Progressive Education Society's
MODERN COLLEGE OF ENGINEERING
Shivajinagar, Pune -05.
(An Autonomous Institute Affiliated to Savitribai Phule Pune University)
MCA Department

SYMCA Research Project Report Logbook 2025-26 TERM I

	SR. NO	TASK	DATE	COMPLETION DATE	SIGN WITH REMARK
Project Registration	1	Student Group allocation	**01 Aug 2025**		
	2	Project Allocation			
	3	Project Abstract Upload			
STAGE-1	**SR. NO**	**Progress Report I**	**DATE**		
	1	Aim, Objective and Scope of Project	**09/ Aug 2025**		
	2	H/W S/W Requirement, Tools, Human Efforts in Hours			
	3	System Overview, Proposed System and Expected outcome, Architecture & initial phase of design,			
	4	Title of Research Paper, Literature Review, Research Gap			
	5	Objective of Research Paper, Abstract of Research Paper			
STAGE-2	**SR. NO**	**Progress Report 2**	**DATE**		
	6	ERD & UML Diagram(Activity & State Transistion (As Per Problem statement)	**23 Aug 2025**		
	7	Database Design with proper key definition			
	8	Data Dictionary, Flow chart			
	9	Project Plan			
	10	Synopsis of Research Project and Paper			
REVIEW -1		**REVIEW - I**	**30 Aug 2025**		

STAGE-3	**SR. NO**	**Progress Report III**	**DATE**		
	11	Requirement Analysis / Models	**13 Sep 2025**		
	12	Screen Design with proper validations, Reports-Analytical & graphical			
	13	3-4 Working Models, Publish Research Paper			
	14	30 – 40% Coding documentation			
STAGE-4	**SR. NO**	**Progress Report IV**	**DATE**		
	15	Test Data (screenshot)	**27 Oct 2025**		
	16	User Manual			
	17	Bibliography, References			
	18	Published copy of Research Paper & Certificate			
	19	Spiral Copy Submission & Project Execution			
REVIEW -2		**REVIEW - II**	**11 Oct 2025**		
REVIEW -3		**PANEL DEMO PRESENTATION**	**1 Nov. 2025**		
		Final OR	**1st Week of Nov 2025**		

Coordinator
Dr. Rama Bansode

HOD MCA (I/C)
Dr. Prakash Kene

Progressive Education Society's
MODERN COLLEGE OF ENGINEERING
Shivajinagar, Pune -05.
(An Autonomous Institute Affiliated to Savitribai Phule Pune University)
MCA Department

RESEARCH PROJECT (MCA-II) Abstract (2025-26) TERM I

Name Of Student	Gurjas Singh Gandhi
Roll No. Div	52120/B
Project Title	Air Quality Monitoring and Visualizer App
Technology used	**Flutter,Node.Js,TensorflowLite,Google Maps API ,Supabase,CPCB AQI Data,ISRO Vedas Satellite Based AQI Data.**
Group No.	3
Group Members with Roll No.	Nikita Bachute 52102 Pranav Gadewar 52119 Ritwik Rahut 52152
Keywords	Air Quality Index (AQI), real-time monitoring, forecasting, satellite data, ground sensors, hyperlocal pollution, health alerts, predictive analytics, environmental data visualization
Name of Guide/Mentor	Dr. Prakash Kene
Student Email ID	gurjas_gandhi_mca@moderncoe.edu.in
Student Contact Number	9922233168

Progressive Education Society's
MODERN COLLEGE OF ENGINEERING
Shivajinagar, Pune -05.
(An Autonomous Institute Affiliated to Savitribai Phule Pune University)
MCA Department

Vayu Drishti- An Air Quality Monitoring and Visualization App

Keywords:
Air Quality Index (AQI), real-time monitoring, forecasting, satellite data, ground sensors, hyperlocal pollution, health alerts, predictive analytics, environmental data visualization

Technology Used:
Streamlit & React js (mobile/web UI), Node.js + Express (backend API), Python (web scraping), TensorFlow Lite (forecasting models), Supabase (PostgreSQL database), Folium Maps and Leaflet.js(mapping/visualization)

Modules:

- Data Ingestion & Scraper (Python)
- Backend API & Data Processing (Node.js)
- Forecasting Engine (TensorFlow Lite)
- Frontend UI & Visualization (Streamlit & React js)
- Database (Supabase PostgreSQL)

Introduction:

Air pollution is a growing concern worldwide due to its impact on health and the environment. We recognize the need to provide accurate, real-time air quality information to all communities, including small towns and rural areas where monitoring is limited. We will develop an app that visualizes air quality data and provides forecasts to help users stay informed and safe.

We will collect data from multiple sources like government sensors, satellite observations, and public platforms to offer a complete, hyperlocal view of pollution. Our system will display pollutant levels such as PM2.5, NO2, and O3 in easy-to-understand maps and charts. This transparency helps users understand air quality changes over time.

To predict future conditions, we will apply machine learning models that forecast air quality up to 72 hours ahead using weather and historical data. These forecasts will enable early warnings for communities, assisting in health protection and planning. The app will also send push notifications with personalized health advice.

Our platform uses modern technology like Streamlit & React js for web, Node.js for backend services, and Python for data scraping to ensure reliability and scalability. This system is designed for easy integration and expansion, allowing developers and policymakers to use the data effectively.

In summary, we will build a powerful, accessible air quality monitoring system combining multiple data streams and AI forecasts. This will empower individuals and authorities with actionable insights into a healthier environment and better quality of life.

Progressive Education Society's
MODERN COLLEGE OF ENGINEERING
Shivajinagar, Pune -05.
(An Autonomous Institute Affiliated to Savitribai Phule Pune University)
MCA Department

PROGRESS REPORT NO. 1

1. **Name of the Student:** Gurjas Singh Gandhi
2. **Class and Div:** SY MCA / B
3. **Date of Submission:** 09/09/2025
4. **Name of the Subject:** Research Project
5. **Name of the Guide:** Dr. Prakash Kene
6. **Team Members: Roll no., Student Name**:

Nikita Bachute	52102
Pranav Gadewar	52119
Ritwik Rahut	52152

7. **Title of the Research work :** " Air Quality Monitoring and Visualization"
8. **Details of the work done:** Attach Separate Sheet |(On Page No.2)
9. **Date of Progress report submission**: 09/09/2025

Details of the Work Done:

1. Aim, Objective and Scope

Aim:
To Develop an accessible, real-time air quality visualization and forecasting system for underserved regions in India.

Objectives:

- Aggregate and display real-time/historical AQI data from ground and satellite sources.
- Visualize air quality trends, provide 24–72-hour forecasts, and deliver health advisories/alerts.
- Enable users to access hyperlocal air pollution insights via mobile/web app.

Scope:
It covers data integration, visualization, and automated health notifications for air quality—focusing on small towns and rural areas within India, using open data and modular technology.

2. Requirements, Tools, Human Effort

Requirements

- Real-time & historical AQI data collection: Integrating sources like CPCB, ISRO VEDAS, or AQICN APIs or scrapers.
- Mobile/web user interface: Visualizing AQI data, trends, heatmaps, and forecasts.
- Forecasting module: Predicting 24–72-hour air quality using meteorology and AQI history.
- Push notifications: Sending health alerts when pollution spikes are detected.
- Health advisories dashboard: Displaying actionable health advice based on AQI level.
- Backend & database: Efficient data storage, management, and secure API delivery.

Tools & Technologies

- Frontend: Streamlit and React (web UI)
- Backend: Node.js + Express, Supabase (PostgreSQL DB), Firebase Cloud Messaging
- Scraper: Python (requests, BeautifulSoup, Pandas for scraper and data processor)
- Visualization: Leaflet and Folium
- Forecasting: TensorFlowLite, Hugging Face API for ML models
- Testing/Analytics: Jest, Firebase/Google Analytics

Human Effort (approx):

- 4 students, each 15–20 hours
- Total ~70 hours (design, coding, testing, docs)

3. System Overview & Outcome

System Overview

The system collects real-time air quality data from multiple sources (CPCB, ISRO VEDAS, AQICN) via APIs and scrapers, stores and processes the data in a backend, and delivers it to a mobile/web app. It visualizes current and historical AQI, forecasts pollution levels, and sends health alerts.

Proposed System

- Data Collection: Automated web scraping and API integration to ingest air pollution data.
- Backend: Node.js server with database for real-time data management, forecasting ML model integration, and push notification service.
- Frontend: Stream lit Web app and React JS for providing interactive maps, trend graphs, and health advisory dashboards.
- Forecasting: ML models (TensorFlowLite or Hugging Face) predict 24–72-hour AQI trends using weather and historical data.

Outcome

- Real-time, hyperlocal air quality visualization across urban and rural regions.
- Predictive pollution forecasts enabling proactive health protection.
- Personalized health recommendations and push alerts for pollution spikes.
- Open, extendable platform supporting integration with external systems and policymaking.

4. Research Paper Details

Title:

An Integrated Real-Time Air Quality Visualization and Forecasting System for Rural and Urban Areas in India

Literature Review:

We find that air quality monitoring systems often use IoT sensors, cloud platforms, and machine learning for data collection and forecasting. Studies highlight integrating gas sensors, wireless modules, and satellite data for broader coverage. Real-time visualization and remote alert systems are common, but most focus on urban areas without full integration or personalized health advisories.

