India’s Air Analysis and Prediction of Air Quality Index (AQI) in Indian cities

Team DataWatch

Tony P Joy (tonyjoy@iisc.ac.in), Ashutosh Mishra (ashutoshmish@iisc.ac.in), Vivek H N (vivekhn@iisc.ac.in) , Neeraj Shete (neerajshete@iisc.ac.in)

Motivation:

Globally, air pollution is a silent killer. The air pollution levels in India are among the highest in the world, posing a heavy threat to the country's health and economy. All of India’s 1.4 billion people are exposed to unhealthy levels of harmful pollutants - emanating from multiple sources. In this project we are trying to analyse and understand the levels of major air pollutants and the resultant AQI (Air Quality Index) across the major cities in India. Through this analysis we intend to identify the patterns and observations related to the AQI that can provide the data backing to prepare a search/ warn system for the citizens on the air quality at a given date and time of the year and remind them about wearing their N95 masks.
While the depleting air quality is a growing concern in India it doesn’t seem to get the required attention both from the responsible sector and the sufferer sector. We believe that the relevant data and live warnings – when incorporated with the daily digital tools like news feeds, tweets and maps can help to bring the desired emphasize and drive actions.

Design Goals:

Analyse and Visualize:

- The major pollutants for each city and compare the average value of each pollutant across the cities.
- Yearly trend for vehicular and industrial pollution to see if we have any cities showing positive of negative trend.
- Analyse city-wise AQI trends and seasonal changes.

Build ML models to:

- Predict the AQI for a given city given the pollutant features.
- Evaluate the performance metrics of different regression models and select the best model for deployment.
- Fine tune the best regression model through hyperparameter tuning.
- Train a LSTM model to forecast AQI for a given duration.
- Utilize spark streaming for real-time prediction of AQI using best model identified.

Features & Data sources:

Data is downloaded from “Central Control Room for Air Quality Management - All India”. Link: CCR (cpcb.gov.in)

We are fetching data for 6 cities – Bengaluru, Hyderabad, Chennai, Mumbai, Kolkata, and Delhi. We are collecting yearly dataset for each of these cities for the years 2019 through 2023. The dataset contains air quality features collected on an hourly basis making it ~8k samples for a city per year (~40k samples across 6 cities).

Features available in the dataset: *Timestamp, Measurement of different pollutants (PM2.5, PM10, NO, NO2, NOx, NH3, SO2, CO, Ozone, Benzene, Toluene, Xylene, O Xylene, Eth-Benzene, MP-Xylene), AT (°C), RH (%), WS (m/s), WD (deg), RF (mm), TOT-RF (mm), SR (W/mt2), BP (mmHg), VWS (m/s))*

Scalability/Performance goals:

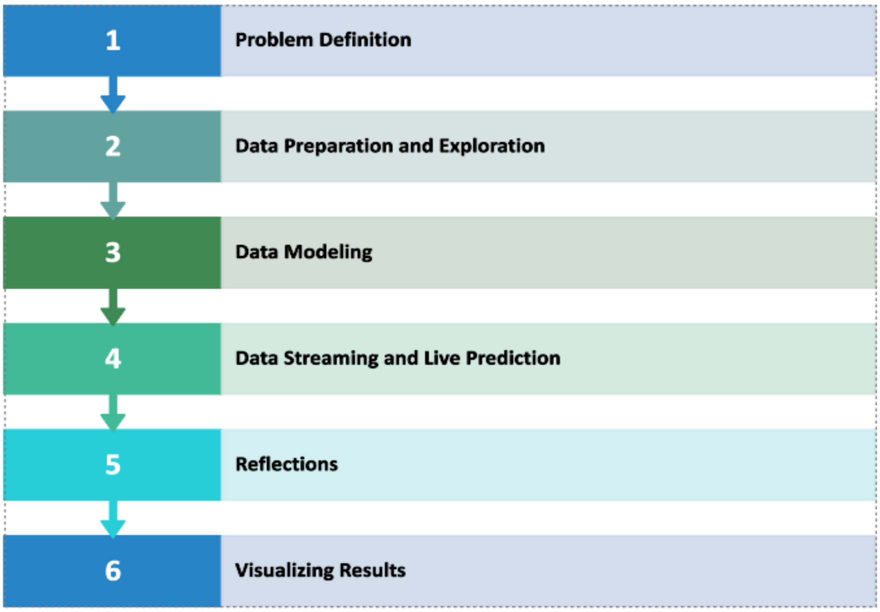
Scalability:

- Models should be scalable to support high data volume and trainable in a distributed computing environment.
- ML Pipeline should be generic to train on additional cities given the same pollutant features.
- Streaming pipeline should be generic enough to support different event sources and sinks – eg: Kafka integration.
- Streaming pipeline should be scalable to handle high input volume and velocity and should support distributed streaming.

Performance goals:

- For the given dataset – Single Model train and validation shouldn't exceed 5 minutes.
- For ensemble models training and validation should be done under 20 minutes.
- Real Time predictions should be in the order of milliseconds for input size < 20 entries.
- Single Model Accuracy on Training Validation should be greater than 80%
- Single Model Accuracy on Testing should be greater than 75%
- Ensemble Model Accuracy on Testing should be greater than 80%.

High-level design:



Data Preparation

The dataset for this study was obtained from the "Central Control Room for Air Quality Management - All India," hosted on the **CPCB** platform. This open-access government repository is well-maintained, providing a comprehensive collection of historical air quality data, making it ideal for our analysis. Six major Indian cities—**Bengaluru, Hyderabad, Chennai, Mumbai, Kolkata, and Delhi**—were selected for focused study. The data spans the years **2019 through 2023**, offering **hourly records**, resulting in approximately **8,000 samples per city per year**.

Data Retrieval and Storage

Datasets were segmented into **training data (2019–2022)** and **testing data (2023)**. These yearly datasets were consolidated into master datasets for training and testing, stored on a shared **Google Drive** for easy access and collaboration.

Data Cleaning

- Threshold-Based Clean-up: Columns with more than 10% missing values were removed to ensure data quality and reduce noise in the dataset.
- Further, missing values were replaced with zeros post sub-indices calculation.

Feature Engineering

The feature AQI was calculated based on the government provided methodology https://cpcb.nic.in/upload/national-air-quality-index/AQI-Calculator.xls.
AQI (Air Quality Index) and **AQI classification** were derived for a holistic view of pollution levels.

Contributions of **vehicular** and **industrial pollution** were estimated to understand sectoral impacts.

- Vehicular pollution = PM2.5 + NOx + CO + SO2
- Industrial Pollution = Ozone + Benzene + NH3 + SO2 + PM10

This meticulous data preparation laid the foundation for developing robust models to analyse and predict air pollution trends in Indian cities.

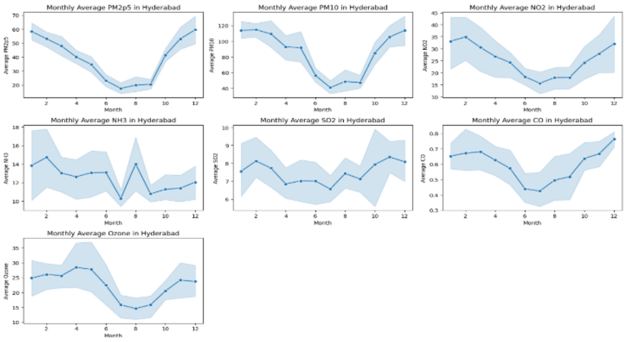
Data Exploration:

