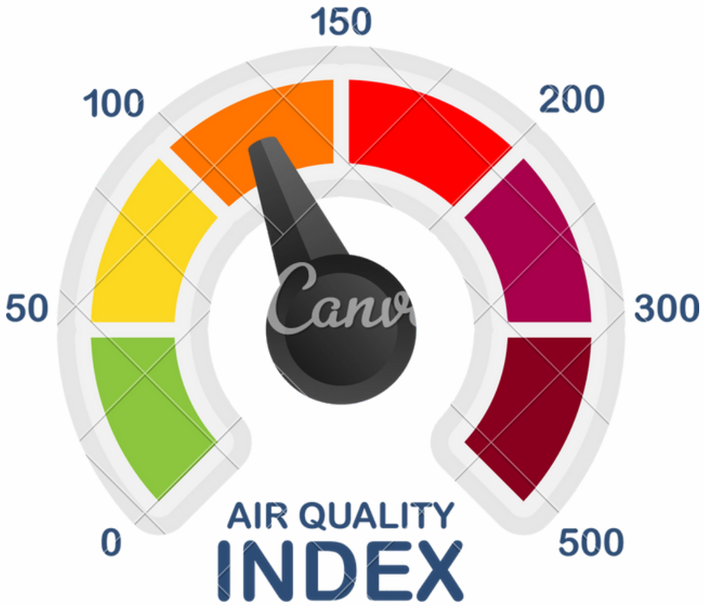
Breathing Easy: Forecasting Air Quality with Real-time and Historical Data

MAHARSH SONI

Project Background & Executive Summary

- Current systems do prediction a 24 hour or 8 hour period before hand
- This leads to poor prediction as external factors can affect the quality of air throughout the day
- Aim - An improved Air Quality Prediction system for California at the the county level using data at latitude-longitude level of granularity for calculation using real-time data



Air Quality Index Readings

AQI Basics for Ozone and Particle Pollution

Daily AQI Color	Levels of Concern	Values of Index	Description of Air Quality
Green	Good	0 to 50	Air quality is satisfactory, and air pollution poses little or no risk.
Yellow	Moderate	51 to 100	Air quality is acceptable. However, there may be a risk for some people, particularly those who are unusually sensitive to air pollution.
Orange	Unhealthy for Sensitive Groups	101 to 150	Members of sensitive groups may experience health effects. The general public is less likely to be affected.
Red	Unhealthy	151 to 200	Some members of the general public may experience health effects; members of sensitive groups may experience more serious health effects.
Purple	Very Unhealthy	201 to 300	Health alert: The risk of health effects is increased for everyone.
Maroon	Hazardous	301 and higher	Health warning of emergency conditions: everyone is more likely to be affected.

Source: https://www.airnow.gov/aqi/aqi-basics/

Literature Review

Name of the Article	About the Dataset	Models Used	Results
Air Quality Prediction using Machine Learning Algorithms – A Review	The findings are from other research papers, the authors have not used their own datasets.	Neural Network, Random Forest (RF), XG-Boost, Artificial Neural Network (ANN), Recurrent neural networks (RNN) etc.	Best performing models used Neural Network: 92.3-99.56% RF: 70-90% XG: 99.51% ANN: 84%
A Machine Learning Model for Air Quality Prediction for Smart Cities	The dataset includes various pollutants like CO, NO2, SO2, and particulate matter (PM2.5 and PM10).	Neural Networks, Multiple Linear Regression (MLR), Support Vector Machine (SVM), and Long Short-Term Memory (LSTM) networks	Best performing models used Neural Networks: 91.62% SVM: 97.3%
Liang et al., (2020) Machine Learning-Based Prediction of Air Quality	Hourly air quality and meteorological data from 79 monitoring stations across Taiwan	Adaptive boosting (AdaBoost), Artificial neural network (ANN), Random forest, Stacking ensemble, and Support vector machine (SVM)	Best performing models used: AdaBoost and Stacking ensemble R2 score : AdaBoost-Linear - 0.734 Stacking ensemble - 0.735
Federated Learning for Air Quality Index Prediction using UAV Swarm Networks	Hourly and daily AQI (Air Quality Index) values for several cities in India gathered from the Central Pollution Control Board of India's official website.	Long short-term memory networks (LSTM) Support Vector Machine (SVM), K-Nearest Neighbour (kNN), XGBoost	LSTM model performed the best with the scores of RMSE: 56.222, MAE: 41.219, and MAPE: 24.184:

HYBRID WATERFALL AND CRISP-DM MODEL

Data Understanding

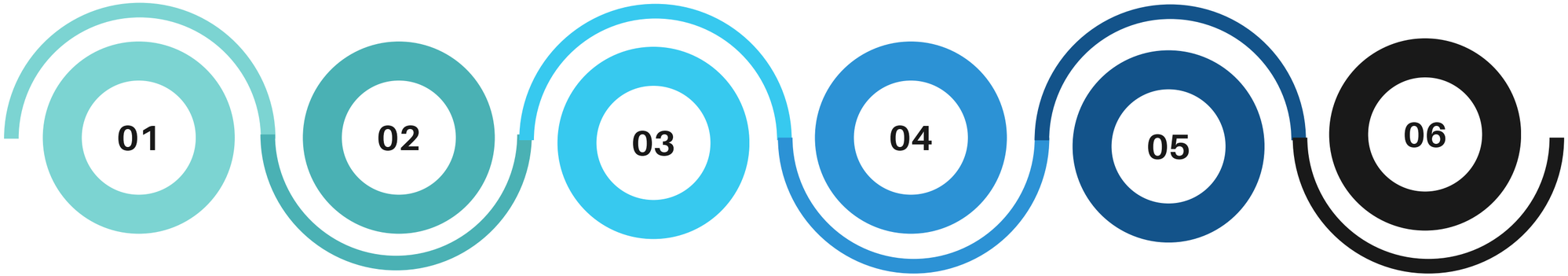
- Collect and explore real-time air quality data.
- Collect historical AQI data.
- Identifying relationships, missing values, and outliers.

Data Modeling

- Choose ML models.
- Train them to predict air quality.
- Test the trained models.

Deployment

- Deploy the model in a web application using Flask/Dash.
- Monitor its performance.
- Manage.
- Keep updated for reliability.



Business Understanding

- Defining the problem statement - Accurate prediction of AQI for lowering pollution and enhancing public health.