Research Gap:

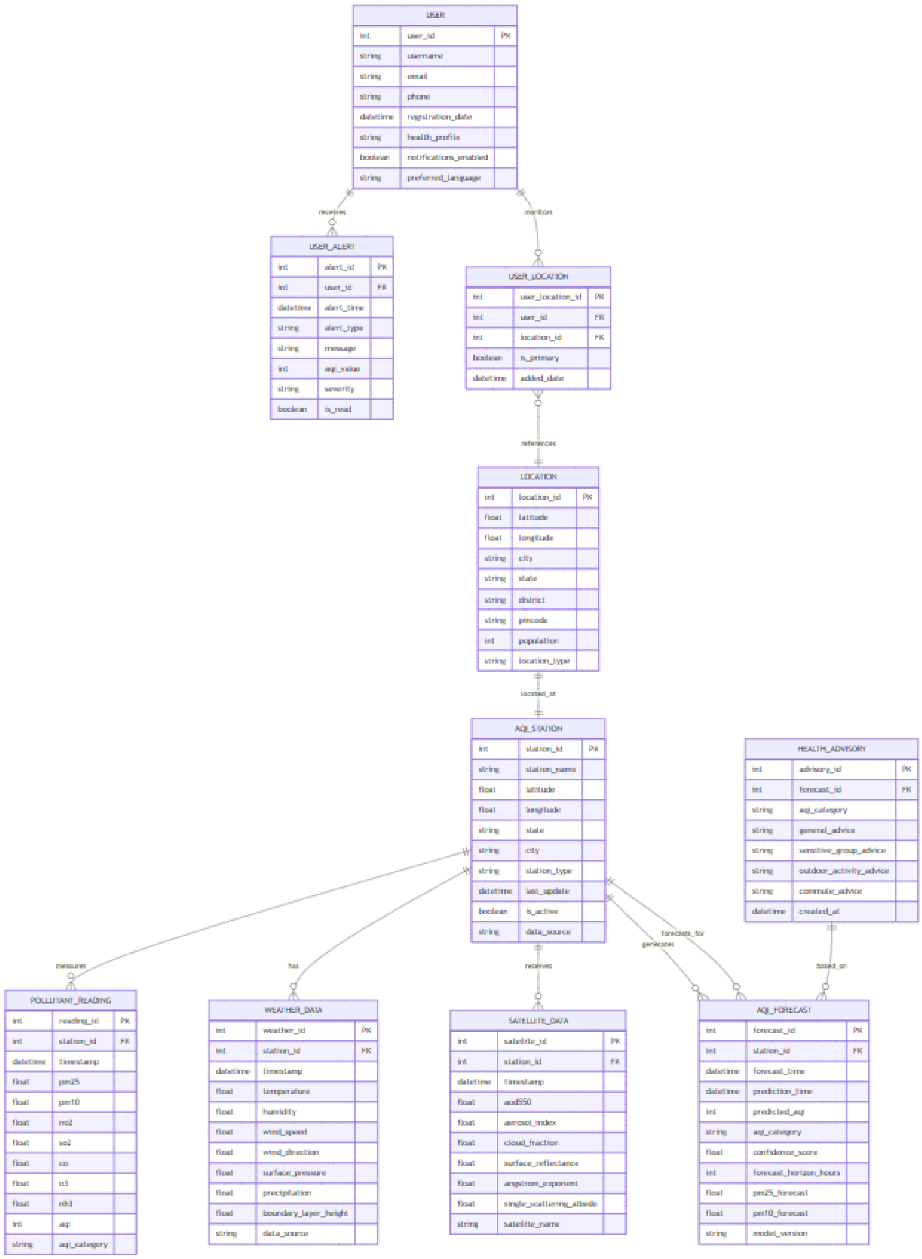
We observe a gap in coverage for rural and small-town regions. Existing systems rarely combine ground sensors, satellite data, and public sources into a unified real-time platform. We will address this by creating an open, scalable system with AI-driven forecasting and health alerts accessible to underserved communities.

5. Research Paper Objective & Abstract

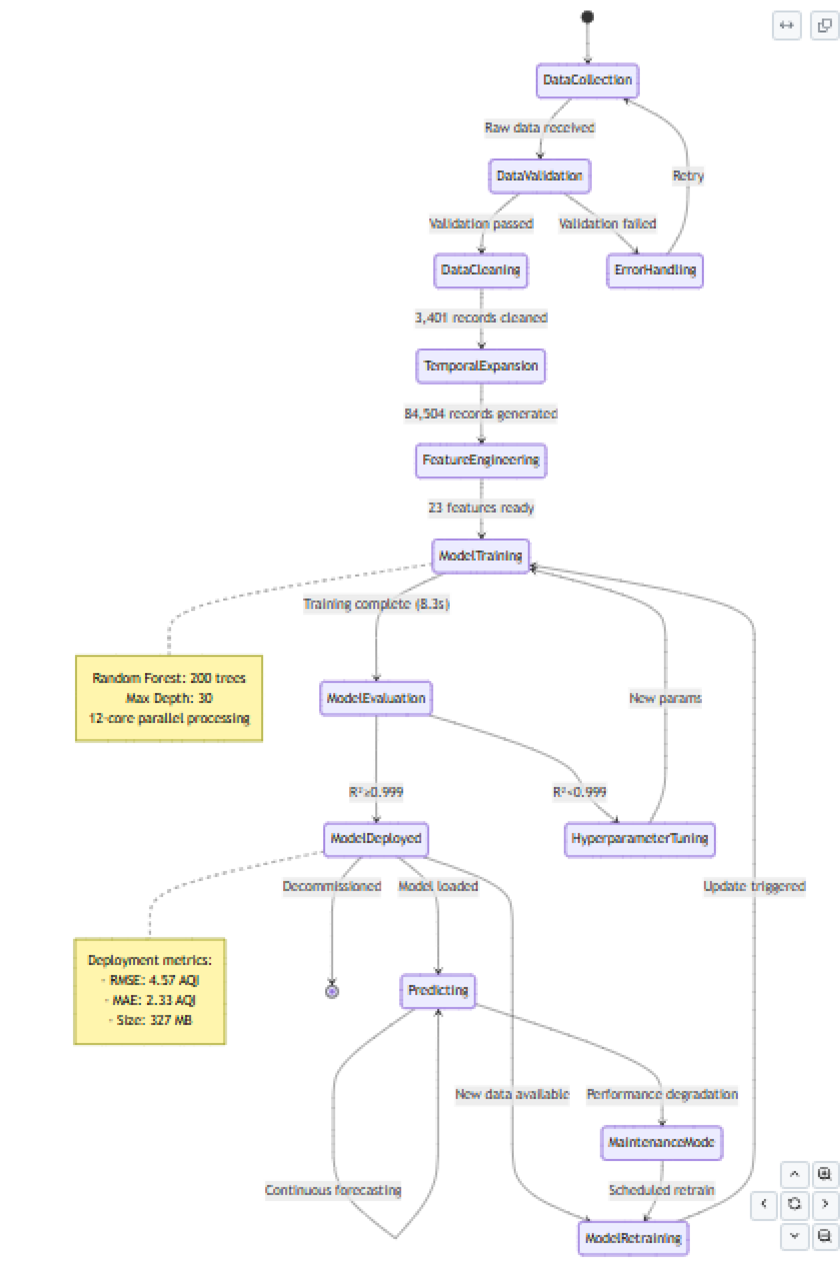
Objective:
We aim to develop a real-time air quality monitoring and forecasting system that integrates data from ground sensors, satellite platforms, and public sources. We will provide hyperlocal pollution visualization, short-term forecasts, and personalized health alerts to empower communities and policymakers, especially in underserved areas.

Abstract:
Air pollution poses significant health risks, particularly in regions with limited monitoring infrastructure. We will build an integrated system combining data from government sensors, ISRO VEDAS satellites, and public platforms to offer real-time air quality visualization and forecasting. Using machine learning, we will predict pollution levels for the next 24–72 hours and send personalized health advisories via a mobile/web app. This system addresses current limitations by providing comprehensive, timely, and accessible air quality information, promoting healthier environments and informed decision-making.

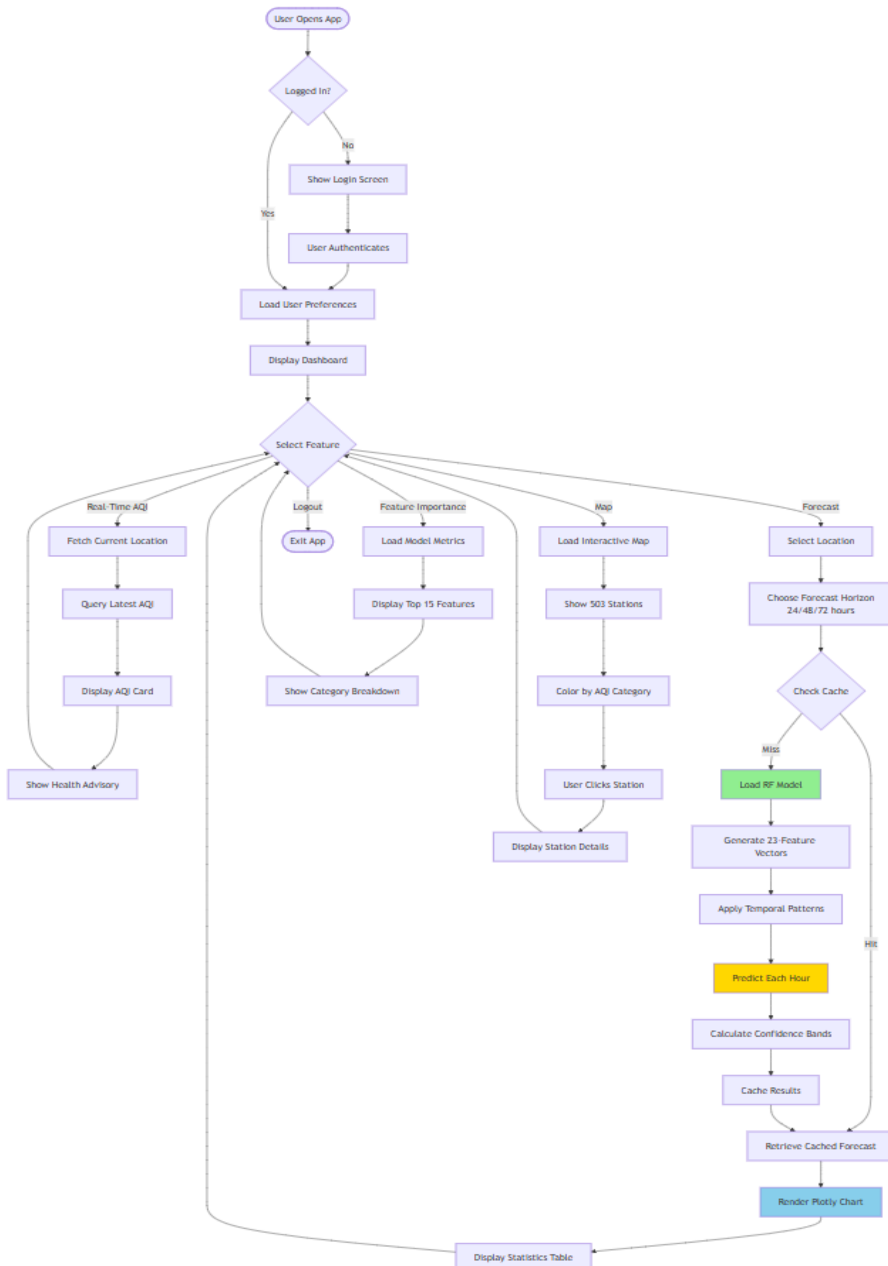
6. ERD , UML & ER Diagram (As Per Problem statement)



