor failures. Additionally, clustering algorithms were used to identify potential anomalies.

The results from the CNN and LSTM models show a significant correlation between windspeed, wind direction, and relative humidity in predicting PM2.5 levels. By forecasting future PM2.5 values, further studies can be conducted, such as creating an alert system for high pollutant levels and suggesting people stay indoors or wear masks.

The results from the DBSCAN and K-Means models indicate that clustering has segregated some data points separately from the rest, suggesting high peaks in the data. Frequent peaks might indicate sensor failure or events like bonfires that caused high pollutant levels. However, our results did not detect any continuous lows or high peaks, indicating no anomalies or sensor failures.

In summary, our project leverages machine learning techniques to monitor air quality, predict future trends, and identify potential anomalies or sensor failures. This approach can contribute to mitigating health and environmental impacts associated with poor air quality.

2. Background and Problem Statement

AirNode, a leading organization in the field of air quality monitoring, uses advanced air quality (AQ) sensors to collect and analyse data from various locations. These sensors provide valuable insights into the extent of pollutants in air such as particulate matter (PM2.5 and PM10), nitrogen dioxide (NO2), and other harmful pollutants.

Despite the advancements in Air Quality (AQ) sensor technology, several challenges need to be addressed to ensure the accuracy and reliability of air quality measurements. Sensor malfunctions, drifts, or other anomalies can compromise the quality of the collected data. More sophisticated analytical techniques are required to distinguish between actual pollution incidents and sensor anomalies. Addressing these issues is crucial for improving air quality monitoring and future decision-making.

To enhance AirNode's sensor's ability to evaluate and mitigate air pollution, this project aims to develop and implement advanced machine learning techniques for:

1. Evaluating AQ sensor measurements and detecting anomalies.
2. Predicting when sensors may fail to take accurate measurements.
3. Classifying air quality measurement anomalies using semi-supervised learning into categories of emission sources and anomalies.

3. Data source and Methodologies

The primary data sources were official reports compiled by individual state pollution control boards across India. These reports, containing air quality measurements and other relevant details, were sent to a central control room managed by the Government of India for air quality monitoring and management at the national level.

Methodology

Data Preprocessing: Preprocessing was done on the data to ensure consistency and accommodate missing numbers by interpolation. To appropriately assess the performance of the model, we divided the data into training and testing sets (80%–20%) and (70%–30%).

Predictive Analysis: We used the Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM) models for predictive analysis. Because LSTM can manage temporal dependencies in sequential data, it is a perfect fit for forecasting future levels of air quality. By

using the CNN model to identify spatial patterns in the data, we were able to increase the accuracy of our predictions.

Clustering Analysis: K-Means and DBSCAN algorithms were used to carry out the clustering study. By using K-Means to group data into clusters according to shared features, common trends in air quality might be found. By detecting outliers that did not fit into any cluster, which may indicate probable sensor malfunctions or uncommon air quality events, DBSCAN, a density-based clustering algorithm, was utilized to detect discrepancies.

Sensor Failure Detection: The residual errors between the actual and anticipated values were analysed, and this allowed us to set an anomaly detection threshold. This threshold is set at ±3 times the residuals' standard deviation. When there was more error than this threshold between the observed and anticipated readings, it suggested that there might have been a malfunction or failure of the sensor.

Model Evaluation: Our models' performance was evaluated using measures including silhouette scores for clustering models and Root Mean Squared Error (RMSE) for regression models, which ensured accurate and dependable predictions and anomaly detections.

4. Data Exploration and Visualization

Dataset Overview

The dataset contains a total of 25 columns or features.

Statistical Summary:

	PM2.5 (µg/m³)	PM10 (µg/m³)	NO (µg/m³)	NO2 (µg/m³)	NOx (ppb)	\
count	7973.000000	7947.000000	8274.000000	8292.000000	8293.000000	
mean	27.378806	64.228514	4.630029	20.418979	25.035350	
std	18.093457	51.739076	6.463143	15.409832	19.318278	
min	0.010000	5.520000	0.015000	0.010000	0.000000	
25%	19.410000	41.405000	2.300000	9.804583	12.247500	
50%	25.690000	54.462500	2.772500	14.995000	17.510000	
75%	33.165000	75.832500	4.469375	27.011250	32.142500	
max	767.735000	972.940000	82.015000	145.485000	156.930000	

	NH3 (µg/m³)	SO2 (µg/m³)	CO (mg/m³)	Ozone (µg/m³)	Benzene (µg/m³)	\
count	8263.000000	7469.000000	8697.000000	8152.000000	7370.000000	
mean	11.093191	19.548847	0.735717	32.301283	0.463858	
std	6.943443	26.027317	0.502356	29.316496	0.174497	
min	0.010000	0.015000	0.000000	0.010000	0.000000	
25%	7.047500	6.452500	0.412500	12.380000	0.445000	
50%	9.167500	8.890000	0.685000	24.776250	0.500000	
75%	13.710000	24.545000	0.915000	44.316875	0.502500	
max	121.155000	180.492500	9.880000	199.440000	2.272500	

	...	MP-Xylene (µg/m³)	AT (°C)	RH (%)	WS (m/s)	WD (deg)	\
count	...	7355.000000	0.0	8707.000000	8707.000000	8707.000000	
mean	...	0.308548	NaN	76.290942	1.743988	193.585882	
std	...	0.101008	NaN	16.180808	0.317540	68.989816	
min	...	0.042500	NaN	25.190000	0.592500	0.085000	
25%	...	0.292500	NaN	63.912500	1.527500	150.265000	
50%	...	0.300000	NaN	80.757500	1.860000	192.345000	
75%	...	0.307500	NaN	91.157500	1.983333	235.730000	
max	...	1.530000	NaN	99.840000	2.517500	354.000000	