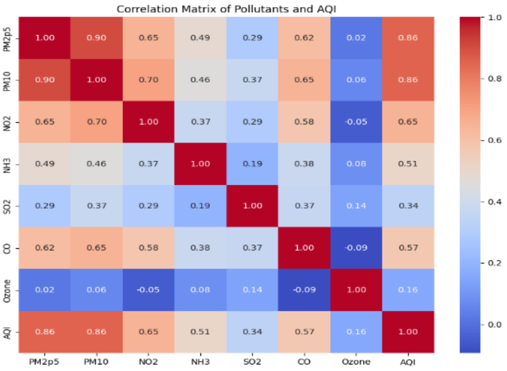
Distribution and Trends of Pollutants

City-wise plots were generated to visualize the distribution of pollutants, revealing the dominant pollutants in each city. Further analysis included:

- **Average levels** of major pollutants across cities.
- Year-over-year changes in pollutant levels to identify trends over time.
- **Monthly averages** and **seasonal variations**, highlighting peak pollution periods for each city.

Study showed that PM pollutants exhibit higher levels in the winter months and lower during the rains. The values for NH3 and SO2 goes highest during the summer months.





	PM2p5	PM10	NO2	NH3	SO2	CO	Ozone	AQI
PM2p5	1.00	0.90	0.65	0.49	0.29	0.62	0.02	0.86
PM10	0.90	1.00	0.70	0.46	0.37	0.65	0.06	0.86
NO2	0.65	0.70	1.00	0.37	0.29	0.58	-0.05	0.65
NH3	0.49	0.46	0.37	1.00	0.19	0.38	0.08	0.51
SO2	0.29	0.37	0.29	0.19	1.00	0.37	0.14	0.34
CO	0.62	0.65	0.58	0.38	0.37	1.00	-0.09	0.57
Ozone	0.02	0.06	-0.05	0.08	0.14	-0.09	1.00	0.16
AQI	0.86	0.86	0.65	0.51	0.34	0.57	0.16	1.00

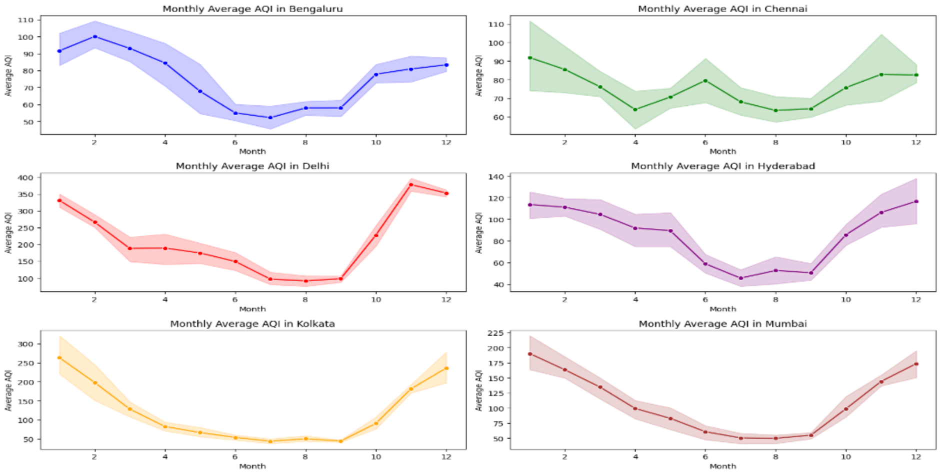
Correlational Analysis of Pollutants against AQI

The Air Quality Index (AQI) has strong positive correlations (above 0.8) with PM2.5 and PM10, indicating that these particulate matter pollutants are major contributors to the overall air quality index. The correlations with other pollutants are more moderate, suggesting AQI is influenced by a combination of different pollutants.