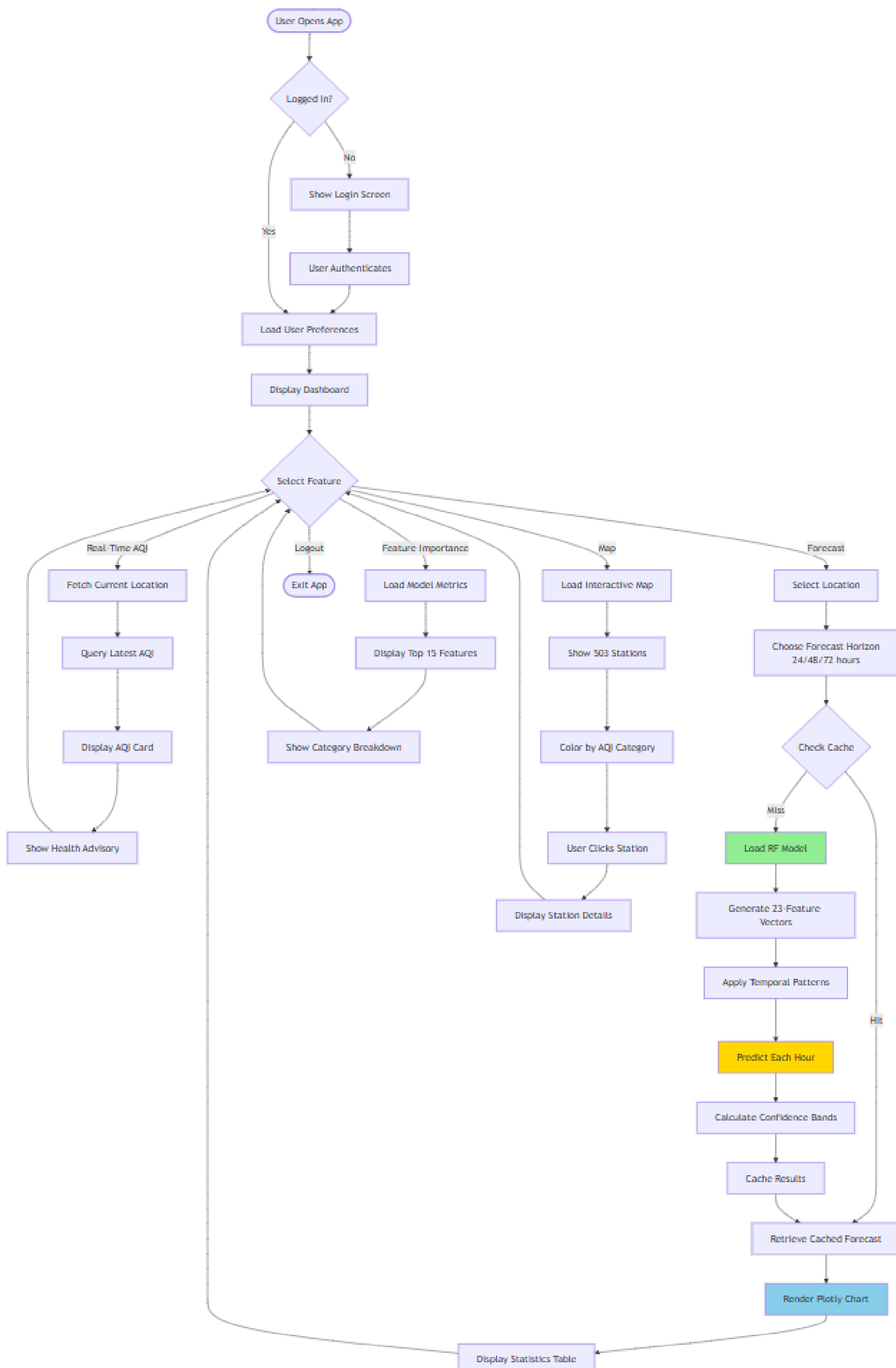
6. State Diagram - Model Training Lifecycle



Activity Diagram

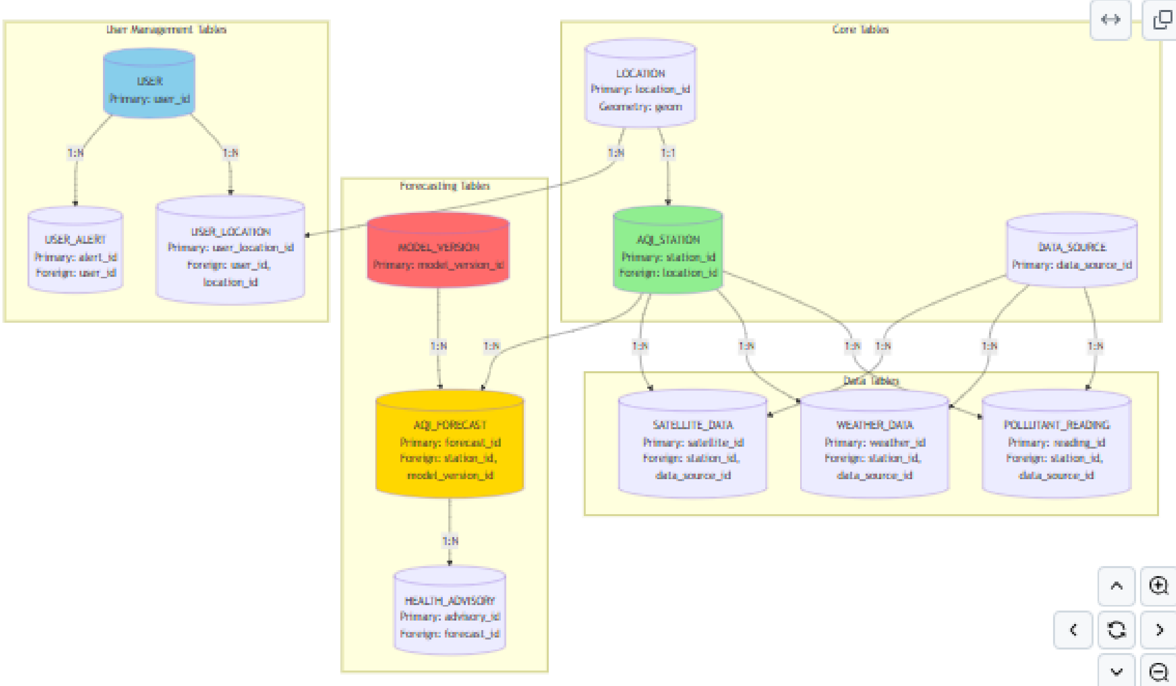

User Opens App
Logged In?
No
Yes
Show Login Screen
User Authenticates
Load User Preferences
Display Dashboard
Select Feature
Real-Time AQI
Logout
Feature Importance
Map
Forecast
Fetch Current Location
Exit App
Load Model Metrics
Load Interactive Map
Select Location
Query Latest AQI
Display Top 15 Features
Show 503 Stations
Choose Forecast Horizon 24/48/72 hours
Display AQI Card
Show Category Breakdown
Color by AQI Category
Check Cache
Miss
Hit
Show Health Advisory
User Clicks Station
Load RF Model
Display Station Details
Generate 23-Feature Vectors
Apply Temporal Patterns
Predict Each Hour
Calculate Confidence Bands
Cache Results
Retrieve Cached Forecast
Render Plotly Chart
Display Statistics Table

7. Database Design with proper key definition



Progressive Education Society's
MODERN COLLEGE OF ENGINEERING
Shivajinagar, Pune -05.
(An Autonomous Institute Affiliated to Savitribai Phule Pune University)
MCA Department

Database Schema - Relational Model



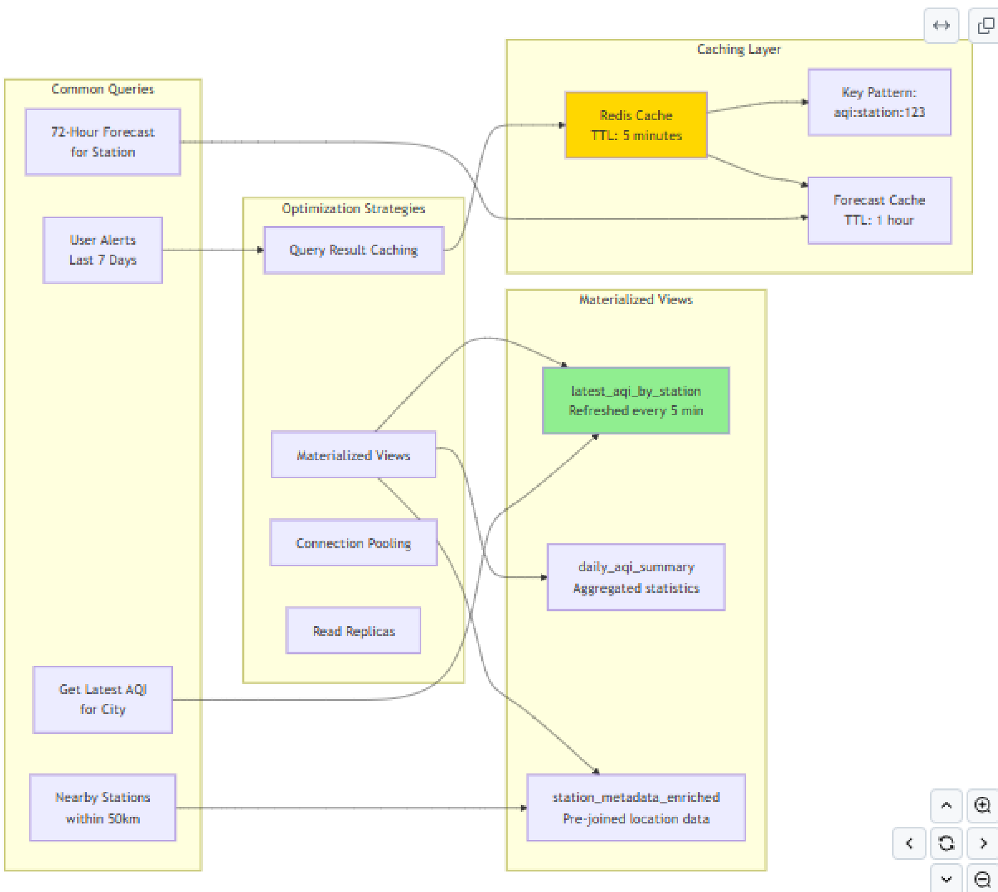
Database Statistics

Component	Count/Size	Description
Total Tables	13	Core + Data + User tables
Partitioned Tables	2	pollutant_reading, aqi_forecast
Materialized Views	1+	latest_aqi_by_station, daily_summary
Indexes	25+	Primary, Foreign, Composite, Spatial
Triggers	3+	Auto-updates, alerting, auditing
Database Size (Est.)	50-100 GB	With 6 months historical data
Daily Growth	~500 MB	503 stations × 24 hours × 3 tables
Partitions	Quarterly	3-month rotation for time-series
Backup Size	~200 GB	Full + incremental backups
Read Replicas	2	For load balancing