	RF (mm)	TOT-RF (mm)	SR (W/mt2)	BP (mmHg)	VWS (m/s)
count	0.0	0.0	5794.000000	8707.000000	8707.000000
mean	NaN	NaN	119.366165	1006.614149	-0.485019
std	NaN	NaN	193.033613	4.884281	0.017203
min	NaN	NaN	0.010000	800.000000	-0.502500
25%	NaN	NaN	1.580000	1003.625000	-0.500000
50%	NaN	NaN	26.877500	1006.475000	-0.490000
75%	NaN	NaN	140.483125	1009.625000	-0.475000
max	NaN	NaN	918.605000	1016.550000	-0.392500

Figure 1: Descriptive summary statistics

Missing Value Analysis

The image below represents the missing value analysis of the data, where we can observe that some features have many missing values, nearly 100%.

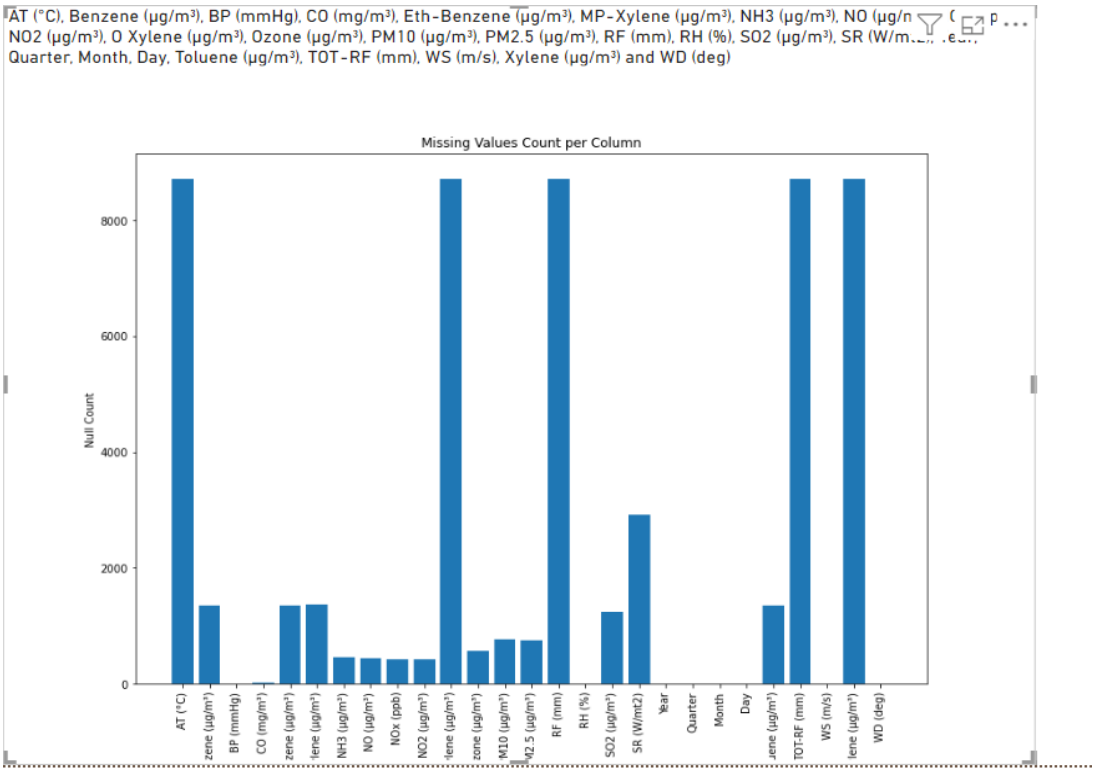

Figure 2: Missing value analysis

Missing Values:	
PM2.5 (µg/m³)	787
PM10 (µg/m³)	813
NO (µg/m³)	486
NO2 (µg/m³)	468
NOx (ppb)	467
NH3 (µg/m³)	497
SO2 (µg/m³)	1291
CO (mg/m³)	63
Ozone (µg/m³)	608
Benzene (µg/m³)	1390
Toluene (µg/m³)	1390
Xylene (µg/m³)	8760
O Xylene (µg/m³)	8760
Eth-Benzene (µg/m³)	1401
MP-Xylene (µg/m³)	1405
AT (°C)	8760
RH (%)	53
WS (m/s)	53
WD (deg)	53
RF (mm)	8760
TOT-RF (mm)	8760
SR (W/mt2)	2966
BP (mmHg)	53
VWS (m/s)	53
dtype: int64	

Figure 3: Missing values of different parameters

Basic Exploration of PM2.5, PM10, and NO2 Feature Distributions over Time

Analysis regarding the distribution of PM2.5, PM10, and NO2 features over the given time period.