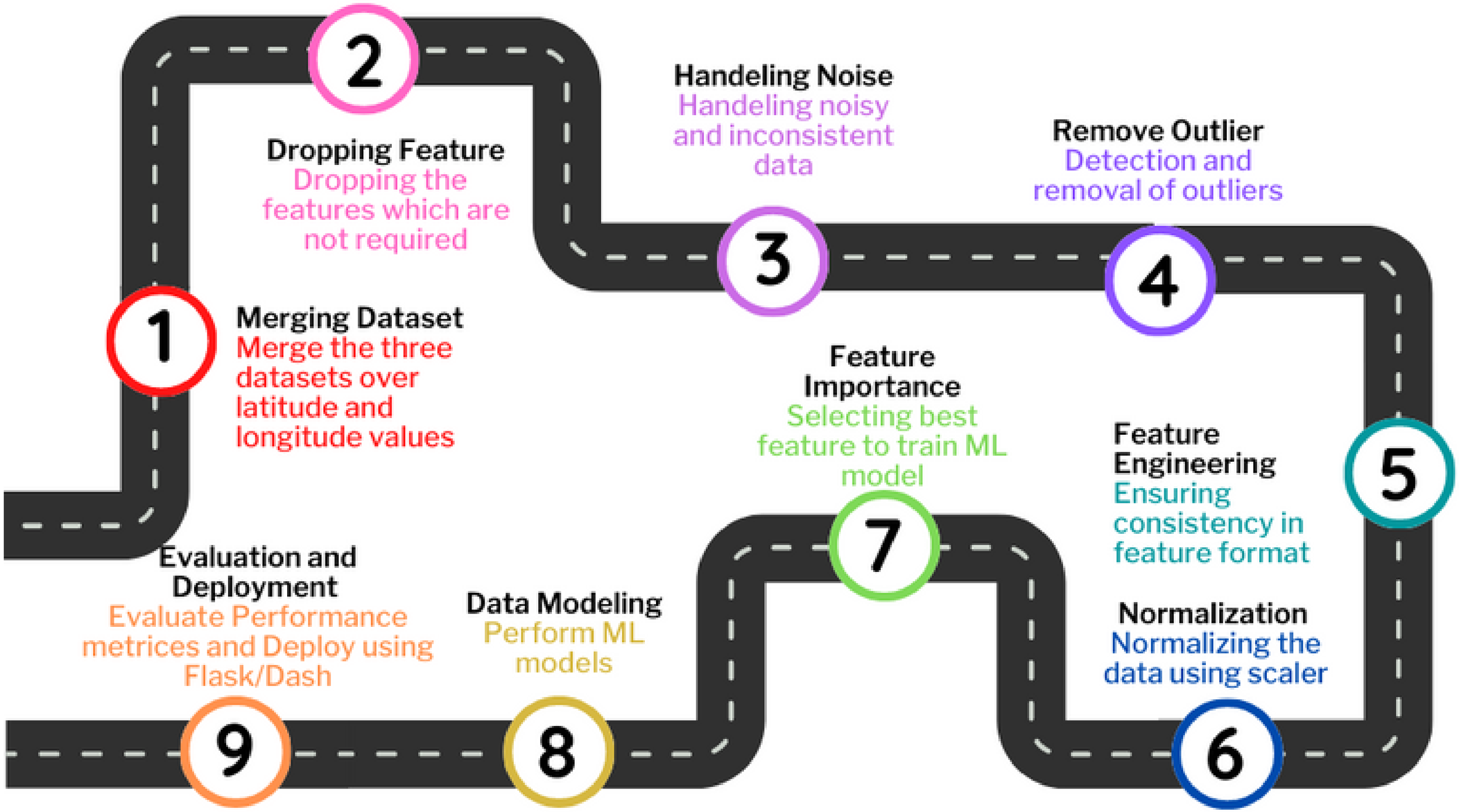
Data Preparation

- Prepare data for machine learning by cleaning, selecting features, handling missing values, and scaling.

Data Evaluation

- Evaluate the model's performance using fresh data, taking into account parameters like accuracy, precision, and recall.
- Determine what needs to be improved..

PROCESS FLOW DIAGRAM



PROJECT RESOURCES AND REQUIREMENTS

Function	Resource Type	Resources
Testing and Training Model	Hardware	CPU/GPU
Development	Software	Google Colab, Python, pandas, NumPy, XGboost, Sklearn and other Libraries
Visualization	Software	Matplotlib, Seaborn
Documentation	Software	Canva, Google Docs, Grammarly

PROJECT RESOURCES AND REQUIREMENTS

Function	Resource Type	Resources
Collaboration	Software	Zoom
Workflow	Software	Trello
Data Storage	Software	Google Drive

Data Collection Process

Data Source

Data collected from EPA and AirNow website - More than 2000 monitoring sites

Locations

All of California listed by Latitude and Longitude

Data Files

Ozone_California_data, AirNow, PM2.5_California_data - CSV Format

TIme Period of Data

We have collected data starting 2020 till present

RAW DATA

Raw datasets of Ozone, PM2.5 and Live from AirNow API.

- Ozone Dataset: 190583 Rows & 20 Columns

Date	Source	Site ID	POC	Daily Max	UNITS	DAILY_AQ	Site Name	DAILY_OB	PERCENT_	AQS_PAR/	AQS_PAR/	CBSA_COI	CBSA_NAI	STATE_CO	STATE	COUI
1/1/2020	AQS	60010007	1	0.025	ppm	23	Livermore	17	100	44201	Ozone	41860	San Franci	6	California	
1/2/2020	AQS	60010007	1	0.017	ppm	16	Livermore	17	100	44201	Ozone	41860	San Franci	6	California	
1/3/2020	AQS	60010007	1	0.013	ppm	12	Livermore	17	100	44201	Ozone	41860	San Franci	6	California	
1/4/2020	AQS	60010007	1	0.028	ppm	26	Livermore	17	100	44201	Ozone	41860	San Franci	6	California	
1/5/2020	AQS	60010007	1	0.031	ppm	29	Livermore	17	100	44201	Ozone	41860	San Franci	6	California	
1/6/2020	AQS	60010007	1	0.031	ppm	29	Livermore	17	100	44201	Ozone	41860	San Franci	6	California	
1/7/2020	AQS	60010007	1	0.032	ppm	30	Livermore	17	100	44201	Ozone	41860	San Franci	6	California	