Analysis of City-wise AQI Values

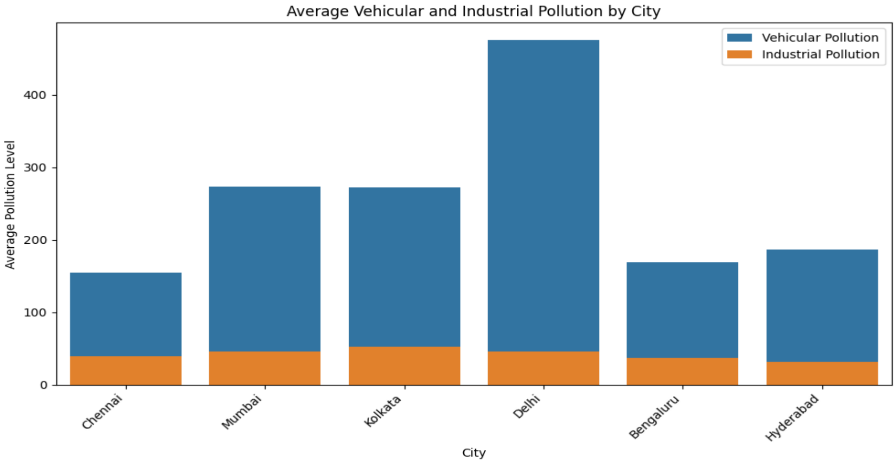
City-specific AQI values were explored through:

- **Monthly AQI distributions**, showcasing variability throughout the year.
- **City-wise AQI comparisons**, identifying cities with the highest and lowest air quality levels.



Vehicular and Industrial Pollution Studies

To understand the contribution of anthropogenic activities:



As plotted above, Delhi stands out with the highest pollution levels, driven by vehicular emissions that far exceed other cities, underscoring the city's critical need for interventions in traffic and transportation management This exploratory analysis provided valuable insights into spatial, temporal, and seasonal air quality dynamics, forming a basis for deeper analysis and predictive modelling.

Data Modelling:

The modelling process focused on developing, training, and comparing a variety of machine learning models to predict air quality trends. The dataset, spanning **2019–2022**, was used for training and testing, while the **2023 dataset** served as a validation set.

Model Selection

Four categories of models were explored:

- **Linear Regression** techniques, including Lasso and Ridge, to capture linear relationships.

- **Tree-based Regression** models like Decision Trees and Random Forests for their ability to handle non-linear patterns.
- **Ensemble Models**, such as Random Forest (with Randomized Search for hyperparameter tuning) known for their robustness and high predictive power.
- **Time Series Forecasting** using PyTorch's LSTM to model temporal dependencies in the data.

Big data platform used:

Apache pyspark, sparlML, sparkStreaming, PyTorch

ML Methods used:

LinearRegression, DecisionTreeRegressor, GeneralizedLinearRegression., ,RandomForestRegressor, RGSCV, LSTM

Evaluation Metrics:

Root Mean Squared Error (RMSE) and **training/testing accuracy** to measure performance. After training and testing, the best-performing model was selected and subjected to hyperparameter tuning to optimize its performance.

Validation

The selected best model (Random Forest regressor) was validated on the **2023 dataset**, with its accuracy on unseen data providing a final measure of its predictive reliability. This structured approach ensured comprehensive comparison and fine-tuning of models, enabling the identification of the most suitable algorithm for forecasting air quality. The process emphasized iterative improvement, leveraging both simple and advanced techniques for optimal results.

Experiment design:

- Each of the identified Single ML models were trained using master train dataset [2019-2022 data].
- A 20% split was performed on the train dataset for validation.
- Trained models were validated using validation split and training accuracy was computed.
- Trained models were then tested using unseen test data [2023 dataset] and testing accuracy was computed.
- Tabulated validation and testing accuracies of different models to identify the best model.
- Identified best model(Random Forest) was then tuned using RGSCV to further improve accuracy.
- LSTM model was trained using train dataset and made to generate forecasts for whole of 2023.
- LSTM generated forecasts for 2023 was plotted against actual AQI values to visualize performance.
- Simulated an input event stream using an application to send hourly rows containing pollutant measurements from test[2023] dataset, and save them to a shared input data source.
- Setup spark streaming to consume new CSV files from shared input store, convert it into spark dataframe and use the best model identified to predict AQI values for the input batch. Spark Streaming will generate an output dataframe containing predicted AQI values for the input batch and write it to a shared output store(sink).
- Simulated a visualization application to monitor the shared output store and for a new csv file – read and plot the predicted AQI values per city for the hour of the Timestamp.