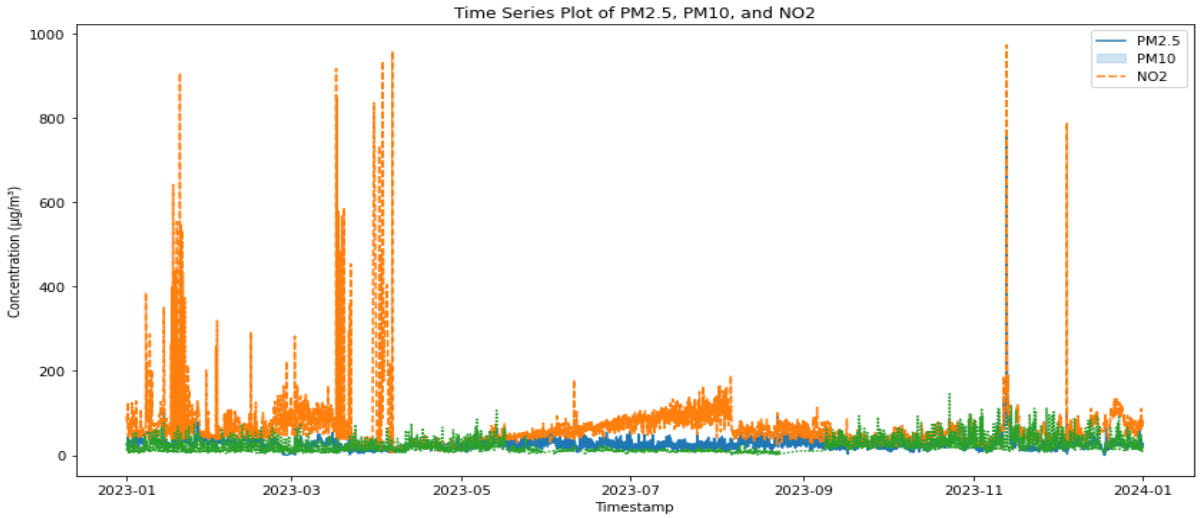

Figure 4: Time series plot

Correlation Matrix

A correlation matrix was generated to identify potential multicollinearity among the predictor variables and to explore the strength and direction of relationships between them.

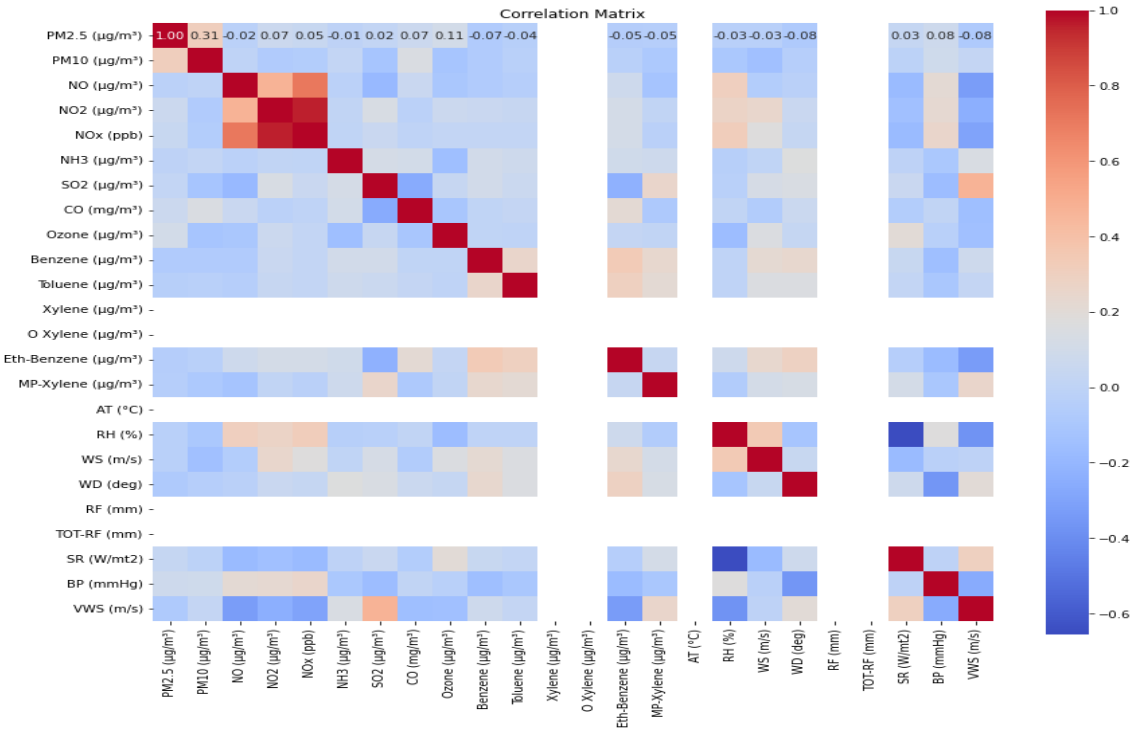

Figure 5: Correlation heat map

5. Modelling and Analysis

5.1 Long Short-Term Memory (LSTM)

The LSTM model was selected for our study because it can effectively predict PM2.5 levels in air quality datasets by capturing long-term dependencies in sequential data. LSTM networks are a kind of recurrent neural network (RNN) that preserves information across extended sequences to learn from previous observations and make precise predictions about the future. The LSTM model had high generalization skills with relatively low RMSE values for both training and test datasets, especially with variable test sizes. It was able to predict future PM2.5 levels by using prior air quality measurements.

