भारतीय विज्ञान संस्थान

DA 231o: Data Engineering at Scale
Course Project Final Presentation

Forecasting India's Air

- AQI Patterns and Real-Time Pollution Alerts

Team DataWatch

Ashutosh Mishra, ashutoshmish@iisc.ac.in

Neeraj K Shete, neerajshete@iisc.ac.in

Tony P Joy, tonyjoy@iisc.ac.in

Vivek H N, vivekhn@iisc.ac.in

Workflow

1. Problem Definition
2. Data Preparation and Exploration
3. Data Modeling
4. Data Streaming and Live Prediction
5. Reflections
6. Visualizing Results

Problem Definition

Objective	Build ML models that can Predict and Forecast the AQI value based on the measured air pollutant parameters.
Deployment Model	The models can be deployed as a standalone dashboard or can be integrated with existing weather forecasting systems.
Data Science Category	This is a classical supervised regression problem. Can be used with both offline and online data input.
KPI Benchmarks (Best Model)	Validation Accuracy KPI: 84 % Test Accuracy KPI: 83.88 %
Stretch Goal	Prediction on live data using Kafka and Spark Streaming. Timeseries Forecasting of AQI values.
Manual Approach	Mean of historical data for similar hour and city could be used as a prediction measure.
Assumptions	Model runtime is not a huge concern for hourly streams. Data taken for training is accurate. There will be authorized methods to subscribe for online data.

Data Collection and Preparation

Data Collection

Government Data – ("Central Control Room for Air Quality Management - All India")

Link: CCR (cpcb.gov.in)

Open and Free to use.

Well Maintained along with sufficient historical data.

Data Retrieval and Storage

Shortlisted 6 cities to focus on – Bengaluru, Hyderabad, Chennai, Mumbai, Kolkata and Delhi.

Collecting yearly dataset for each of these cities for the years 2019 through 2023 (hourly data).

~8k samples for a city per year.

Dedicated Train [2019 - 22] and Test [2023] Datasets.

Combined datasets to form train and test master datasets of all cities.

Stored in shared google drive.

Data Cleaning

Set threshold of 10% and delete columns based on missing value threshold

Convert Timestamp column to datetime format.

Feature Engineering

Rolling average computation using sliding window

Sub Index Calculation

AQI and AQI Class calculation

Vehicular and Industrial pollution calculation.

Data Exploration

Study of major pollutants per city

- Distribution of pollutants
- Monthly Averages
- Seasonal Trends

Study impact of pollutants on AQI

- Correlational analysis of major pollutants on AQI value.
- Heatmap of Feature Correlation.