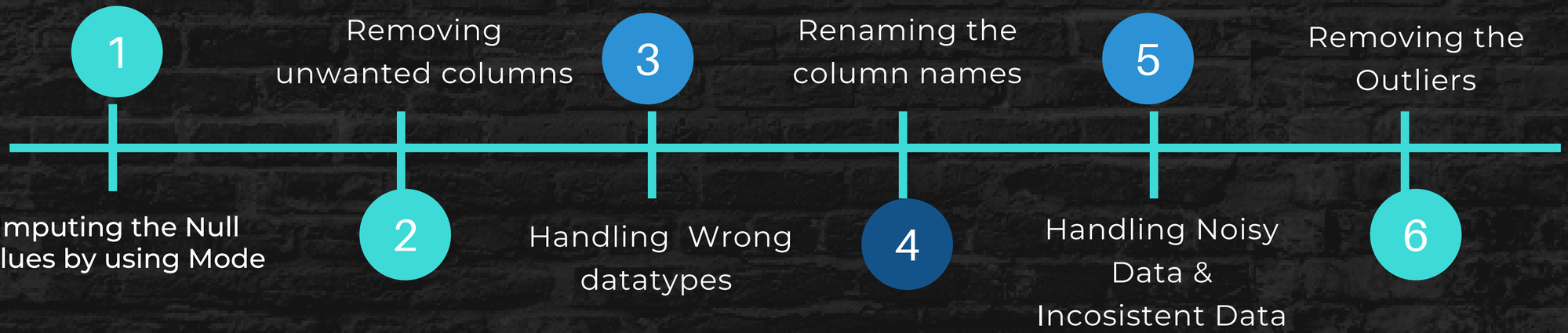
- PM2.5 Dataset: 183768 Rows & 20 Columns

Date	Source	Site ID	POC	Daily Mea	UNITS	DAILY_AQ	Site Name	DAILY_OB	PERCENT_	AQS_PAR/	AQS_PAR/	CBSA_COI	CBSA_NAI	STATE_CO	STATE	CC
1/1/2020	AQS	60010007	3	8.6	ug/m3 LC	36	Livermore	1	100	88101	PM2.5 - Lc	41860	San Franci	6	California	
1/2/2020	AQS	60010007	3	4.5	ug/m3 LC	19	Livermore	1	100	88101	PM2.5 - Lc	41860	San Franci	6	California	
1/3/2020	AQS	60010007	3	14.2	ug/m3 LC	55	Livermore	1	100	88101	PM2.5 - Lc	41860	San Franci	6	California	
1/4/2020	AQS	60010007	3	10.9	ug/m3 LC	45	Livermore	1	100	88101	PM2.5 - Lc	41860	San Franci	6	California	
1/5/2020	AQS	60010007	3	7.8	ug/m3 LC	33	Livermore	1	100	88101	PM2.5 - Lc	41860	San Franci	6	California	
1/6/2020	AQS	60010007	3	6.2	ug/m3 LC	26	Livermore	1	100	88101	PM2.5 - Lc	41860	San Franci	6	California	
1/7/2020	AQS	60010007	3	6.9	ug/m3 LC	29	Livermore	1	100	88101	PM2.5 - Lc	41860	San Franci	6	California	

- Live Dataset from AirNow API: 2 Rows & 10 Columns

DateObserved	HourObserved	LocalTimeZone	ReportingArea	StateCode	Latitude	Longitude	ParameterName	AQI	Concentration
5/8/2023	8	PST	San Jose	CA	37.33	-121.9	O3	24	0.028
5/8/2023	8	PST	San Jose	CA	37.33	-121.9	PM2.5	18	4.6

Pre-processing



Data Transformation

MERGED DATASET

	Date	Site ID	Ozone	DAILY_AQI_VALUE_x	DAILY_OBS_COUNT_x	SITE_LATITUDE_x	SITE_LONGITUDE_x
0	01/01/2020	60010007	0.025	23	17	117	21
1	01/02/2020	60010007	0.017	16	17	117	21
2	01/03/2020	60010007	0.013	12	17	117	21
3	01/04/2020	60010007	0.028	26	17	117	21
4	01/05/2020	60010007	0.031	29	17	117	21

PM2.5	DAILY_AQI_VALUE_y	DAILY_OBS_COUNT_y	SITE_LATITUDE_y	SITE_LONGITUDE_y
8.6	36	1	117	21
4.5	19	1	117	21
14.2	55	1	117	21
10.9	45	1	117	21
7.8	33	1	117	21