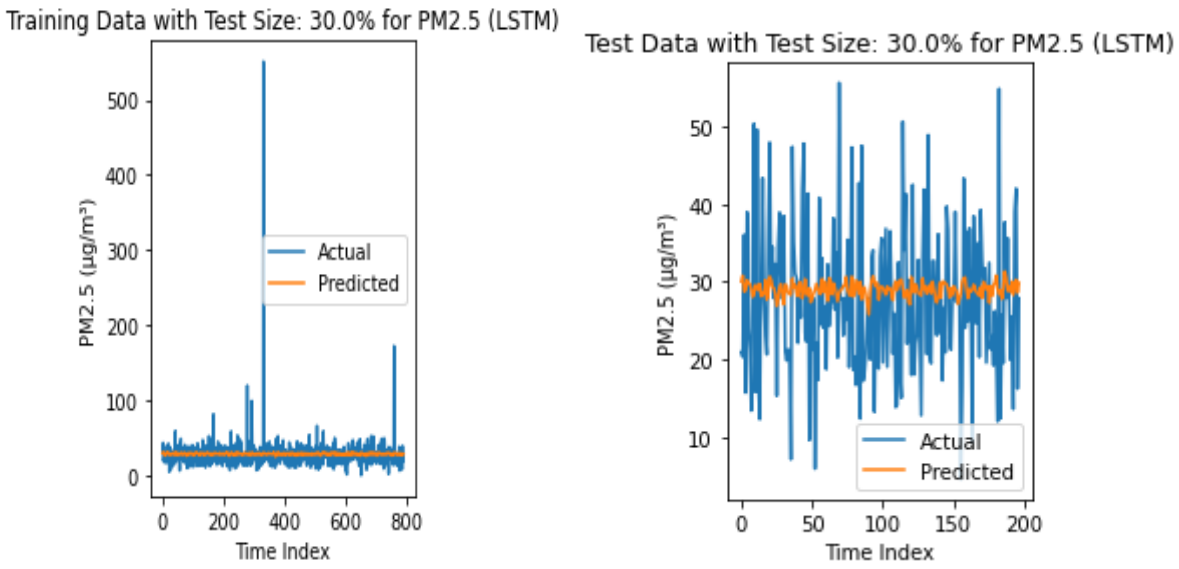

Figure 6: Actual vs Predicted graph of LSTM

5.2 Convolutional Neural Network (CNN)

The CNN model was employed to identify complex patterns present in the data related to air quality. CNNs are excellent at capturing spatial hierarchies, which are evident in time series data as well. They have historically been employed for visual information. In our investigation, this model showed to be especially effective; when applied to a 20% test set, the CNN achieved the lowest test RMSE of **15.510**. This better performance suggests that CNNs can model complex patterns in air quality data very well, which can result in forecasts that are more reliable and precise.

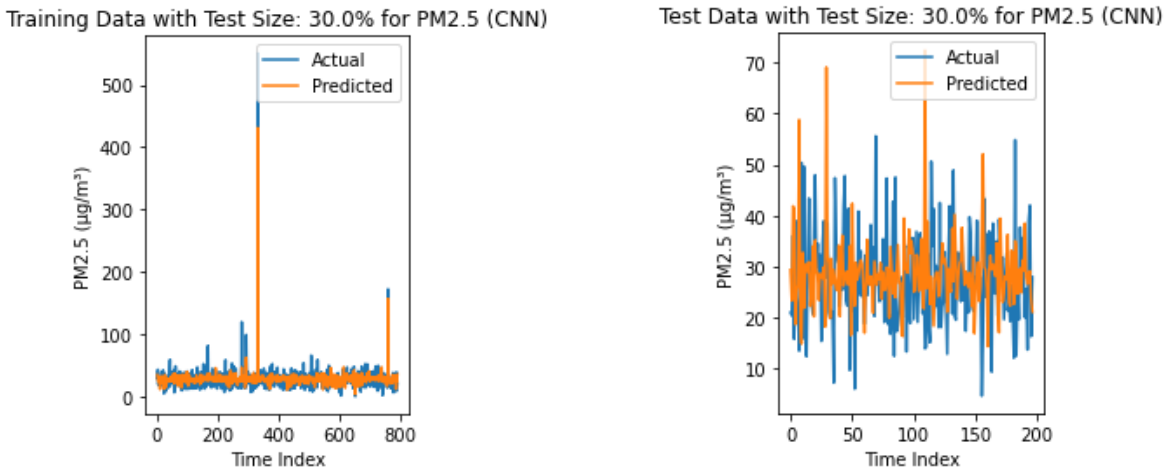

Figure 7: Actual vs Predicted graph of CNN

5.3 K-Means

The K-Means technique was used to divide the air quality data into different categories according to wind speed and levels of PM2.5 to discover anomalies. With a silhouette score of **0.390**, K-Means, a centroid-based clustering method, offered decent clustering effectiveness. K-Means assisted in locating possible anomalies in the air quality data by highlighting clusters that differed noticeably from the norm. These unusual patterns may have been associated with specific times or circumstances.

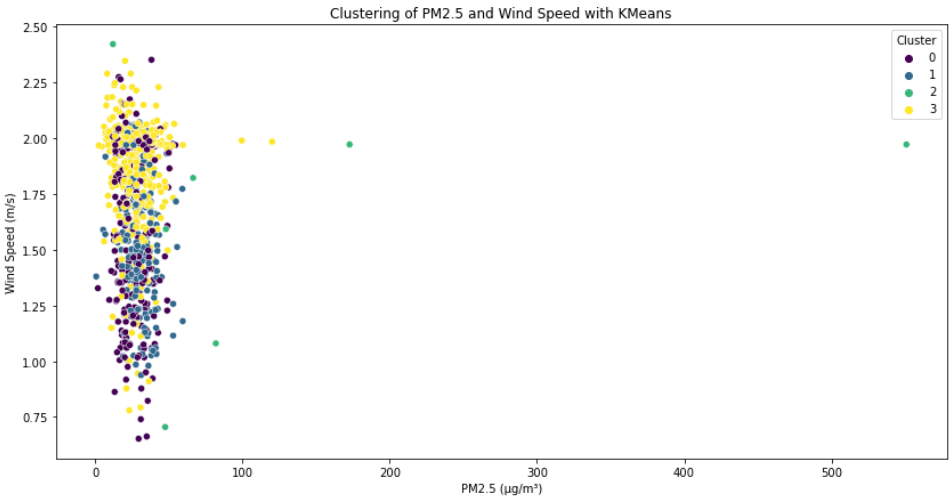

Figure 8: K-means Clustering of PM2.5

5.4 DBSCAN

Finally, to find clusters with different densities and outliers, the Density-Based Spatial Clustering of Applications with Noise (DBSCAN) technique was employed. DBSCAN outperformed K-Means, obtaining a **0.630** silhouette score, which was higher. This suggests

that DBSCAN performed better at handling noise in the air quality data and differentiating across clusters with varying densities. DBSCAN is an asset for identifying significant clusters and patterns in the dataset and for gaining a deeper understanding of the distribution and variability of air quality measurements due to its robustness and outlier management capabilities.