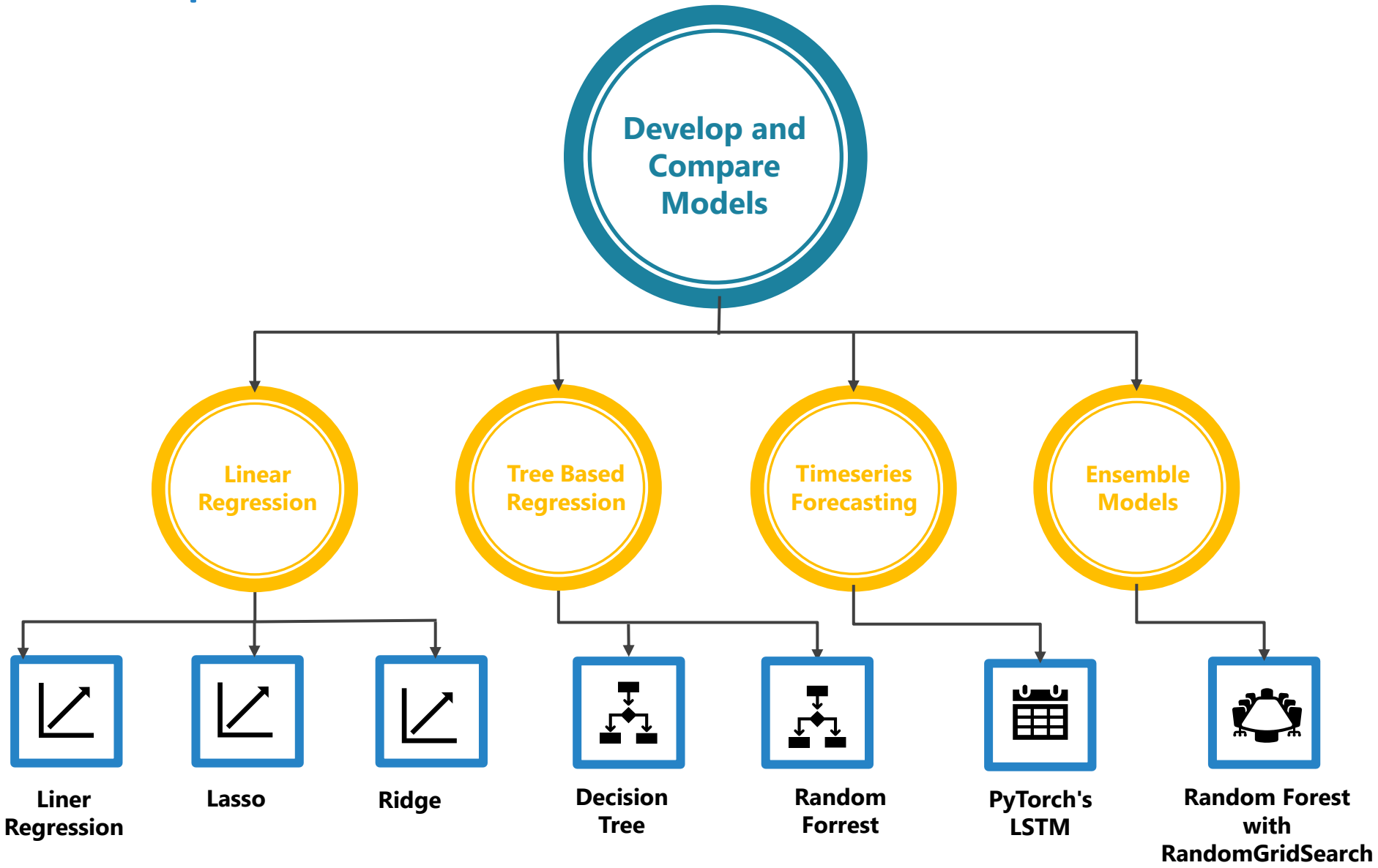
Study of City wise AQI Values

- Monthly AQI Distribution
- AQI Class Distribution
- Change over the years – Impact of lockdown

Study of Vehicular and Industrial Pollution

- City wise Average
- Monthly Trends

Model Development



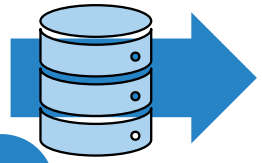
Data Streaming and Live Prediction

Producer

- Generate hourly pollutant data *
- Generate consolidated CSV for all cities.
- Write consolidated hourly CSV to Input Store

* Hourly Samples from 2023 Dataset

Shared Input Store

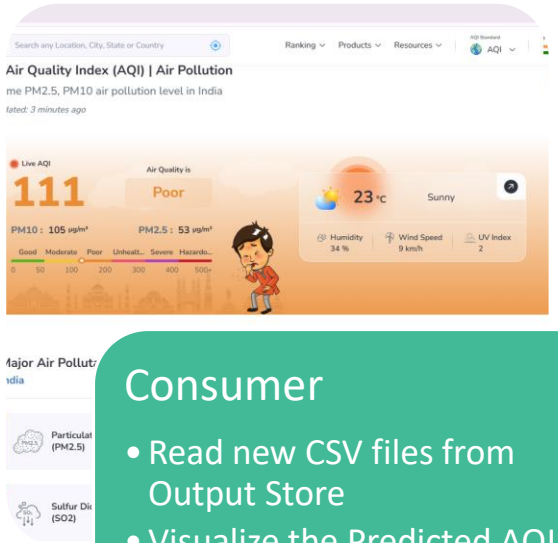


APACHE Spark™

Spark Streaming

- Read new CSV files from Input Store
- Convert to Dataframe.
- Predict AQI with saved model.
- Generate Output Dataframe.
- Write Output Dataframe as CSV to Output Store.