Final merged data - 128640 rows and 12 columns

Feature Selection

Descriptive Variables

Latitude of the monitoring site

Longitude of the monitoring site

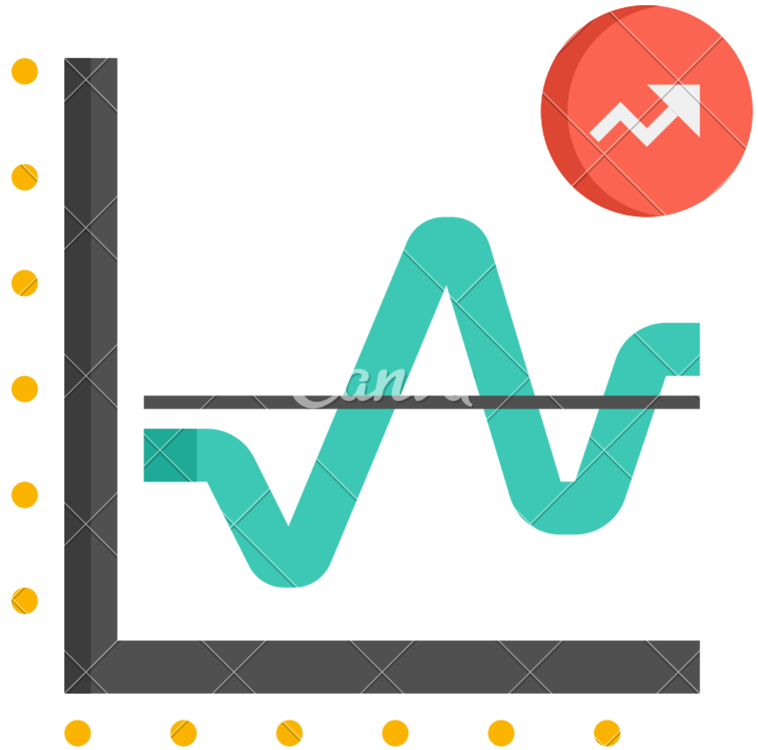
Ozone Concentration

PM2.5 Concentration

Target Variable

Daily AQI value of Ozone

Daily AQI value of PM2.5

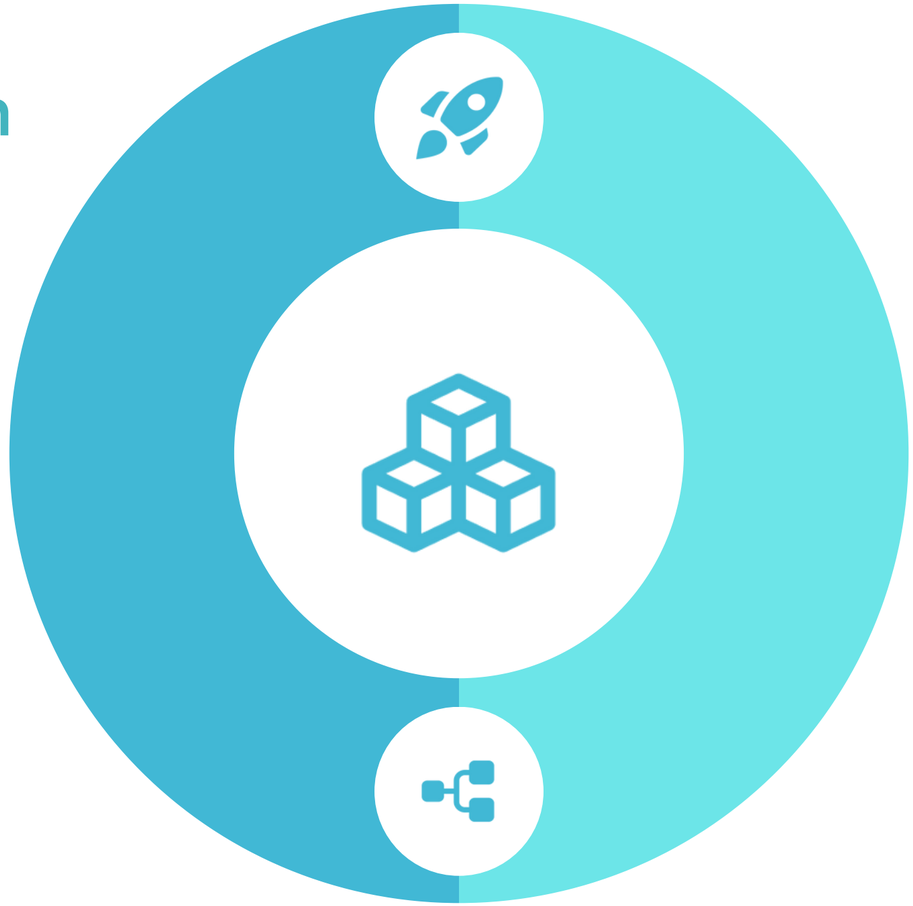


MODEL PROPOSAL

01 - Linear Regression

Supervised ML Model, which aims to find a best fit linear line.