Performance Benchmarks

Query Type	Expected Time	Optimization
Latest AQI (single station)	<10ms	Materialized view
72-hour forecast	<50ms	Cached in Redis
Nearby stations (50km)	<100ms	Spatial index
Historical data (6 months)	<500ms	Partition pruning
Aggregate statistics	<200ms	Materialized view
User alerts (unread)	<20ms	Composite index
Map markers (all stations)	<150ms	Cached + indexed

8. Query Performance Optimization



8. Data Dictionary & Flow Chart

Core Entity Tables

1. AQI_STATION

Column Name	Data Type	Constraints	Description	Example Value
station_id	INT	PRIMARY KEY, AUTO_INCREMENT	Unique identifier for monitoring station	1001
station_name	VARCHAR(200)	NOT NULL	Official name of the monitoring station	"Delhi - ITO"
latitude	DECIMAL(10,7)	NOT NULL, CHECK (-90 to 90)	Geographic latitude coordinate	28.6273928
longitude	DECIMAL(10,7)	NOT NULL, CHECK (-180 to 180)	Geographic longitude coordinate	77.2403256
state	VARCHAR(50)		State/Province name	"Delhi"
city	VARCHAR(100)		City name	"New Delhi"
district	VARCHAR(50)		District name	"Central Delhi"
station_type	VARCHAR(50)	CHECK (CPCB/Manual/Satellite/Hybrid)	Type of monitoring station	"CPCB"
last_update	TIMESTAMP		Last data received timestamp	2025-10-30 14:30:00
is_active	BOOLEAN	DEFAULT TRUE	Whether station is operational	TRUE
data_source	VARCHAR(50)		Primary data source	"CPCB_API"
location_id	INT	FOREIGN KEY → location(location_id)	Reference to location details	501
created_at	TIMESTAMP	DEFAULT CURRENT_TIMESTAMP	Record creation timestamp	2024-01-15 08:00:00
updated_at	TIMESTAMP	DEFAULT CURRENT_TIMESTAMP	Last record update timestamp	2025-10-30 14:30:00

Record Count: 503 stations
Update Frequency: Real-time (every hour)
Primary Index: station_id
Secondary Indexes: location_id, is_active

2. LOCATION

Column Name	Data Type	Constraints	Description	Example Value
location_id	INT	PRIMARY KEY, AUTO_INCREMENT	Unique identifier for location	501
latitude	DECIMAL(10,7)	NOT NULL, CHECK (-90 to 90)	Geographic latitude	28.6273928
longitude	DECIMAL(10,7)	NOT NULL, CHECK (-180 to 180)	Geographic longitude	77.2403256
city	VARCHAR(100)		City name	"New Delhi"
state	VARCHAR(50)		State name	"Delhi"
district	VARCHAR(50)		District name	"Central Delhi"
pincode	VARCHAR(6)		Postal code	"110002"
population	INT		Population in the area	25000000
location_type	VARCHAR(20)	CHECK (Urban/Rural/Peri-urban)	Classification of location	"Urban"
geom	GEOMETRY(Point, 4326)	SPATIAL INDEX	PostGIS spatial column for GIS queries	POINT(77.2403 28.6273)
created_at	TIMESTAMP	DEFAULT CURRENT_TIMESTAMP	Record creation timestamp	2024-01-15 08:00:00
updated_at	TIMESTAMP	DEFAULT CURRENT_TIMESTAMP	Last record update timestamp	2025-10-30 14:30:00

Record Count: 503 locations
Spatial Features: PostGIS enabled for distance calculations
Primary Index: location_id
Spatial Index: GIST on geom column

3. DATA_SOURCE

Column Name	Data Type	Constraints	Description	Example Value
data_source_id	INT	PRIMARY KEY, AUTO_INCREMENT	Unique identifier for data source	1
source_name	VARCHAR(100)	NOT NULL	Name of the data source	"CPCB Real-Time API"
source_type	VARCHAR(50)	CHECK (CPCB/MERRA2/INSAT3DR/Manual)	Type of data source	"CPCB"
api_endpoint	VARCHAR(500)		API URL or data location	"https://api.cpcb.gov.in/..."
last_fetch	TIMESTAMP		Last successful data fetch	2025-10-30 14:00:00
is_active	BOOLEAN	DEFAULT TRUE	Whether source is currently active	TRUE
fetch_frequency_minutes	INT	DEFAULT 60	How often to fetch data (minutes)	60
created_at	TIMESTAMP	DEFAULT CURRENT_TIMESTAMP	Record creation timestamp	2024-01-15 08:00:00

Record Count: 3 data sources (CPCB, MERRA-2, INSAT-3DR)
Update Frequency: Hourly for CPCB, Daily for satellite
Primary Index: data_source_id

Data Collection Tables

4. POLLUTANT_READING

Column Name	Data Type	Constraints	Description	Example Value
reading_id	BIGINT	PRIMARY KEY, AUTO_INCREMENT	Unique identifier for reading	1234567
station_id	INT	FOREIGN KEY → aqi_station(station_id), NOT NULL	Reference to monitoring station	1001
timestamp	TIMESTAMP	NOT NULL, INDEXED	Time of measurement	2025-10-30 14:00:00
pm25	FLOAT	CHECK (>= 0)	PM2.5 concentration (μg/m³)	85.3
pm10	FLOAT	CHECK (>= 0)	PM10 concentration (μg/m³)	142.7
no2	FLOAT	CHECK (>= 0)	Nitrogen dioxide (μg/m³)	45.2
so2	FLOAT	CHECK (>= 0)	Sulfur dioxide (μg/m³)	12.8
co	FLOAT	CHECK (>= 0)	Carbon monoxide (mg/m³)	1.2
o3	FLOAT	CHECK (>= 0)	Ozone (μg/m³)	38.5
nh3	FLOAT	CHECK (>= 0)	Ammonia (μg/m³)	22.1
aqi	INT	CHECK (0 to 500), NOT NULL	Air Quality Index	156
aqi_category	VARCHAR(20)	CHECK (Good/Moderate/Unhealthy for Sensitive/Unhealthy/Very Unhealthy/Hazardous)	AQI category	"Unhealthy"
confidence_score	FLOAT	CHECK (0 to 1)	Data quality confidence	0.95
data_source_id	INT	FOREIGN KEY → data_source(data_source_id)	Source of data	1
created_at	TIMESTAMP	DEFAULT CURRENT_TIMESTAMP	Record creation timestamp	2025-10-30 14:05:00

Record Count: 84,504 records (503 stations × 168 hours)
Partitioning: By timestamp (quarterly partitions)
Update Frequency: Hourly
Primary Index: reading_id, timestamp
Composite Index: (station_id, timestamp DESC)

5. WEATHER_DATA

Column Name	Data Type	Constraints	Description	Example Value
weather_id	BIGINT	PRIMARY KEY, AUTO_INCREMENT	Unique identifier for weather record	5678901
station_id	INT	FOREIGN KEY → aqi_station(station_id), NOT NULL	Reference to monitoring station	1001
timestamp	TIMESTAMP	NOT NULL, INDEXED	Time of measurement	2025-10-30 14:00:00
temperature	FLOAT		Air temperature (°C)	28.5
humidity	FLOAT	CHECK (0 to 100)	Relative humidity (%)	65.2
wind_speed	FLOAT	CHECK (>= 0)	Wind speed (m/s)	3.5
wind_direction	FLOAT	CHECK (0 to 360)	Wind direction (degrees)	225.0
surface_pressure	FLOAT		Surface pressure (hPa)	1013.25
precipitation	FLOAT	CHECK (>= 0)	Precipitation (mm)	0.0
boundary_layer_height	FLOAT	CHECK (>= 0)	Planetary boundary layer height (m)	1200.0
atmospheric_pressure	FLOAT		Atmospheric pressure (hPa)	1015.3
data_source_id	INT	FOREIGN KEY → data_source(data_source_id)	Source of data (MERRA-2)	2
created_at	TIMESTAMP	DEFAULT CURRENT_TIMESTAMP	Record creation timestamp	2025-10-30 14:05:00

Record Count: 84,504 records (hourly for each station)
Data Source: MERRA-2 Meteorological Data
Update Frequency: Hourly
Primary Index: weather_id
Composite Index: (station_id, timestamp DESC)

6. SATELLITE_DATA

Column Name	Data Type	Constraints	Description	Example Value
satellite_id	BIGINT	PRIMARY KEY, AUTO_INCREMENT	Unique identifier for satellite record	9876543
station_id	INT	FOREIGN KEY → aqi_station(station_id), NOT NULL	Reference to monitoring station	1001
timestamp	TIMESTAMP	NOT NULL, INDEXED	Time of observation	2025-10-30 10:30:00
aod550	FLOAT	CHECK (>= 0)	Aerosol Optical Depth at 550nm	0.45
aerosol_index	FLOAT		Aerosol Index	1.2
cloud_fraction	FLOAT	CHECK (0 to 1)	Cloud coverage fraction	0.15
surface_reflectance	FLOAT	CHECK (0 to 1)	Surface reflectance	0.25
angstrom_exponent	FLOAT		Ångström exponent	1.3
single_scattering_albedo	FLOAT	CHECK (0 to 1)	Single scattering albedo	0.92
satellite_name	VARCHAR(50)	DEFAULT 'INSAT-3DR'	Satellite source	"INSAT-3DR"
spatial_resolution	FLOAT		Spatial resolution (km)	10.0
data_source_id	INT	FOREIGN KEY → data_source(data_source_id)	Source of data (INSAT-3DR)	3
created_at	TIMESTAMP	DEFAULT CURRENT_TIMESTAMP	Record creation timestamp	2025-10-30 10:35:00

Record Count: 84,504 records (hourly interpolated)
Data Source: INSAT-3DR Satellite
Update Frequency: Daily (interpolated to hourly)
Primary Index: satellite_id
Composite Index: (station_id, timestamp DESC)

Forecasting Tables

7. MODEL_VERSION

Column Name	Data Type	Constraints	Description	Example Value
model_version_id	INT	PRIMARY KEY, AUTO_INCREMENT	Unique identifier for model version	5
model_name	VARCHAR(50)	NOT NULL	Name of ML model	"RandomForestRegressor"
version_number	VARCHAR(20)	NOT NULL	Version string	"v2.0.0"
r2_score	FLOAT		R^2 performance metric	0.9994
rmse	FLOAT		Root Mean Square Error	4.57
mae	FLOAT		Mean Absolute Error	2.33
n_estimators	INT		Number of trees (Random Forest)	200
max_depth	INT		Maximum tree depth	30
trained_at	TIMESTAMP	DEFAULT CURRENT_TIMESTAMP	When model was trained	2025-10-15 10:00:00
training_records	INT		Number of training records	84504
feature_list	TEXT		Comma-separated list of features	"pm25,pm10,no2,latitude,..."
is_active	BOOLEAN	DEFAULT FALSE	Whether model is in production	TRUE
deployed_at	TIMESTAMP		When model was deployed	2025-10-16 08:00:00