Shared Output Store



Consumer

- Read new CSV files from Output Store
- Visualize the Predicted AQI for each city *

* New bar plots

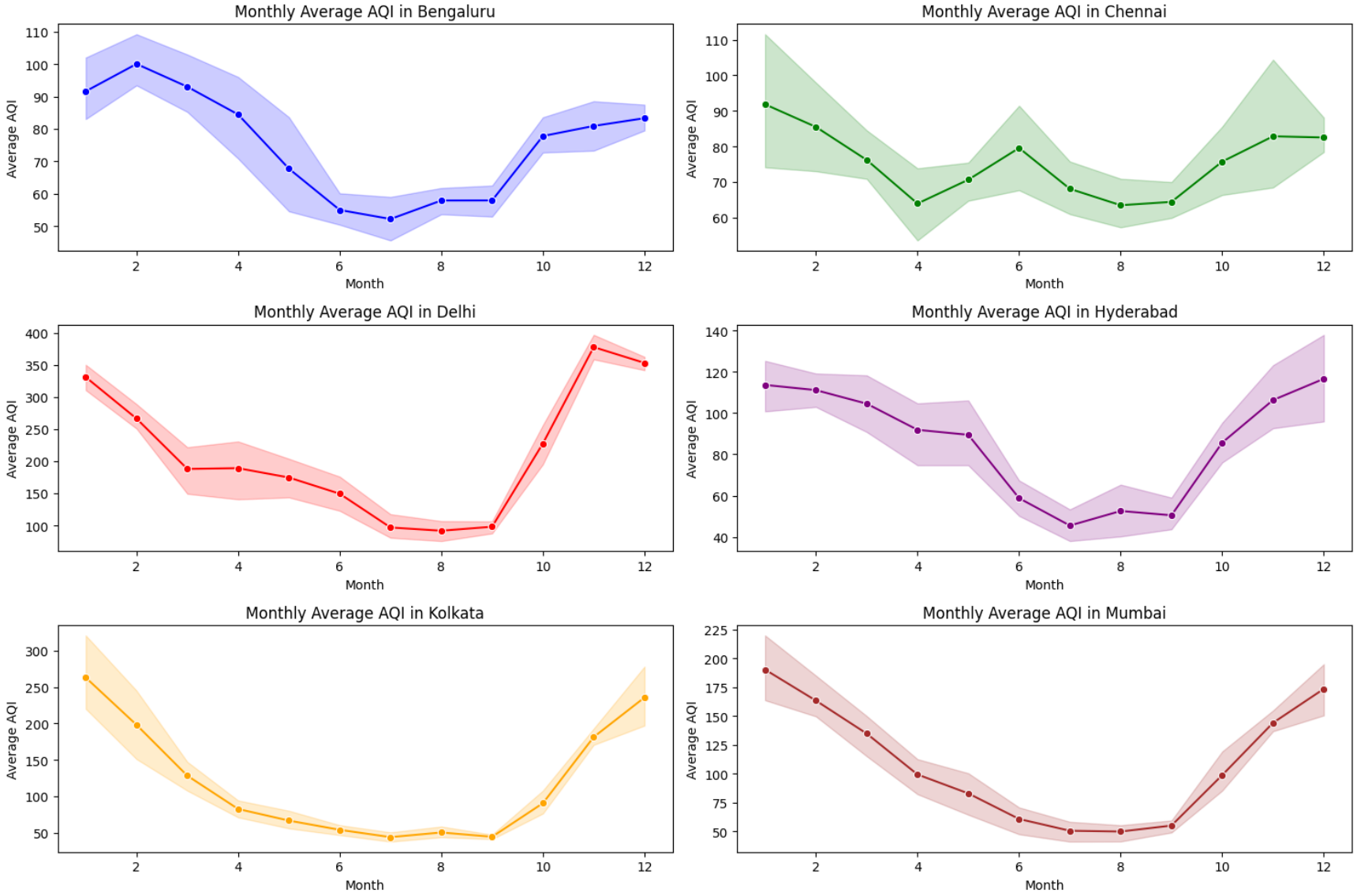
Reflection and Inference : Data Preparation and EDA