- Implements gradient boosting trees.
- Characteristics:
  - Easy to implement
  - Avoid overfitting
  - updated easily using stochastic gradient descent.



02 - Adaboost with LR

Each weak learner is a linear regression model that tries to find the best-fit line that minimizes the mean squared error. The weighted predictions of all the linear regression models are combined to make the final prediction.

- Accurate Predictions
- Avoids Overfitting
- Easy to Implement
- Updated Easily

MODEL PROPOSAL

03 - Neural Networks

It recognizes patterns and makes predictions from data.

- Characteristics:
  - Inspired by the human brain
  - Recognize complex patterns in data
  - Require large labeled data and are computationally expensive
  - Have different architectures with unique strengths and weaknesses.

04 - LSTM

It captures long-term dependencies in sequential data.

- Characteristics:
  - Good for sequential data with long-term dependencies
  - Memory and forgetfulness
  - Used for language, speech, and time series
  - Require large labeled data and can be computationally expensive.

MODEL PROPOSAL

05 - XGBoost

An Ensemble Learning technique that uses boosting.

- Implements gradient boosting trees.
- Characteristics:
  - Performance
  - Parallelization
  - Computation Speed

06 - Random Forest

An Ensemble learning technique that uses bagging.

- Implements decision trees on randomly selected samples.
- combines results from different models and takes average.

Training & Test Datasets

X_train

Rows: 128640
Columns: 3

X_test

Rows: 1
Columns: 3

y_train

Rows: 128640
Columns: 3

y_test

Rows: 1
Columns: 3

We used time-series cross validation to train the models

Model Evaluation: Ozone

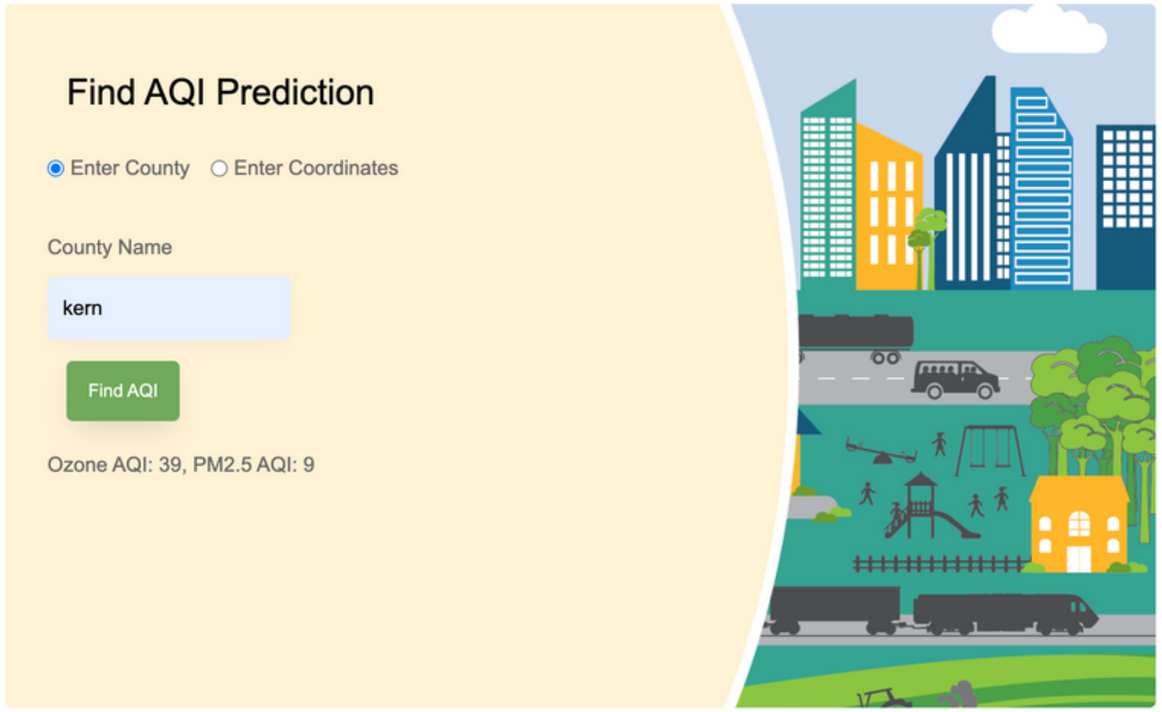
Models	R2	MSE	RMSE	MAE
Linear Regression	0.68	211.75	14.55	5.21
Adaboost with LR	0.78	140.5	11.85	5.47
LSTM	0.70	197.82	14.06	4.93
XGBoost	0.86	94.56	9.72	1.91
Random Forest	0.84	104.82	10.24	2.12
Neural Network	0.62	248.84	15.77	5.50