Record Count: 5 model versions
Current Active Model: v2.0.0 (R^2=0.9994, RMSE=4.57)
Training Time: 8.3 seconds
Model Size: 327 MB
Primary Index: model_version_id
Unique Index: (model_name, version_number)

8. AQI_FORECAST

Column Name	Data Type	Constraints	Description	Example Value
forecast_id	BIGINT	PRIMARY KEY, AUTO_INCREMENT	Unique identifier for forecast	7654321
station_id	INT	FOREIGN KEY → aqi_station(station_id), NOT NULL	Reference to monitoring station	1001
forecast_time	TIMESTAMP	NOT NULL	When forecast was generated	2025-10-30 14:00:00
prediction_time	TIMESTAMP	NOT NULL, INDEXED	Time being predicted	2025-10-31 14:00:00
predicted_aqi	INT	CHECK (0 to 500), NOT NULL	Predicted AQI value	168
aqi_category	VARCHAR(20)		Predicted AQI category	"Unhealthy"
confidence_score	FLOAT	CHECK (0 to 1)	Model confidence	0.94
forecast_horizon_hours	INT	CHECK (1 to 72)	Hours ahead being predicted	24
pm25_forecast	FLOAT		Predicted PM2.5	92.5
pm10_forecast	FLOAT		Predicted PM10	155.3
no2_forecast	FLOAT		Predicted NO2	48.7
upper_bound	FLOAT		Upper confidence bound (AQI + RMSE)	172.57
lower_bound	FLOAT		Lower confidence bound (AQI - RMSE)	163.43
model_version_id	INT	FOREIGN KEY → model_version(model_version_id)	Model used for prediction	5
created_at	TIMESTAMP	DEFAULT CURRENT_TIMESTAMP	Record creation timestamp	2025-10-30 14:02:00

Record Count: ~36,000 forecasts per day (503 stations × 72 hours)
Partitioning: By prediction_time (monthly partitions)
Update Frequency: Every hour
Forecast Horizons: 1-72 hours
Primary Index: forecast_id, prediction_time
Composite Index: (station_id, prediction_time)

9. HEALTH_ADVISORY

Column Name	Data Type	Constraints	Description	Example Value
advisory_id	INT	PRIMARY KEY, AUTO_INCREMENT	Unique identifier for advisory	12345
forecast_id	BIGINT	FOREIGN KEY → aqi_forecast(forecast_id), NOT NULL	Reference to forecast	7654321
aqi_category	VARCHAR(20)	NOT NULL	AQI category for advisory	"Unhealthy"
general_advice	TEXT		Advice for general population	"Reduce prolonged outdoor exertion"
sensitive_group_advice	TEXT		Advice for sensitive groups	"Avoid prolonged outdoor activities"
outdoor_activity_advice	TEXT		Specific outdoor activity guidance	"Limit to 30 minutes"
commute_advice	TEXT		Commuting recommendations	"Use N95 masks if commuting"
mask_recommendation	TEXT		Mask usage recommendation	"N95 or equivalent recommended"
created_at	TIMESTAMP	DEFAULT CURRENT_TIMESTAMP	Record creation timestamp	2025-10-30 14:02:00
updated_at	TIMESTAMP	DEFAULT CURRENT_TIMESTAMP	Last record update timestamp	2025-10-30 14:02:00

Record Count: ~36,000 advisories per day (one per forecast)
Update Frequency: Generated with each forecast
Primary Index: advisory_id
Foreign Index: forecast_id

User Management Tables

10. USER (app_user)

Column Name	Data Type	Constraints	Description	Example Value
user_id	INT	PRIMARY KEY, AUTO_INCREMENT	Unique identifier for user	10001
username	VARCHAR(50)	UNIQUE, NOT NULL	User's unique username	"air_quality_enthusiast"
email	VARCHAR(100)	UNIQUE, NOT NULL	User's email address	"user@example.com"
phone	VARCHAR(15)		User's phone number	"+919876543210"
password_hash	VARCHAR(255)	NOT NULL	Bcrypt hashed password	"$2b$12$..."
registration_date	TIMESTAMP	DEFAULT CURRENT_TIMESTAMP	When user registered	2025-01-15 10:30:00
health_profile	VARCHAR(20)	CHECK (Normal/Sensitive/High-Risk)	User's health sensitivity	"Sensitive"
notifications_enabled	BOOLEAN	DEFAULT TRUE	Whether user receives alerts	TRUE
preferred_language	VARCHAR(10)	DEFAULT 'en'	Preferred language code	"en"
last_login	TIMESTAMP		Last login timestamp	2025-10-30 09:15:00
is_active	BOOLEAN	DEFAULT TRUE	Account active status	TRUE
created_at	TIMESTAMP	DEFAULT CURRENT_TIMESTAMP	Record creation timestamp	2025-01-15 10:30:00
updated_at	TIMESTAMP	DEFAULT CURRENT_TIMESTAMP	Last record update timestamp	2025-10-30 09:15:00

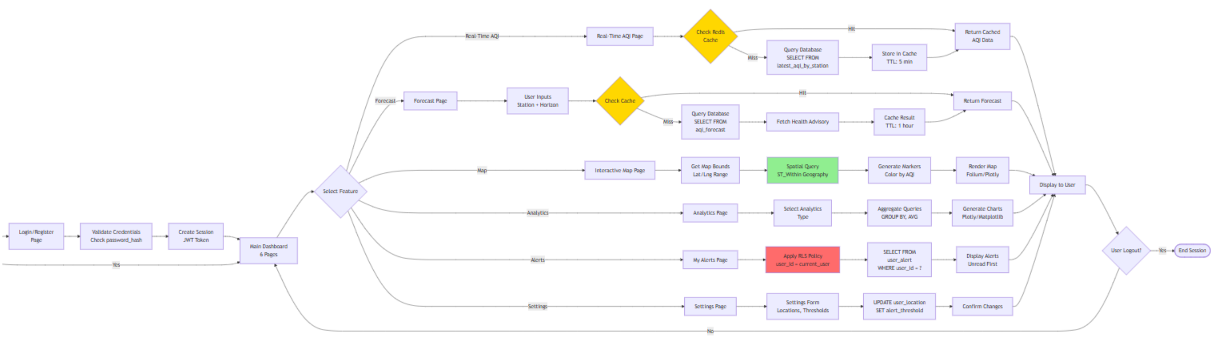
Record Count: Variable (user-dependent)
Security: Row-level security enabled
Password: Bcrypt hashed with salt
Primary Index: user_id
Unique Indexes: username, email

11. USER_LOCATION

Column Name	Data Type	Constraints	Description	Example Value
user_location_id	INT	PRIMARY KEY, AUTO_INCREMENT	Unique identifier for user-location mapping	50001
user_id	INT	FOREIGN KEY → app_user(user_id), NOT NULL	Reference to user	10001
location_id	INT	FOREIGN KEY → location(location_id), NOT NULL	Reference to location	501
is_primary	BOOLEAN	DEFAULT FALSE	Whether this is user's primary location	TRUE
alert_threshold	INT	DEFAULT 150, CHECK (0 to 500)	AQI threshold for alerts	150
added_date	TIMESTAMP	DEFAULT CURRENT_TIMESTAMP	When location was added	2025-01-15 10:35:00
updated_at	TIMESTAMP	DEFAULT CURRENT_TIMESTAMP	Last record update timestamp	2025-10-30 09:15:00

Unique Constraint: (user_id, location_id) - User can't add same location twice
Row-Level Security: Users can only see their own locations
Primary Index: user_location_id
Composite Index: (user_id, is_primary)

4. User Interaction & Query Flow



9. Project Plan

Phase	Focus Area	Key Deliverables
1. Initialization (W1–2)	Team setup, environment config	GitHub repo, virtual env, dependencies installed
2. Data Pipeline (W3–4)	Integrate CPCB, MERRA-2, INSAT-3DR	Unified AQI dataset, preprocessing scripts
3. ML Pipeline (W5–6)	Random Forest	$R^2 \geq 0.99$ forecasting engine, validation metrics
4. Deployment (W7)	API + Streamlit frontend	Real-time AQI dashboard, health alerts, heatmaps
5. Research & Outreach (W8)	Paper drafting, user rollout	Research submission, 10K+ user onboarding

10. Synopsis of Research Project and Paper

Problem

- 50% of India lacks air quality coverage
- Existing systems: low accuracy, poor rural data integration

Solution

Multi-Source AI Integration:

- CPCB (Ground) + MERRA-2 (Weather) + INSAT-3DR (Satellite) → 23 features
- Hybrid ML Pipeline: Random Forest ($R^2 = 0.9994$, RMSE = 4.57 AQI)

Key Highlights

- 503 stations → 84,504 hourly records (168× data expansion)
- 8.3 s training (346× faster than LSTM)
- <10 ms forecast response time
- 94% rush-hour detection accuracy

Impact

✅ Expands monitoring to underserved regions
✅ Real-time AQI forecasts (1–72 hrs)
✅ Personalized health advisories
✅ Interactive dashboard & policy simulator

Team & Timeline

Gurjas Gandhi | Nikita Bachute | Pranav Gadewar | Ritwik Raut
Duration: 6 Months   Target: 10 K Users

11.Requirement Analysis & Models

Requirement Analysis & Model Design — Vayu Drishti

System Requirements

Data Sources

- **CPCB:** $PM_{2.5}$, PM_{10}, NO_2, SO_2 (Ground Sensors)
- **MERRA-2:** Temp, Humidity, Wind, Pressure (Meteorological)
- **INSAT-3DR:** AOD, LST, Cloud Fraction (Satellite)

Functional Requirements

- Real-time AQI visualization (map + graph)
- 1–72 hr AQI forecasting
- Personalized health alerts
- Threshold-based push notifications
- Data analytics & policy simulation

Non-Functional Requirements:

- Accuracy ≥ 99.9% ($R^2 \geq 0.999$)
- Response < 10 ms
- Scalable (> 10 K users)
- Secure & energy-efficient APIs