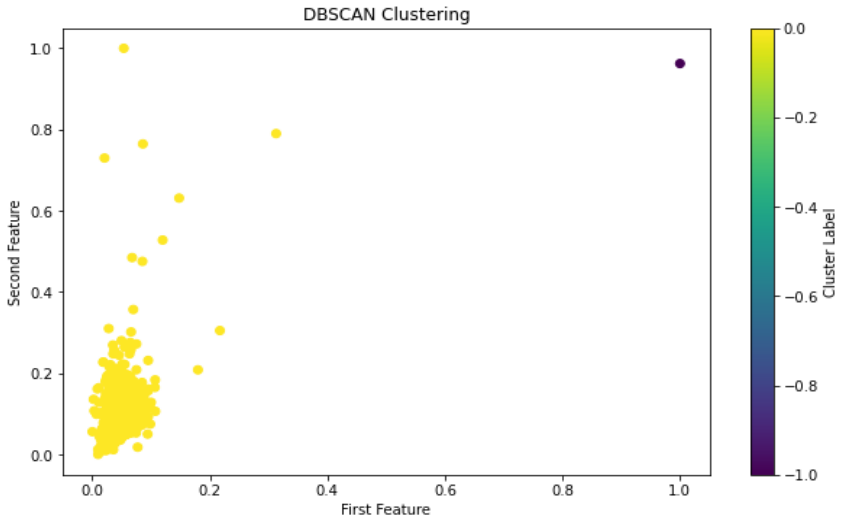

Figure 9: DBSCAN Clustering

The clustering visualization with K-Means showed distinct groupings of data points, highlighting the relationship between PM2.5 levels and wind speed. The DBSCAN clustering further refined these groupings, identifying more precise and meaningful clusters that can inform targeted environmental interventions.

Evaluation

All the four models discussed above were built successfully to obtain the results that would depict how well the model contributes to our objectives.

Model	Train RMSE	Test RMSE	Silhouette Score
LSTM (20% test size)	16.970	15.754	N/A
LSTM (30% test size)	16.946	15.701	N/A
CNN (20% test size)	15.693	15.510	N/A
CNN (30% test size)	16.743	15.617	N/A
K-Means	N/A	N/A	0.390
DBSCAN	N/A	N/A	0.630

Table 1: Evaluation metrics

The performance of each model is summed up in the evaluation table. The accuracy of LSTM and CNN models in predicting PM2.5 levels was evaluated by utilizing RMSE values for both training and test sets. CNN demonstrated its efficacy by achieving the lowest test RMSE, particularly when testing with a 20% sample size. With a higher silhouette score for clustering than K-Means, DBSCAN scored better in terms of cluster integrity and separation. These findings highlight how well one model complements the others in our thorough investigation of air quality.

6. Insights and Interpretation

Results of all the models that were performed are compared to figure out which would be the best model.

Upon assessing the LSTM model, we found that the test RMSE was 15.754 and the train RMSE was 16.970. The next step of increasing the test size to 30%, the train RMSE was slightly improved to 16. 946 and the test RMSE was improved to 15.701. These findings helped in proving that, the model performs better with bigger test set and effectively generalizes to new data without any overfitting, this indicates a fair balance between training and testing data. Depending on the test size, the CNN model showed a significant difference in performance. The test RMSE was 15. 617 and the train RMSE was 16.743. On the other hand, the CNN model produced the lowest test RMSE of all configurations with a test size of 20%, achieving a test RMSE of 15. 510 and a train RMSE of 15.693. This suggests that the CNN model **outperforms** the LSTM model when applied to a smaller test set, demonstrating its superior ability to identify complex patterns in the air quality data.

As a result, the CNN model with a 20% test size was chosen for air quality forecasts since it offered the **best predictive accuracy** and the **lowest test RMSE** of 15.510. With comparable train and test RMSE values, which show less overfitting, both the **CNN** and **LSTM** models showed strong generalization ability. As **DBSCAN** has a **higher silhouette score** than other algorithms, it is better suited for detecting meaningful clusters in the air quality dataset when it comes to clustering jobs.

Future Prediction:

LSTM and CNN model was further developed to predict the future air quality value that is for the next 100 hours.