Model Evaluation: PM2.5

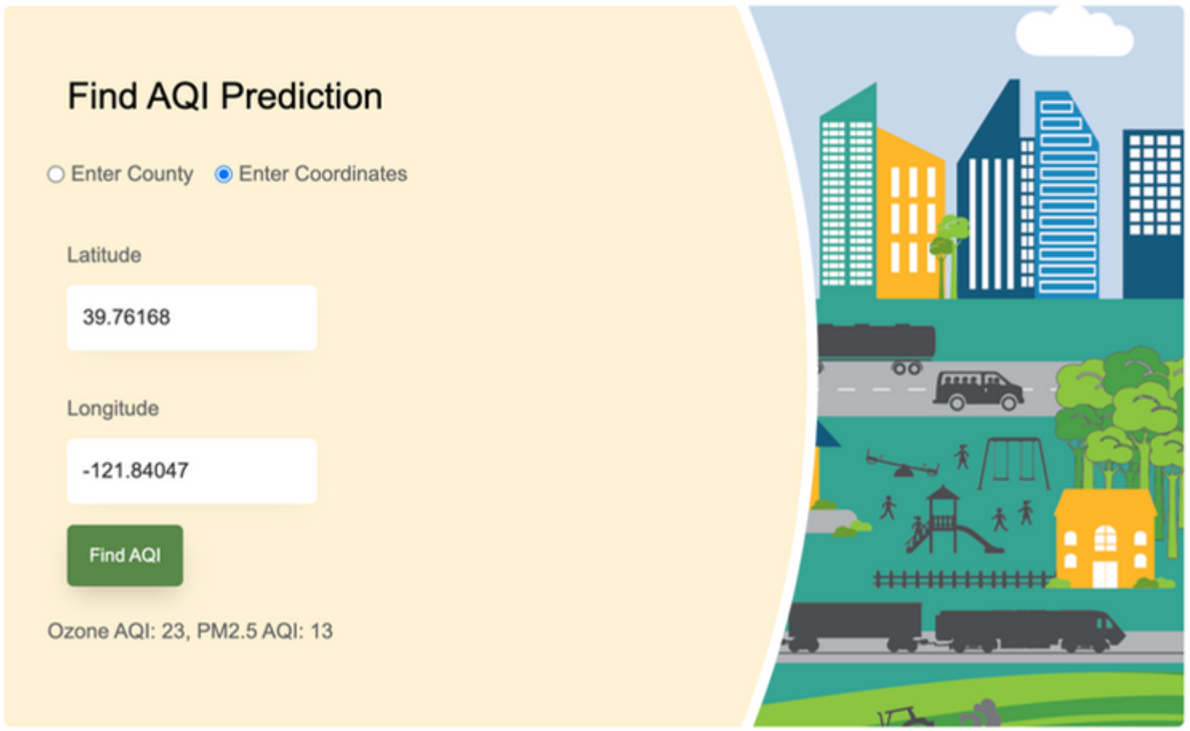
Models	R2	MSE	RMSE	MAE
Linear Regression	0.59	279.5	16.71	4.56
Adaboost with LR	0.68	216.10	14.7	4.74
LSTM	0.95	31.36	5.60	1.02
XGBoost	0.98	13.25	3.64	0.22
Random Forest	0.98	13.51	3.67	0.224
Neural Networks	0.93	48.80	6.98	0.78

User Interface

Breathing Easy



Breathing Easy



THANK YOU