Reflection and Inference:

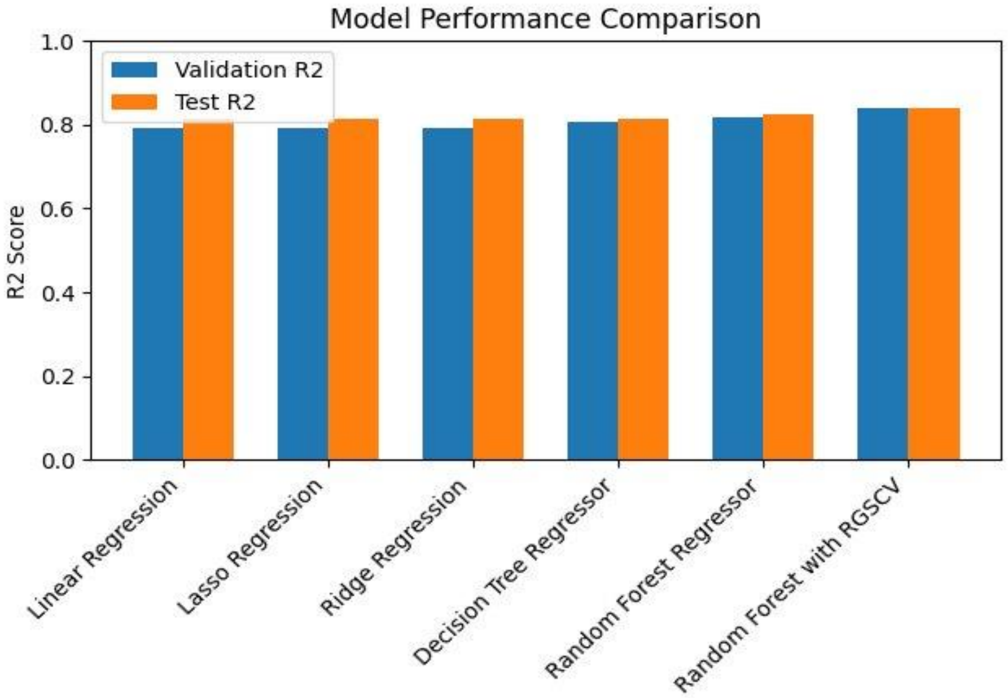
EDA and data Preparation metrics

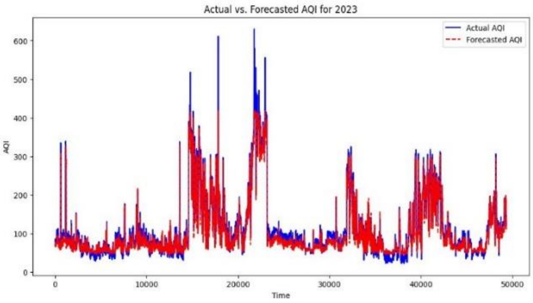
Data Preparation and EDA metrics	**Success Criteria**	**Results Shown**
Data Cleaning and Preprocessing	Achieve 90%+ clean, usable data across selected cities and pollutants	Successfully cleaned the dataset with 0% missing values.