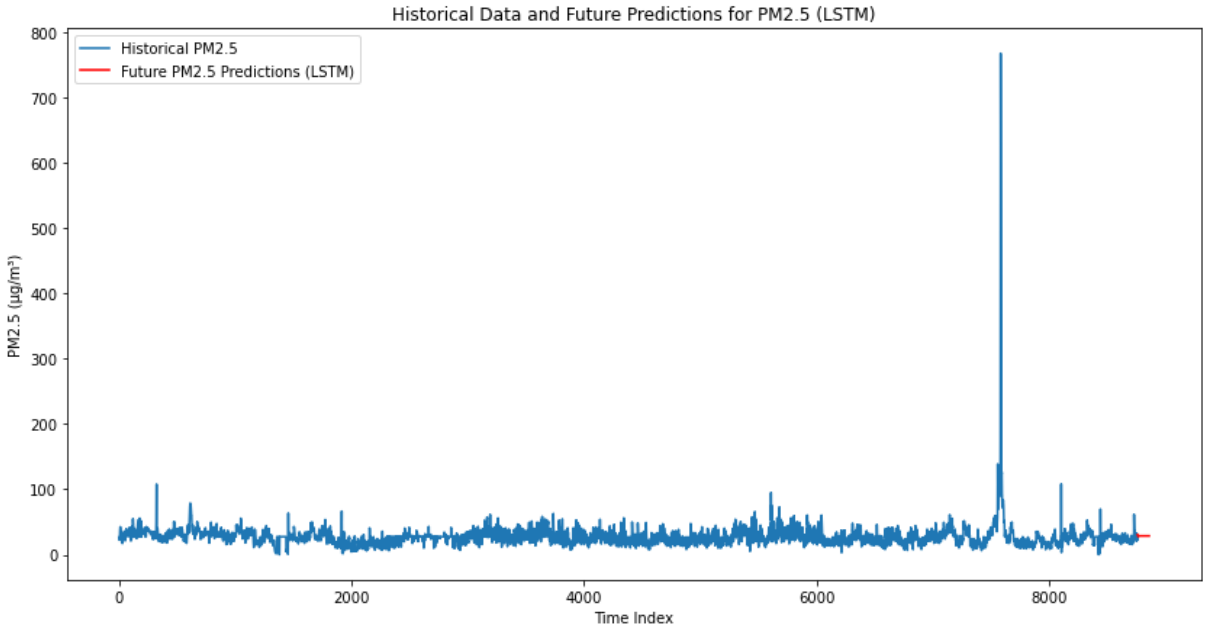

Figure 10: Future values of PM2.5 using LSTM

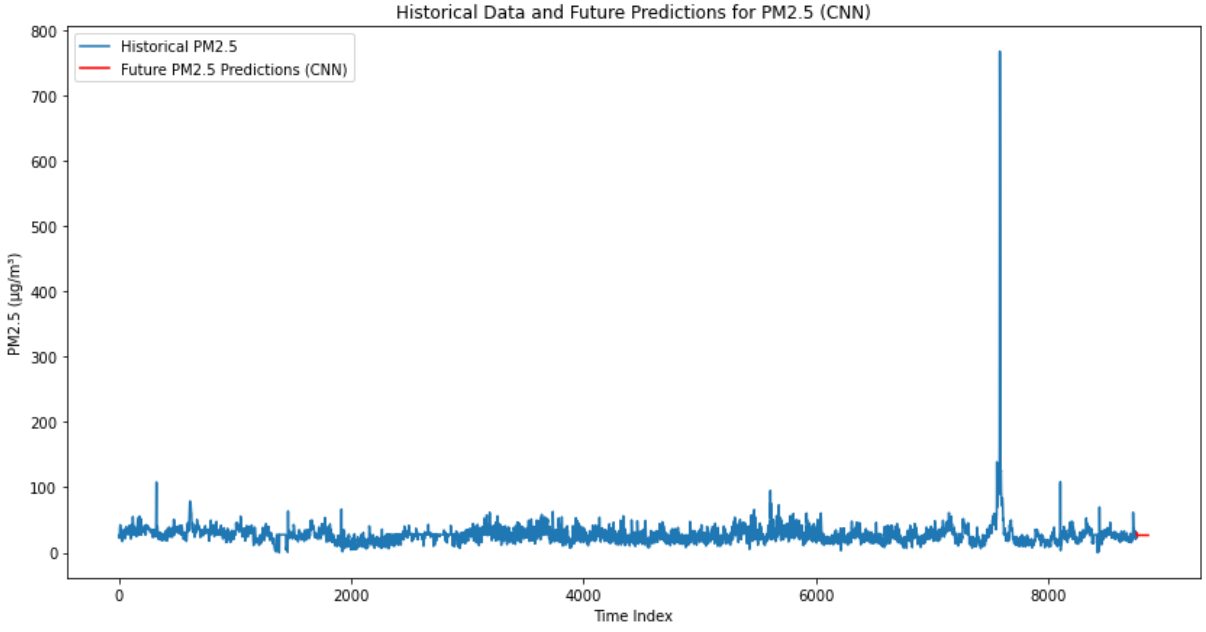

Figure 11: Future values of PM2.5 using CNN

Actual and anticipated PM2.5 levels closely match on the graph for forecasting future air quality values over the next 100 hours using LSTM and CNN models. Both models demonstrate a reasonable level of accuracy in predicting short-term air quality, which is important for prompt environmental monitoring and action. They also successfully capture trends and fluctuations.

Senor failure detection:

Apart from doing predictive and clustering analysis, we also focused on identifying sensor malfunctions, which is an essential part of maintaining accurate air quality monitoring. By forecasting PM2.5 levels and contrasting them with actual sensor readings, we used the LSTM model and CNN model to find probable sensor failures.