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

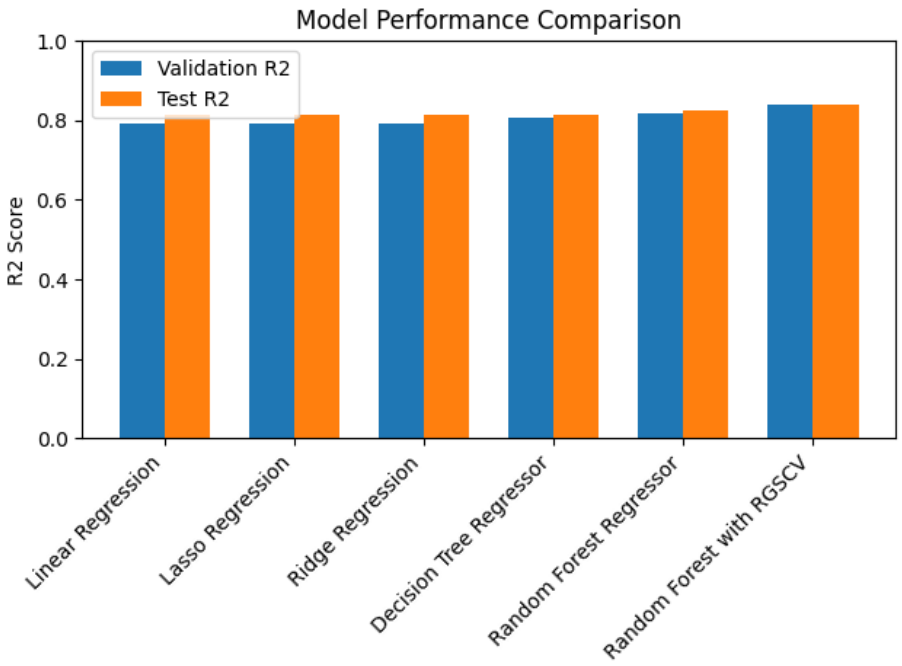
Reflection and Inference : Data Modelling