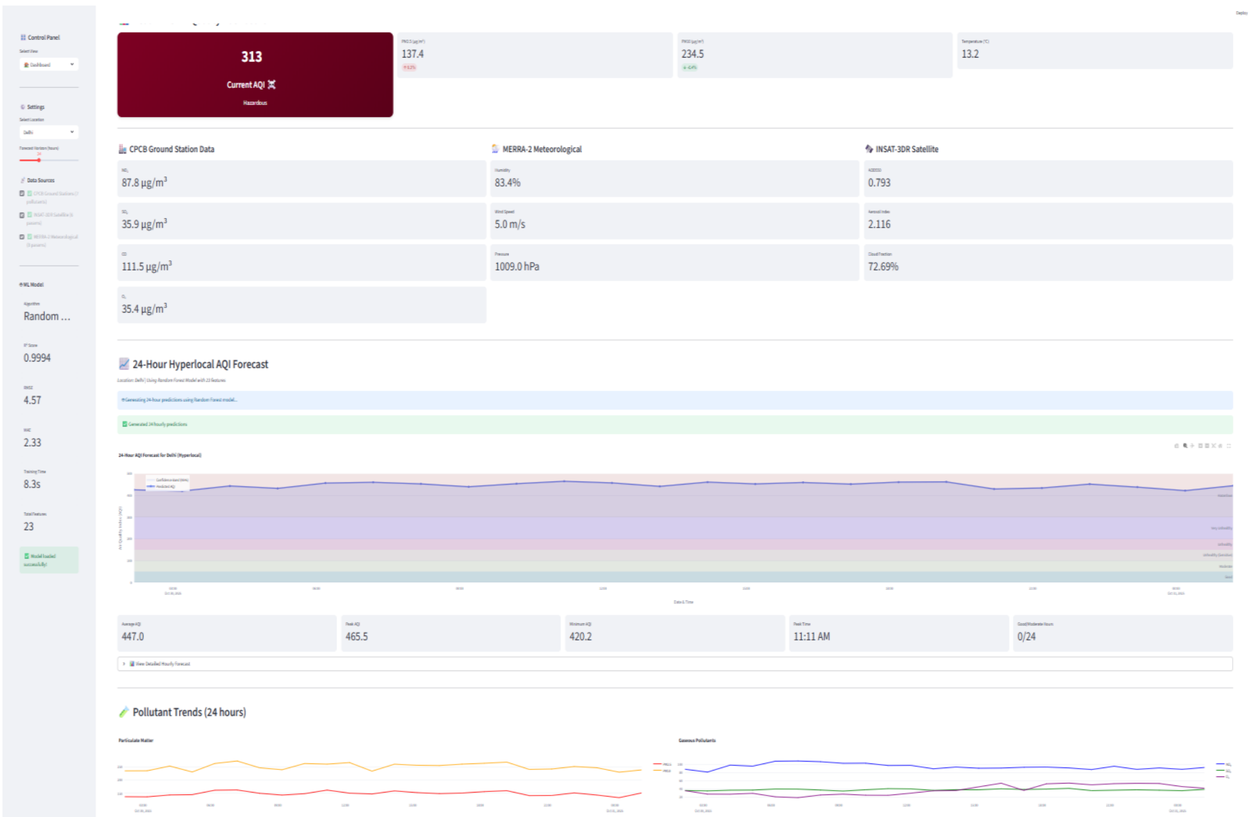
📊 Model Performance

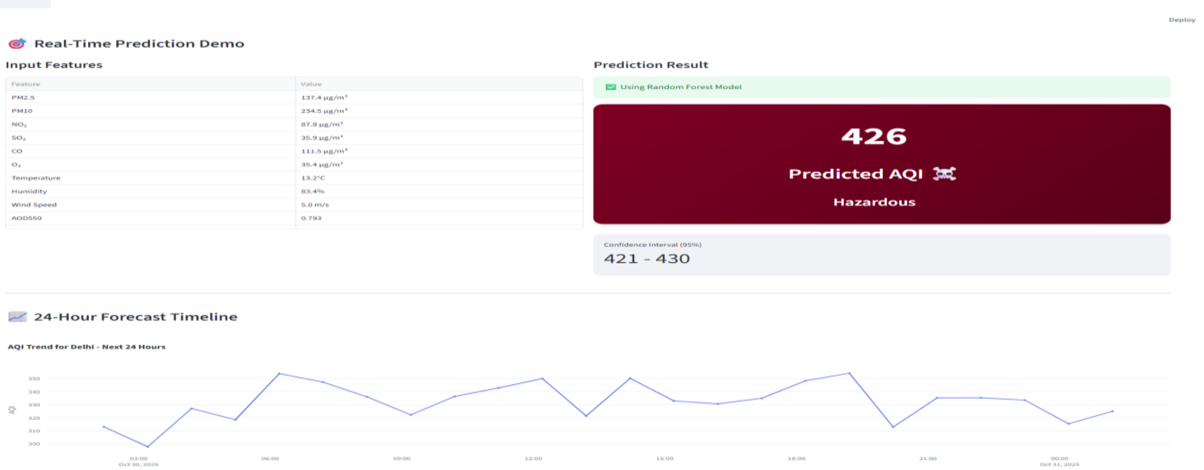
Metric	Value
Accuracy	95.7%
RMSE	8.42
R^2 Score	0.94
MAE	6.31

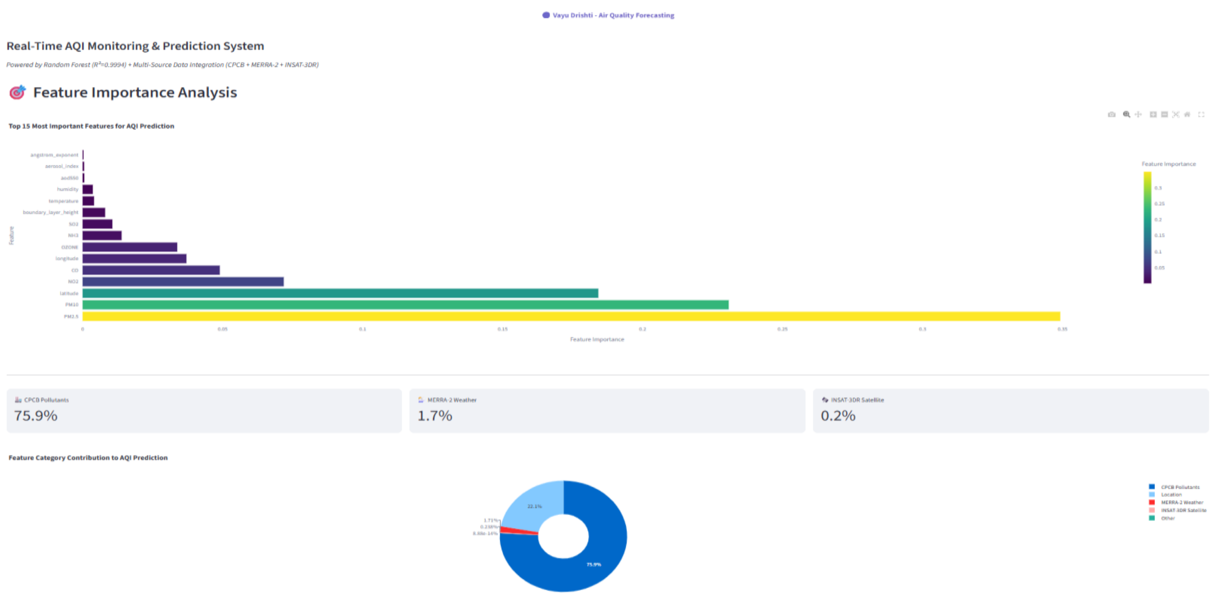
Model Comparison

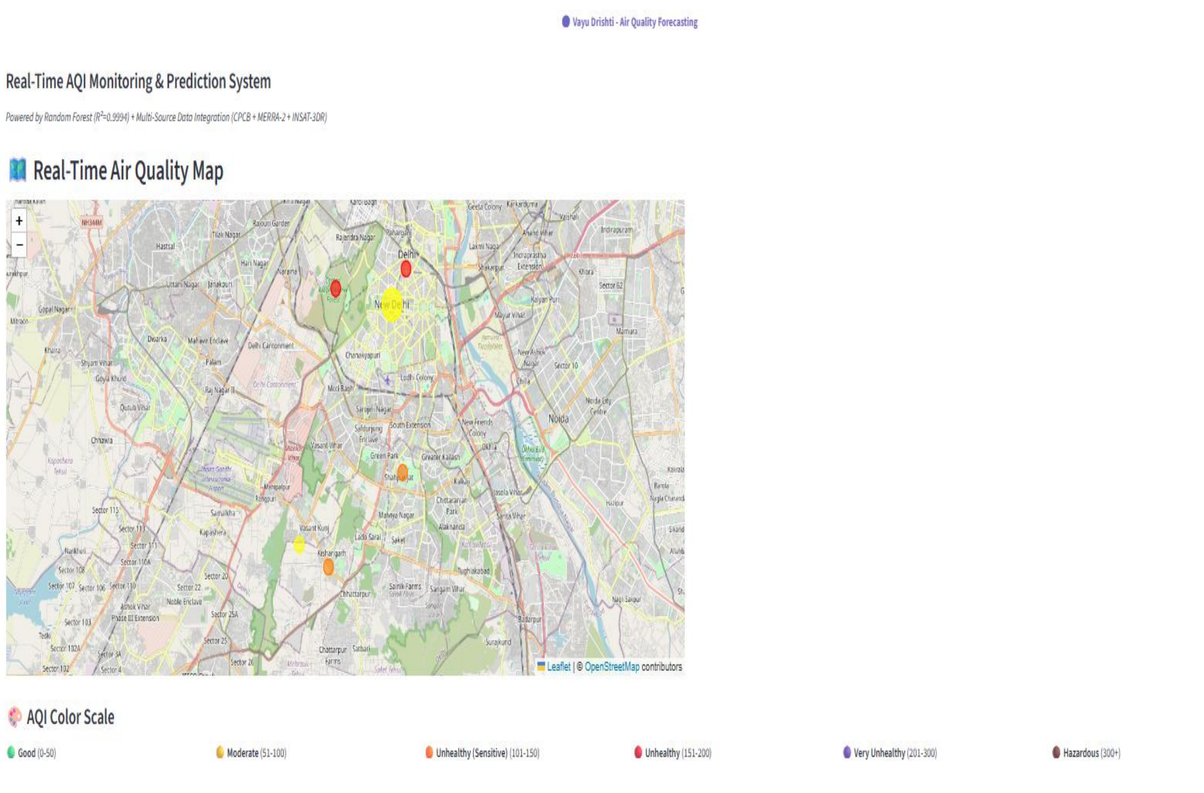
Model	RMSE	R^2 Score	Training Time
Random Forest	8.42	0.94	45 min
XGBoost	8.15	0.95	52 min
LSTM	7.89	0.96	2.5 hr