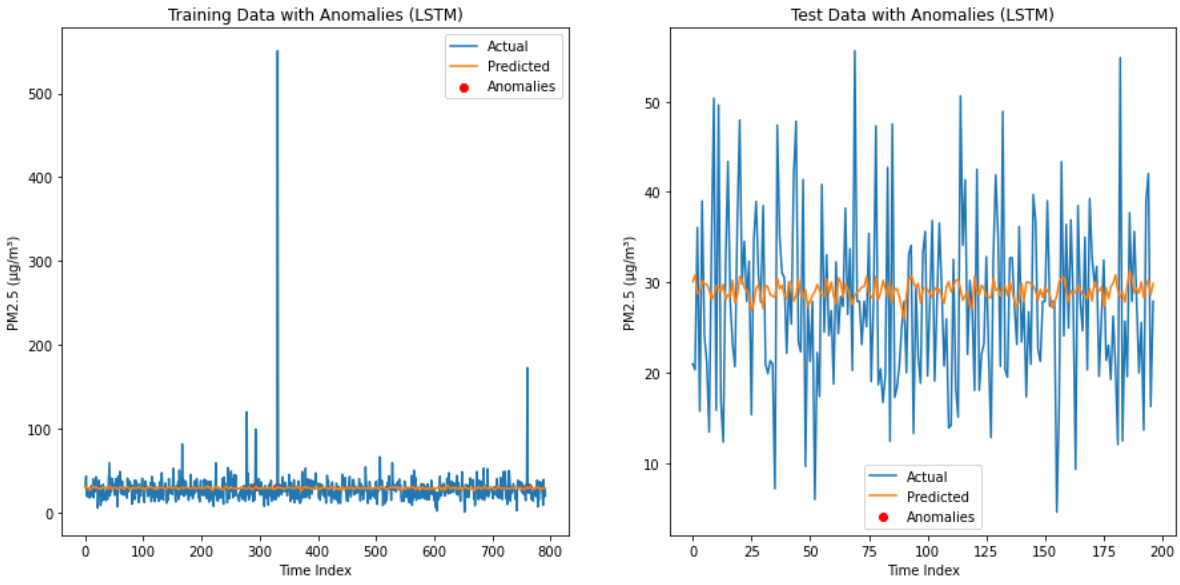

Figure 12: Sensor failure detection using LSTM

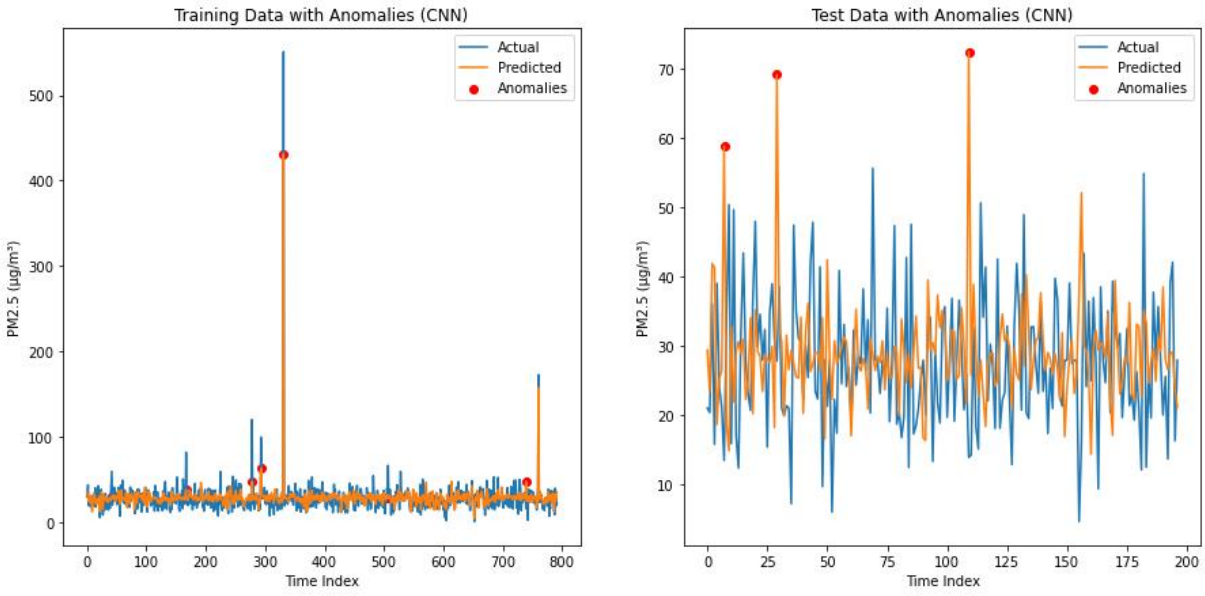

Figure 13: Sensor failure detection using CNN

7. Recommendations

We recommend the following measures to improve air quality monitoring and management:

- Use the CNN and LSTM models to construct a real-time air quality monitoring system. To safeguard the public's health, this system will continuously gather and process data, producing precise forecasts and timely notifications.
- Create an improved visualization dashboard that presents historical patterns, predicted insights, and real-time information in an understandable manner. Lastly, scale this solution to different regions, customizing models for regional variations to support broader environmental monitoring efforts.

8. Limitations

1. **Assumption of Anomalies**: Our approach may miss some forms of anomalies, such as abrupt pollution instances or environmental disruptions since it interprets anomalies in air quality data as sensor faults.
2. **Reliance on historical data:** Models' extensive reliance on past data may cause them to overlook new trends or abrupt changes in air quality that were not recorded in the training set.
3. **Limited Generalizability**: Due to variances in air quality dynamics caused by variables like geography and climate, models may not generalize effectively to different geographic areas or time periods.
4. **External factors Ignored**: The methodology limits the models' accuracy in real-world scenarios by ignoring external elements that can have a substantial impact on air quality, such as weather or industrial activity.
5. **Continuous Refinement**: To overcome these shortcomings and guarantee the models' efficacy in a variety of environmental scenarios, ongoing adaptation and refinement are required.

9. Conclusion

In conclusion, our project provides a thorough approach to efficient air quality monitoring. Both our CNN and LSTM models have performed well:

- The CNN model has the highest predicted accuracy. It has been effective to use DBSCAN for clustering, which brings up certain patterns in the air quality data.