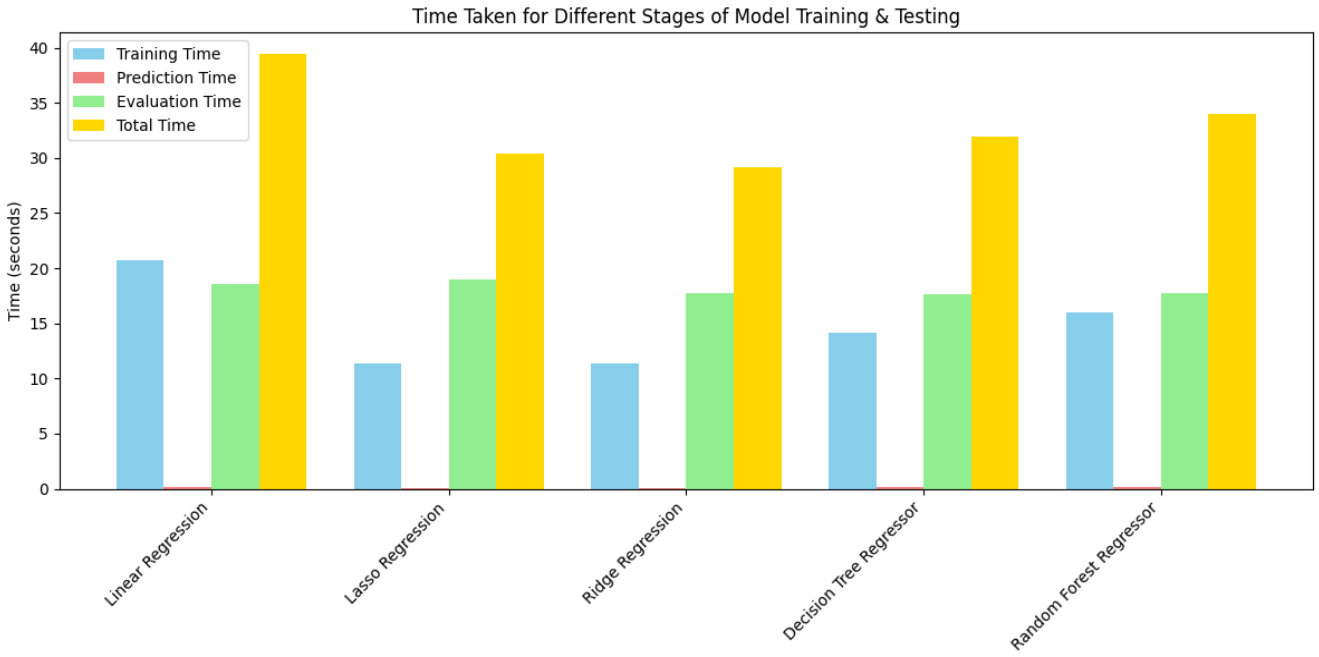
Models	RMSE		Testing Accuracy		Validation Accuracy	
	Achieved	Target	Achieved(%)	Target(%)	Achieved(%)	Target(%)
Linear Regression	39.5	35	79.2	75	81.3	75
Lasso Regression	39.4	35	79.2	75	81.3	75
Ridge Regression	39.5	35	79.2	75	81.2	75
Decision Tree Regressor	37.9	35	80.8	80	81.46	80
Random Forest	37.1	35	81.67	80	82.64	80
Random Forest with Randomized search	34.8	35	83.8	80	84.2	80