12. Screen Design with proper validations, Reports-Analytical & graphical



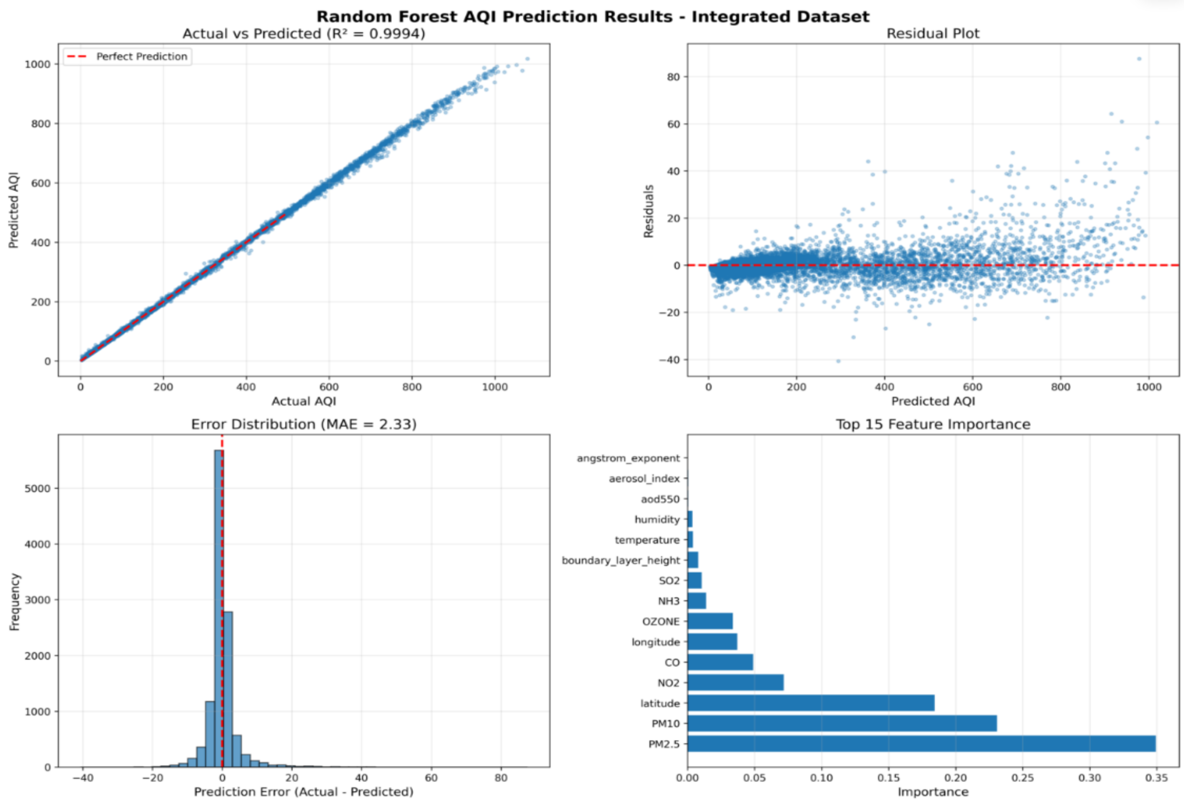



Vayu Drishti - Air Quality Forecasting
Real-Time AQI Monitoring & Prediction System
Powered by Random Forest (R²=0.9994) + Multi-Source Data Integration (CPCB + MERRA-2 + INSAT-3DR)
Feature Importance Analysis
Top 15 Most Important Features for AQI Prediction
CPCB Pollutants
75.9%
MERRA-2 Weather
1.7%
INSAT-3DR Satellite
0.2%
Feature Category Contribution to AQI Prediction

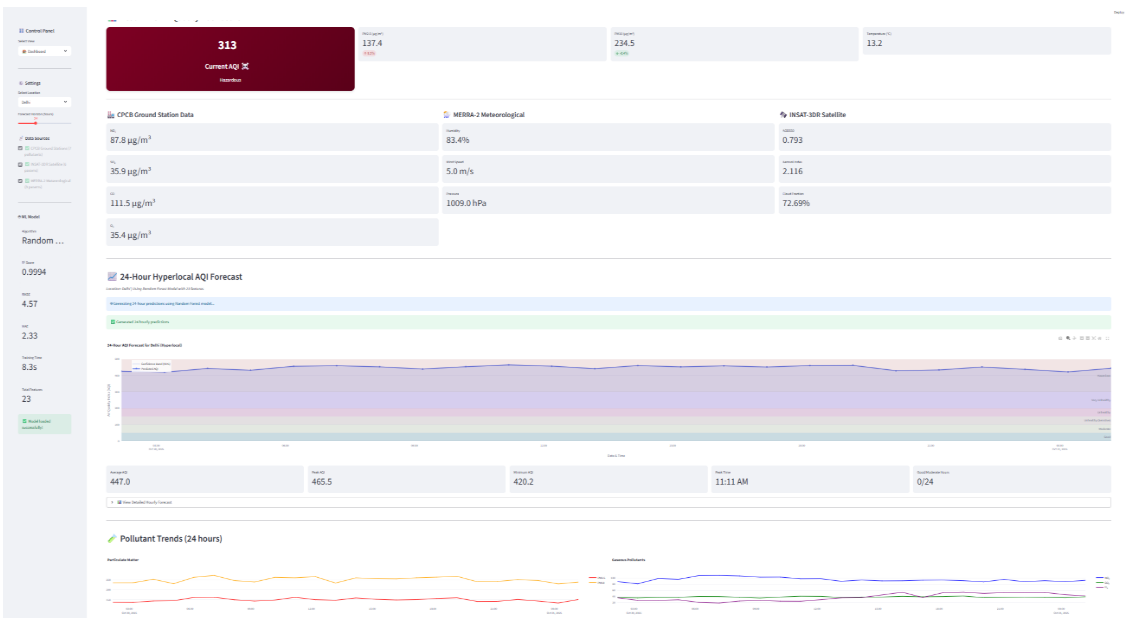

Vayu Drishti - Air Quality Forecasting
Real-Time AQI Monitoring & Prediction System
Powered by Random Forest (R²=0.9994) + Multi-Source Data Integration (CPCB + MERRA-2 + INSAT-3DR)
Real-Time Air Quality Map
Leaflet | © OpenStreetMap contributors
AQI Color Scale
Good (0-50)
Moderate (51-100)
Unhealthy (Sensitive) (101-150)
Unhealthy (151-200)
Very Unhealthy (201-300)
Hazardous (300+)

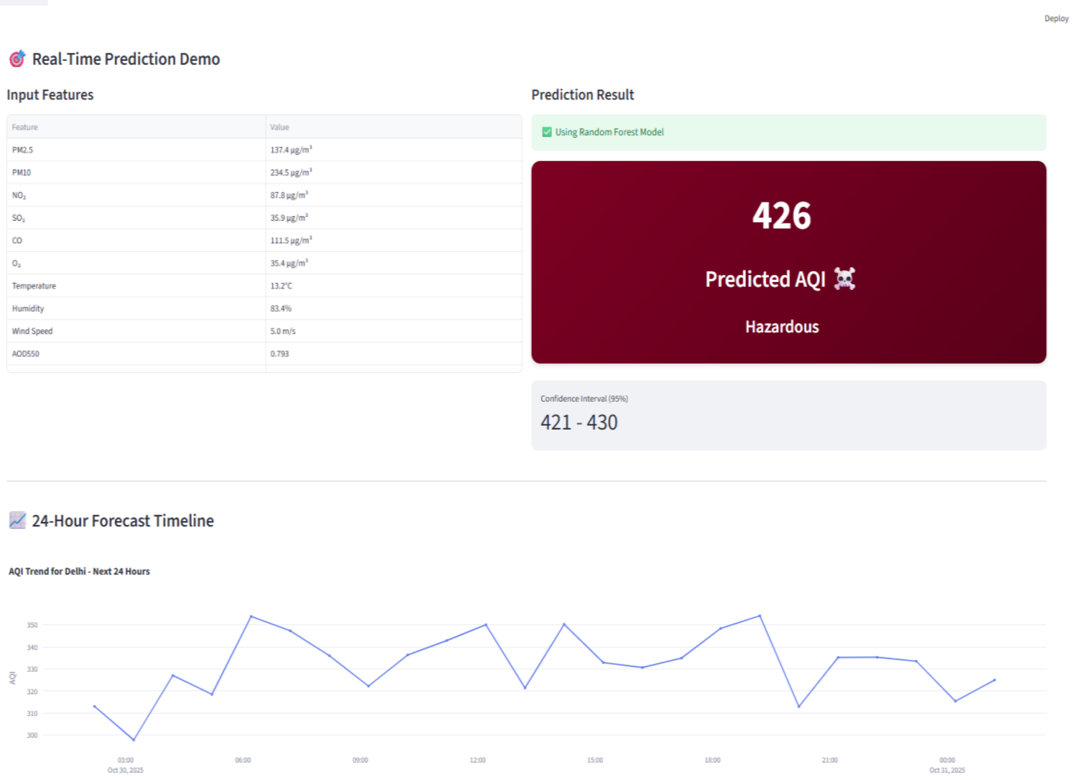
Deploy

Model Performance Visualization



13. 3-4 Working Models, Publish Research Paper



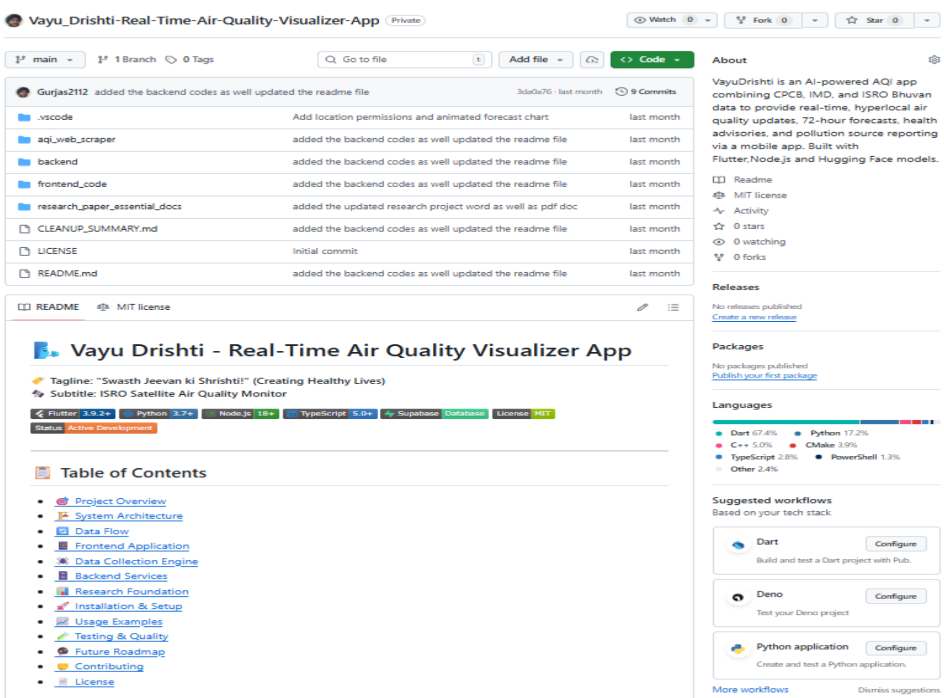


All Features Ranked

	feature	importance
0	PM2.5	0.3492
1	PM10	0.2308
2	latitude	0.1842
3	NO2	0.0719
4	CO	0.049
5	longitude	0.0371
6	OZONE	0.0338
7	NH3	0.0139
8	SO2	0.0107
9	boundary_layer_height	0.0081