- By implementing our recommendations into practice, air quality monitoring will become more dependable and efficient, resulting in communities that are healthier and more sustainable.
- We can solve environmental issues and make major progress in preserving the environment and public health by utilizing advanced analytics and machine learning.

10. References

1. airquality.cpcb.gov.in. (n.d.). CCR. [online] Available at: https://airquality.cpcb.gov.in/ccr/#/caaqm-dashboard-all/caaqm-landing/caaqm-data-repository
2. Samad, A., Garuda, S., Vogt, U. and Yang, B. (2023). Air pollution prediction using machine learning techniques – An approach to replace existing monitoring stations with virtual monitoring stations. *Atmospheric Environment*, [online]310,p.119987. https://doi.org/10.1016/j.atmosenv.2023.119987
3. Gilik, A., Ogrenci, A.S. and Ozmen, A. (2021). Air quality prediction using CNN+LSTM-based hybrid deep learning architecture. Environmental Science and Pollution Research. https://doi.org/10.1007/s11356-021-16227-w
4. Sokhi, R., Moussiopoulos, N., Baklanov, A., Bartzis, J., Coll, I., Finardi, S., Friedrich, R., Geels, C., Grönholm, T., Halenka, T., Ketzel, M., Maragkidou, A., Matthias, V., Moldanova, J., Ntziachristos, L., Schäfer, K., Suppan, P., Tsegas, G., Carmichael, G. and Franco, V. (2022). at-mospheric chemist, awarded the Nobel Prize in Chemistry 1995; Mario Molina (1943-2020), atmospheric chemist, awarded the No-bel Prize in. Atmos. Chem. Phys, [online] 22(7), pp.4615–4703. doi: https://doi.org/10.5194/acp-22-4615-2022.
5. Kumar, K. and Pande, B.P. (2022). Air pollution prediction with machine learning: a case study of Indian cities. International Journal of Environmental Science and Technology. [online] doi: https://doi.org/10.1007/s13762-022-04241-5

11. Appendix

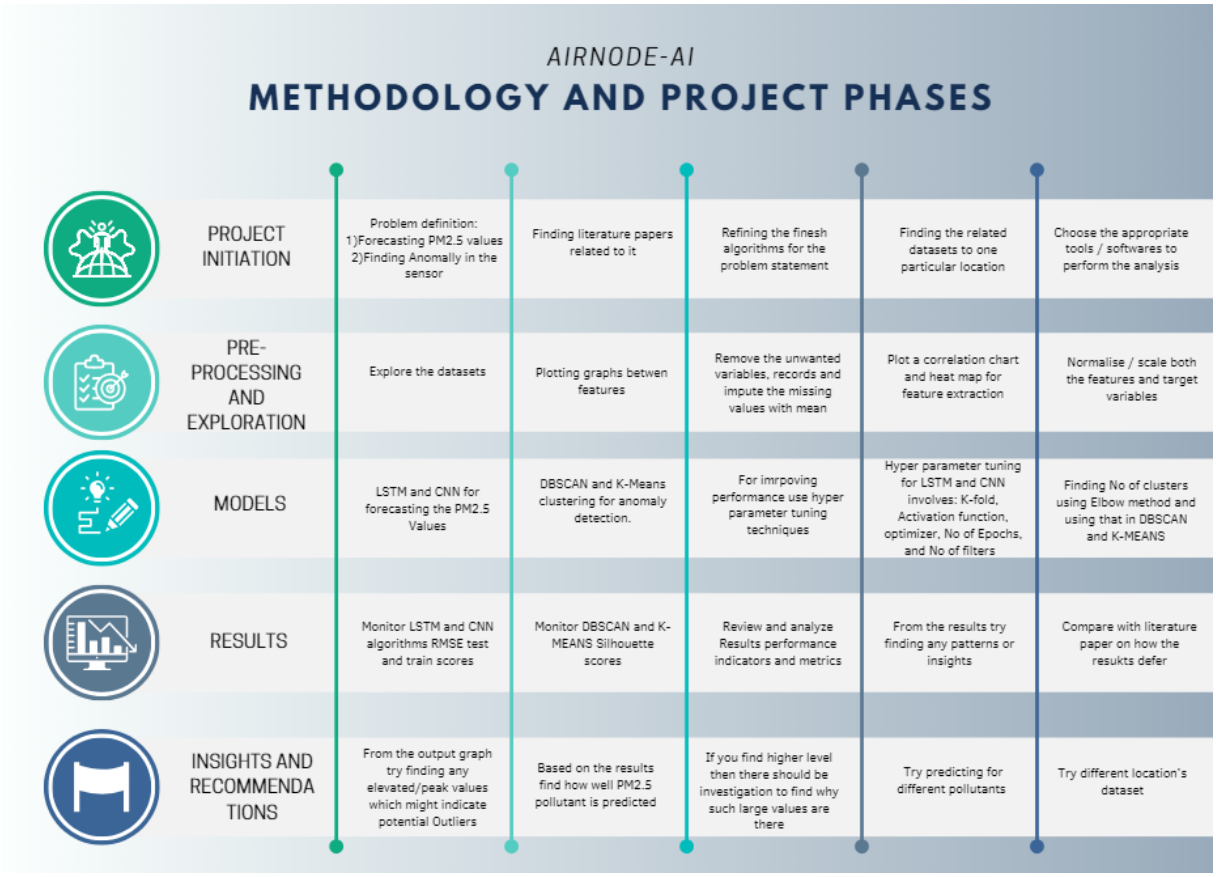

Figure 14: Generalised flowchart of the project