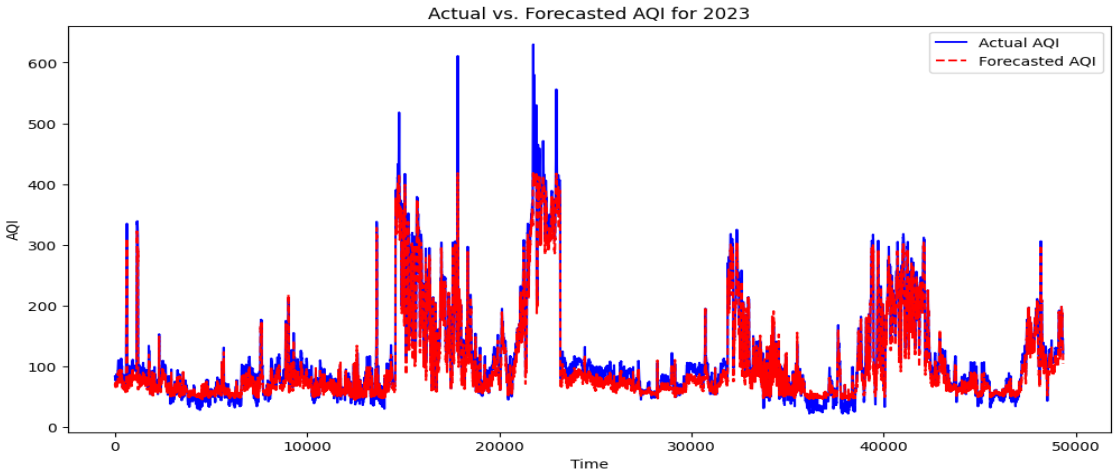
Visualizing Results

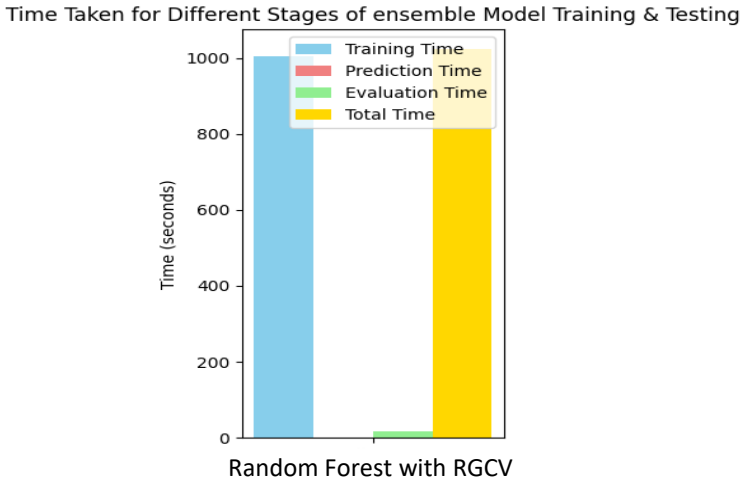


Visualizing Results

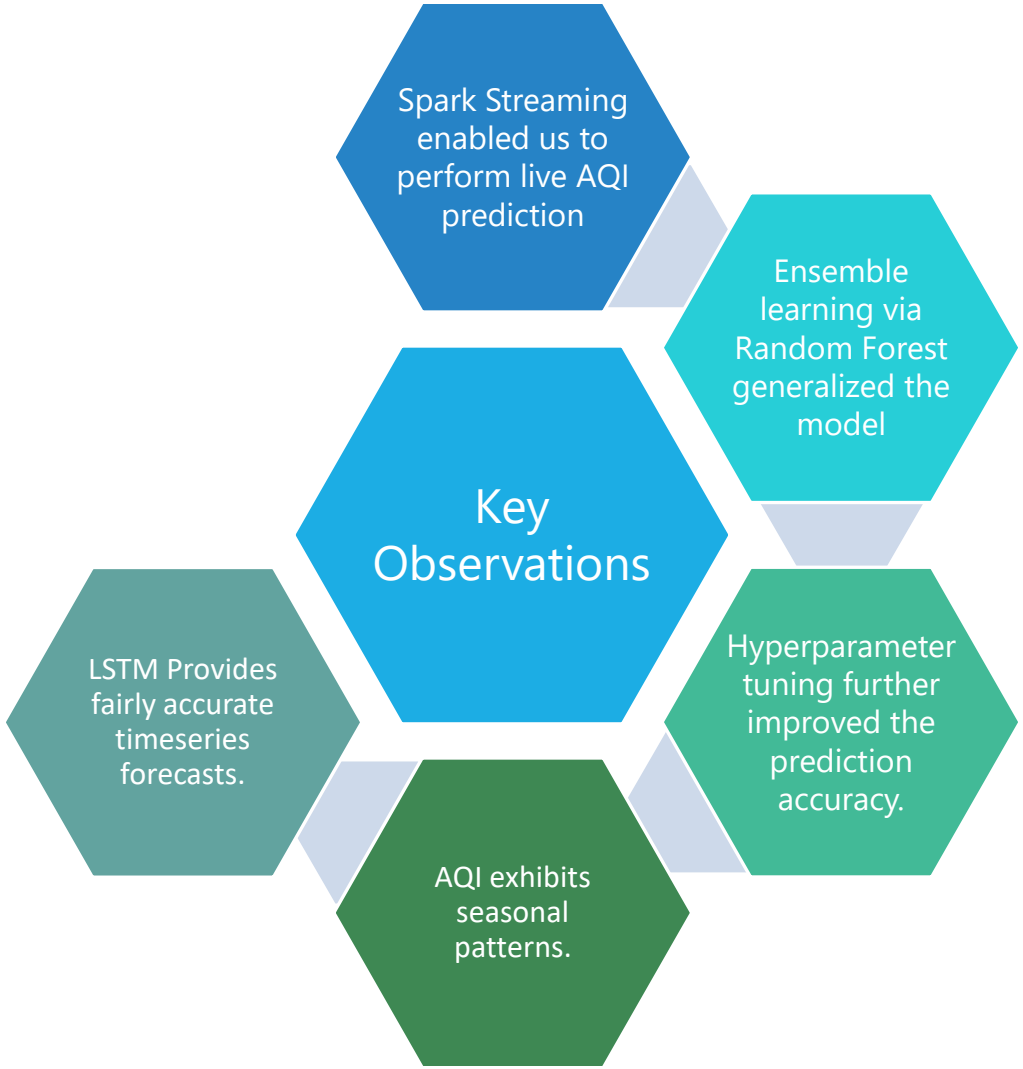








Key observations



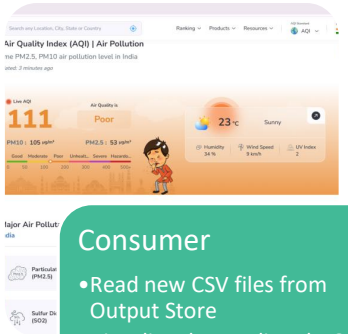
Deployment for Demo

Producer

- Generate hourly pollutant data *
- Generate consolidated CSV for all cities.
- Write consolidated hourly CSV to Input Store