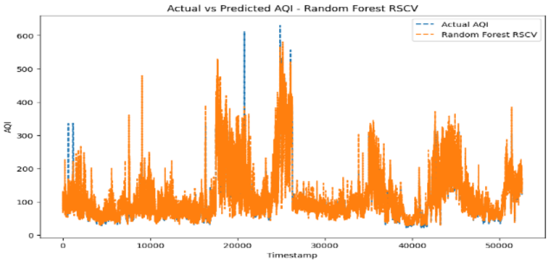
Distribution of Pollutants Across Cities	Understand the distribution of major pollutants for comparison across cities.	Observed flatter distributions in Delhi for PM pollutants; lower values in Bengaluru and Chennai.
Average Value of Major Pollutants for Each City	Visualize and compare average pollutant values across cities.	Found airborne PM as the dominant pollutant; Delhi and Mumbai had the highest averages for most pollutants.
Pollutant Trends (2019-2023)	Compare pollutant trends over the years and identify long-term patterns.	Chennai showed sustained control; Mumbai exhibited upward trends in pollutants; Delhi remained consistently high.
Monthly Pollutant Levels (Seasonal Trends)	Identify seasonal changes and peak months for pollutants.	PM pollutants peaked in winter; NH3 and SO2 levels were highest during summer; monsoon reduced pollution.
AQI Trends Across Cities	Assess monthly AQI trends and identify best and worst months for air quality.	Delhi worsened in winter; June to September showed healthy AQI in most cities except Delhi.
Industrial vs. Vehicular Pollution	Determine dominant pollution sources for each city and their trends.	Industrial pollution exceeded vehicular in all cities; Delhi showed consistently high vehicular pollution levels.
City-Specific Insights	Highlight unique trends for each city.	Chennai and Bengaluru maintained better air quality; Delhi and Mumbai had higher pollutant levels and poor AQI

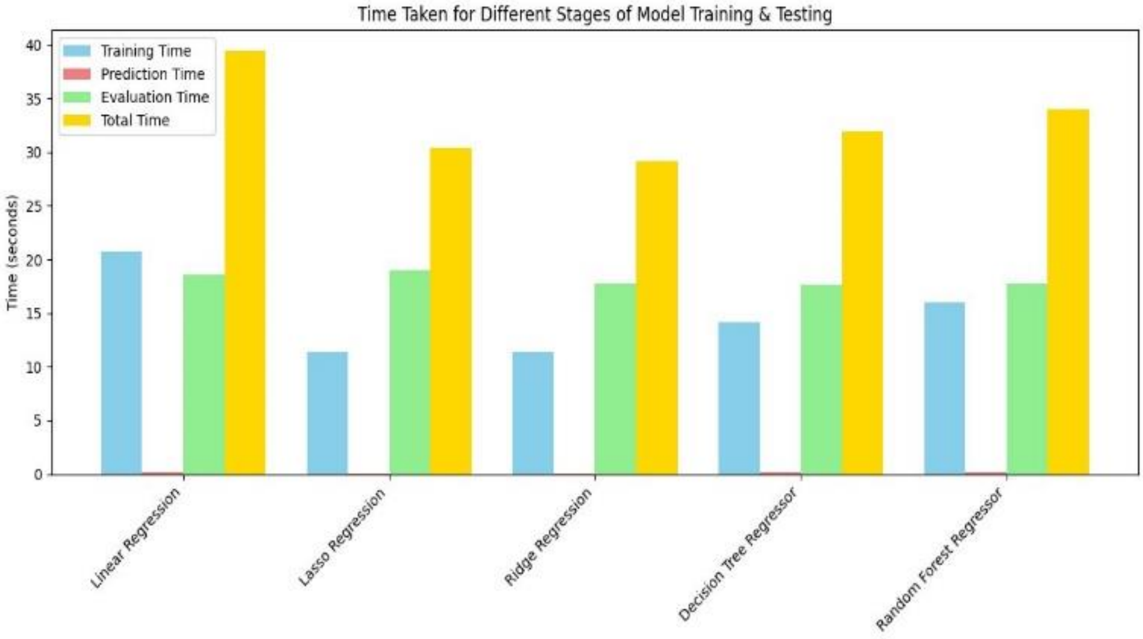
Model Evaluation Metrics:

Models	RMSE		Testing Accuracy		Validation accuracy	
	Achieved	Target	Achieved	Target	Achieved	Target
Linear Regression	39.5	35	0.79	0.75	0.81	0.75
Lasso Regression	39.4	35	0.78	0.75	0.8	0.75
Ridge Regression	39.5	35	0.78	0.75	0.81	0.75
Decision tree regressor	37.9	35	0.8	0.80	0.81	0.70
Random forest with Grid search	37.1	35	0.81	0.80	0.82	0.80
Random Forest with randomized search	34.8	35	0.83	0.80	0.84	0.80


Model Performance Comparison
Validation R2
Test R2
R2 Score
1.0
0.8
0.6
0.4
0.2
0.0
Linear Regression
Lasso Regression
Ridge Regression
Decision Tree Regressor
Random Forest Regressor
Random Forest with RGSCV


Actual vs. Forecasted AQI for 2023
Actual AQI
Forecasted AQI
AQI
600
500
400
300
200
100
0
0
10000
20000
30000
40000
50000
Time


Actual vs Predicted AQI - Random Forest RSCV
Actual AQI
Random Forest RSCV
AQI
600
500
400
300
200
100
0
0
10000
20000
30000
40000
50000
Timestamp


Time Taken for Different Stages of Model Training & Testing
Training Time
Prediction Time
Evaluation Time
Total Time
Time (seconds)
40
35
30
25
20
15
10
5
0
Linear Regression
Lasso Regression
Ridge Regression
Decision Tree Regressor
Random Forest Regressor

Summary:

The report outlines a project focused on analyzing and predicting air quality in six major Indian cities (Delhi, Mumbai, Bengaluru, Hyderabad, Chennai, and Kolkata) using hourly data from 2019 to 2023. It aims to understand pollutant distributions, seasonal trends, and their impact on AQI. The project employs machine learning models like Linear regression, Lasso and Ridge regression, Decision Tree regression, Random Forest regression for AQI prediction and LSTM for AQI forecasting.

Key deliverables include pollutant analysis, AQI trend visualization, and identifying seasonal and city-specific pollution patterns. Future enhancements propose developing a live dashboard for real-time AQI predictions by integrating spark streaming that we developed with kafka event streams.

Successfully built ML model to:

- Predict the AQI for a given city given the pollutant features.
- Evaluate the performance metrics of different regression models and select the best model for deployment.
- Fine tune the best regression model through hyperparameter tuning.
- Forecasted AQI values for an year using LSTM.
- Utilized spark's structured streaming to process real-time(simulated) inputs to demonstrate live AQI prediction.

Scalability:

- Entire project was run on colab notebook and on single system. We did not explore distributed training due to lack of execution environment.
- Spark ML models used in the project support distributed learning using spark dataframes.
- LSTM from PyTorch has inbuilt support for distributed training. Developed code is modular for quick porting.
- For stream Data processing we used Spark Streaming that supports distributed streaming for high volume and velocity of events.
- The ML pipelines are scalable to support data for more cities and the models are easily retrainable.

Performance:

- For single models achieved validation accuracy > 80% and test accuracy > 75%.
- Improved Best Models accuracy using RGSCV to achieve > 80% test accuracy.
- Demonstrated instantaneous live prediction using best model and spark structured streaming.
- Forecasted AQI curve showed reasonable parity with the actual.
- Training and Validation time for single ML models were under 2 minutes.
- Training and Validation time for RGSCV model was under 20 minutes.

Future extensions:

The document envisions significant future enhancements, including the development of a live dashboard to monitor and predict air quality in real time. This dashboard would integrate historical data with live feeds from air quality monitoring stations, offering dynamic visualization of pollutant levels and AQI trends across Indian cities. It aims to provide actionable insights through features like:

1. Real-Time AQI Display: A constantly updated AQI index for each city, allowing users to view current pollution levels and compare them across locations.
2. Predictive Analytics: Provide AQI forecasts for selected duration.
3. Kafka Integration: Integrate developed streaming process with kafka event streams.