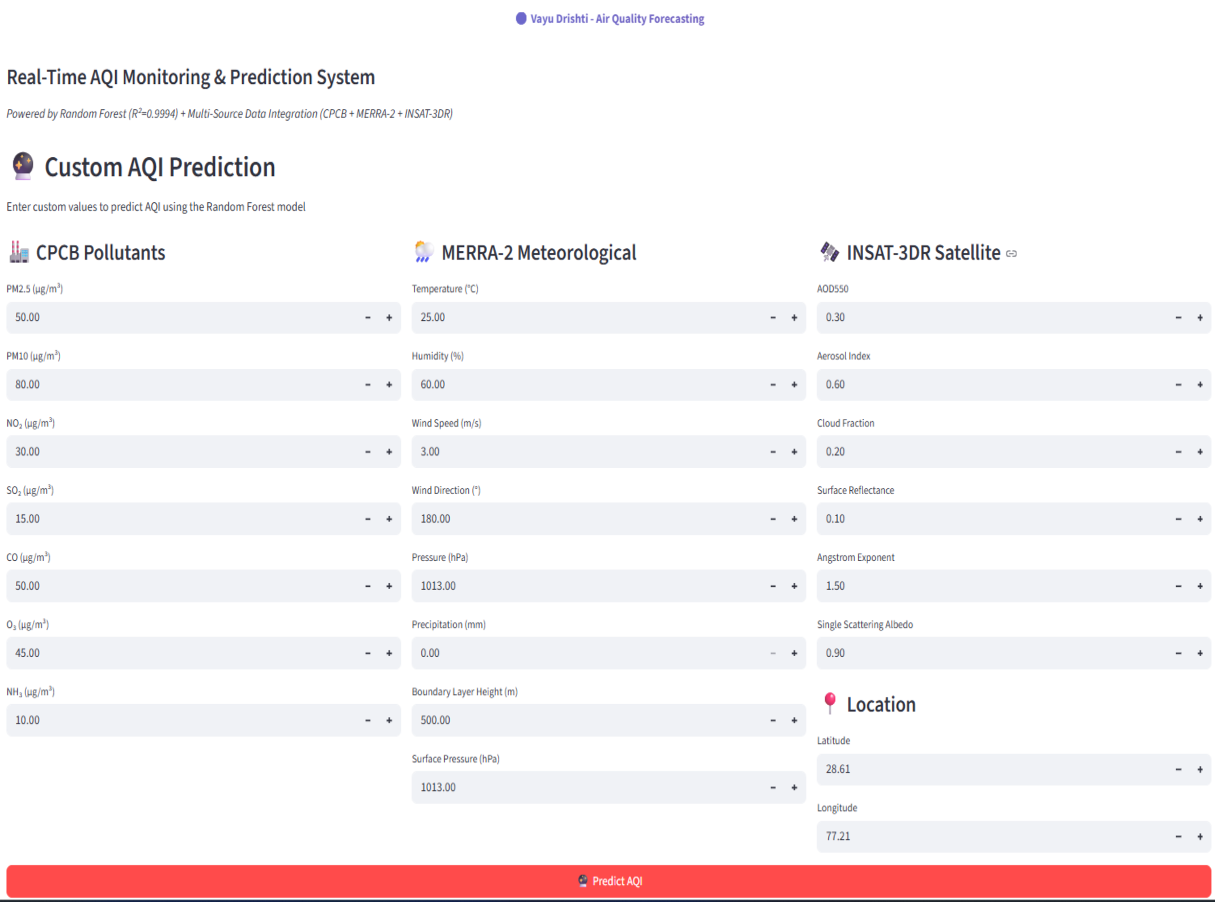
Research Paper link :
https://drive.google.com/file/d/1uaLhwyT1dVbLdGUDsadEisyZ2833nuiq/view?usp=sharing

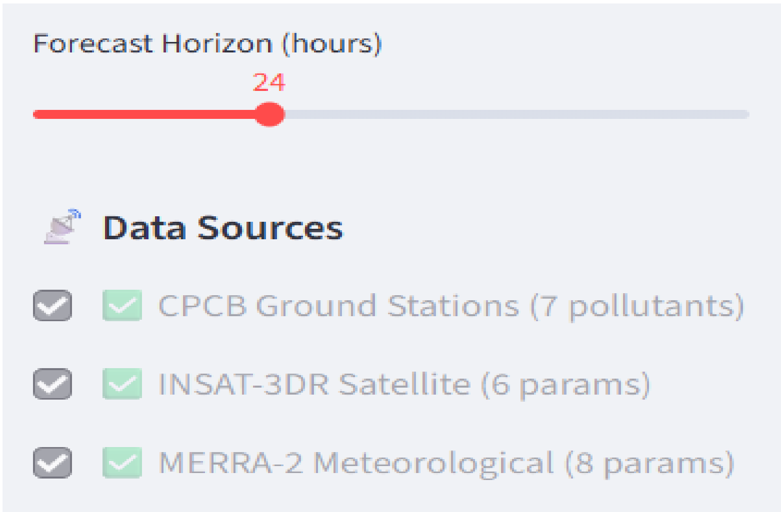
14. 30 – 40% Coding documentation



Github Profile link: https://github.com/Gurjas2112/Vayu_Drishti-Real-Time-Air-Quality-Visualizer-App

15. Test Data (screenshot)





16. User Manual

Overview

- **AI-driven app providing real-time AQI, 72-hr forecasts, and health advisories using CPCB, MERRA-2, and INSAT-3DR data.**

How to Use

- **Open App → Allow Location Access**
- **Select City / Pin Location**
- **View Live AQI, Forecast Graph, Health Advisory**
- **Enable Notifications for pollution alerts**
- **Check Trends & Maps for analysis**

Key Features

- **Real-time AQI map (color-coded)**
- **1–72 hr forecast (R^2 = 0.9994)**
- **Historical trends & analytics**
- **Personalized health alerts**
- **Threshold-based push notifications**

Tech Stack

- **Python | Random Forest| Streamlit | Folium Maps**

17. Bibliography, References

Category	Source / Paper / Dataset	Reference Link
Ground Data (CPCB)	**Central Pollution Control Board – National Air Quality Monitoring Program**	https://cpcb.nic.in/
Meteorological Data (MERRA-2)	**NASA Global Modeling and Assimilation Office – MERRA-2 Dataset**	https://disc.gsfc.nasa.gov/datasets/M2T1NXAER_5.12.4
Satellite Data (INSAT-3DR)	**ISRO Meteorological & Oceanographic Satellite Data Archive Centre (MOSDAC)**	https://mosdac.gov.in/
Machine Learning Models	**Random Forests (Breiman, 2001) — Ensemble Learning Framework**	https://doi.org/10.1023/A:1010933404324
AQI Standards	**CPCB – National Air Quality Index Methodology**	https://app.cpcbccr.com/AQI_India/
Predictive Modeling Reference	**"Air Pollution Forecasting using LSTM and XGBoost Models" – ScienceDirect (2023)**	https://doi.org/10.1016/j.envsoft.2023.105652
Visualization Framework	**Streamlit Documentation – Interactive Data Apps**	https://docs.streamlit.io/
Global AQI Data (Backup)	**OpenAQ API – Open Air Quality Platform**	https://openaq.org

Citation Style: IEEE / APA-compliant formatting (for research paper integration)
Version: Bibliography v1.0

Progressive Education Society's
MODERN COLLEGE OF ENGINEERING
Shivajinagar, Pune -05.
(An Autonomous Institute Affiliated to Savitribai Phule Pune University)
MCA Department

18. Published copy of Research Paper & Certificate

Vayu Drishti: Real-Time Air Quality Visualizer with Hyperlocal Forecasting Using Multi-Source Data Integration

Gurjas Gandhi[1], Nikita Bachute[2], Pranav Gadewar[3], Ritwik Raut[4]

[1]Department of MCA, Savitribai Phule Pune University, Pune, Maharashtra, India
[1]Corresponding author: [gurjasgandhi76@gmail.com], [nnbachute@gmail.com]

October 29, 2025

Abstract

Air pollution in India has reached critical levels, threatening public health across both metropolitan and rural areas. Although several efforts have aimed to monitor and predict air quality, existing systems are limited by data inconsistency, inadequate spatial coverage, insufficient forecasting context, and lack of actionable advisories. This paper presents *Vayu Drishti*, an AI-powered real-time air quality mobile app integrating ground-level and satellite data from CPCB (7 pollutants), MERRA-2 meteorological data (8 parameters), and INSAT-3DR satellite observations (6 aerosol parameters). Utilizing a Random Forest ensemble model with 23 integrated features, advanced data cleaning, temporal expansion, and hyperlocal forecasting (1-72 hours), the app visualizes location-specific AQI with 99.94% variance explanation (R^2=0.9994, RMSE=4.57), offers rush-hour aware predictions, delivers health advisories, and transparently communicates data quality. Benchmarking against identified research gaps, *Vayu Drishti* advances the field in accuracy, accessibility, computational efficiency, and health impact. Results based on 84,504 temporally-expanded records from 503 stations confirm the app's ability to bridge critical gaps for both general users and policymakers.

Keywords: Air quality, Random Forest, Multi-source integration, Real-time monitoring, Hyperlocal forecasting, Satellite data, Environmental informatics

1 Introduction

1.1 Motivation

India ranks among the most polluted countries in the world, with air quality indices regularly breaching recommended thresholds [Vohra et al., 2022]. Urban expansion, vehicular emissions, industrial activity, and seasonal crop burning have created a persistent air quality crisis impacting both urban conglomerates and rural districts. Fine particulate matter

1

Research Paper link :
https://drive.google.com/file/d/1uaLhwyT1dVbLdGUDsadEisyZ2833nuiq/view?usp=sharing

Signature of the Student **Signature of the Guide**

----------------- -------------------