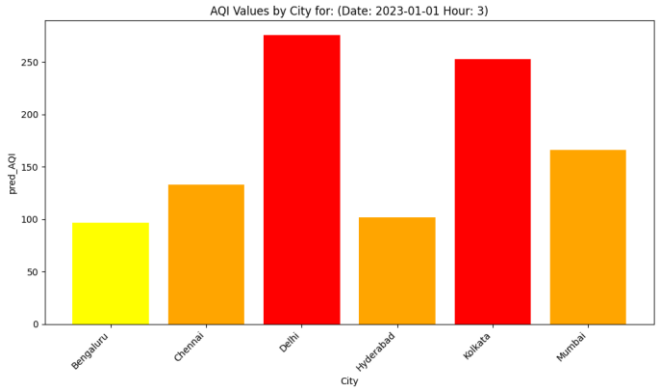
Spark Streaming

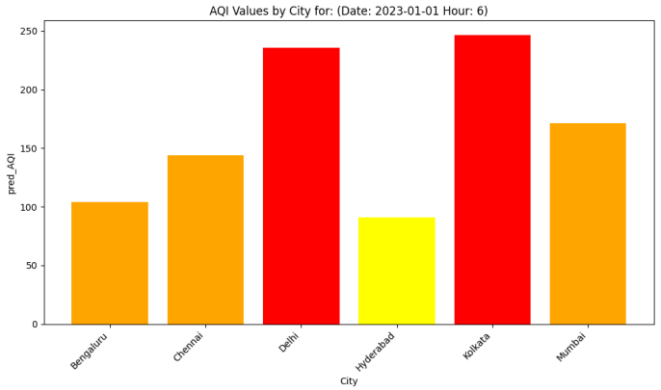
- Read new CSV files from Input Store
- Convert to Dataframe.
- Predict AQI with saved model.
- Generate Output Dataframe.
- Write Output Dataframe as CSV to Output Store.

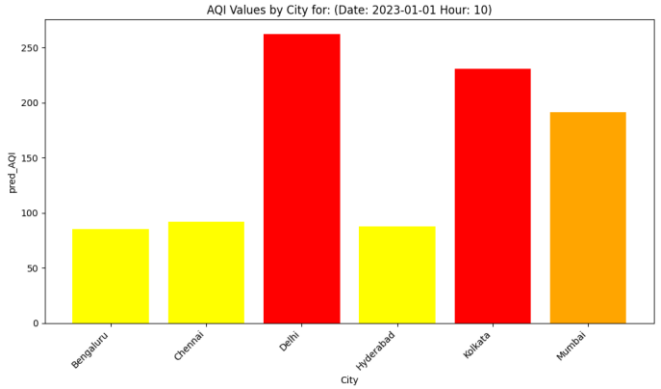


Consumer

- Read new CSV files from Output Store
- Visualize the Predicted AQI for each city *



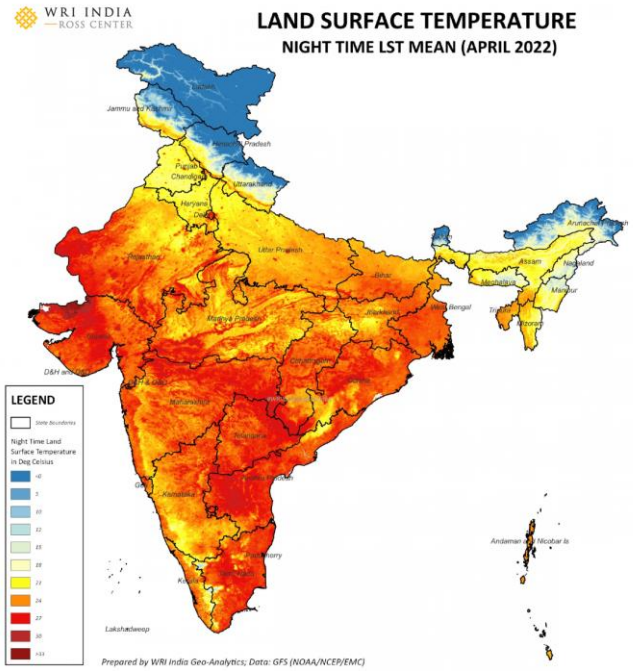


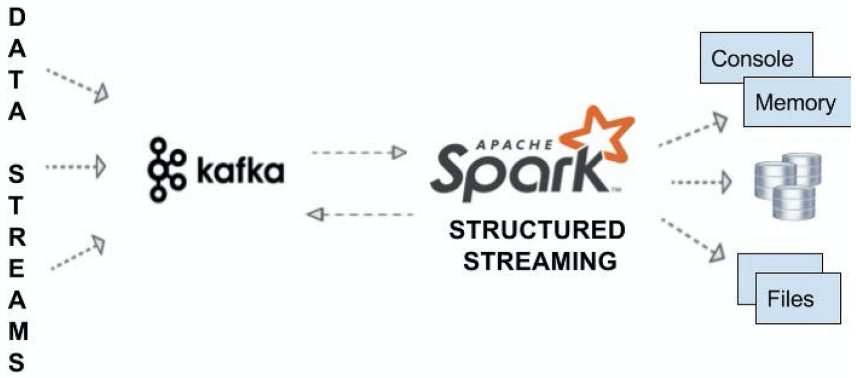


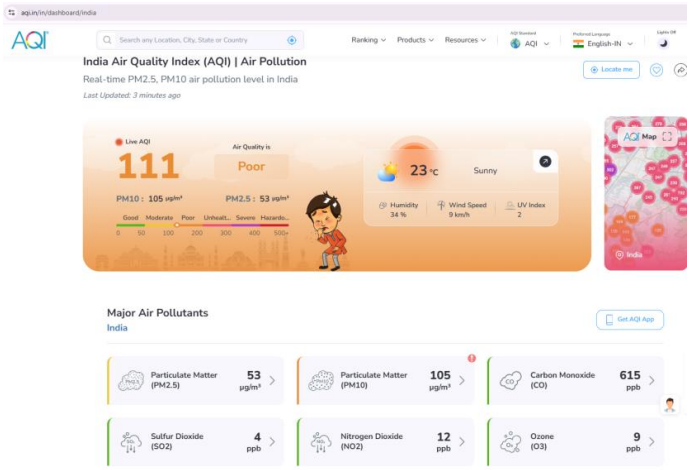
Enhancements/Future work

City wise MODELS

Continuous Training and Integration







LIVE DASHBOARD

Forecasts, Prediction and Live Measurements

Thank You !!