AQI Prediction Using Machine Learning in Indian Cities

CONTENTS

01

Introduction

Overview of the Project

The project focuses on predicting the Air Quality Index (AQI) using machine learning techniques specifically tailored for various tailored for various cities across India.

Presented by: kunal ,krishnam,prateek ,prateek dubey,krishna,harshit

Course :Introduction to Ai

Institution: KIET Group of institution

Date: [Insert Date] 27 may 2025

02

Problem Statement

Why Predict AQI?

Air Pollution Concern

Air pollution is a growing concern in in many urban areas in India, leading to leading to various health issues.

Air Quality Index

The Air Quality Index (AQI) serves as a standardized measure to assess how polluted the air is in specific locales.

AQI Prediction

Timely prediction of AQI can greatly assist in public health planning and enable individuals to make informed decisions regarding outdoor activities.

03

Dataset Overview

Dataset Description

1

Data Source

Source: Kaggle – Air Quality Data in India

2

Time Frame

Time Period: 2015–2020

3

Cities Included

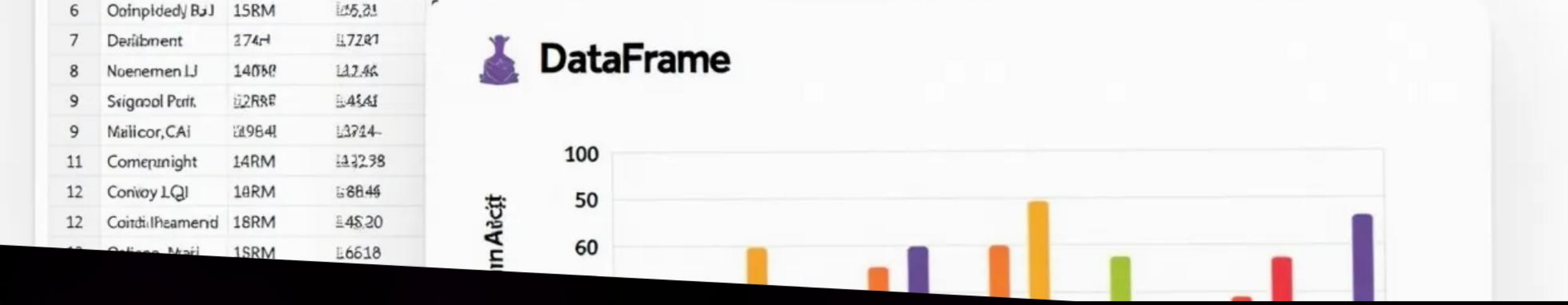
Cities: Multiple major Indian Indian cities across the country country

4

Key Features

Key Features Included:

- PM2.5, PM10, NO2, SO2, CO, O3
- Date, City, AQI



Dataset Description

Data Visualization

Visualization of the dataset can include:

.regression graph representing Actual vs predicted Aqi

04

Data Preprocessing

Data Cleaning & Preparation

Data Integrity

Removed any missing and duplicate entries to ensure data integrity.

Date Conversion

Converted string date formats into datetime objects for easier for easier analysis.

NaN Values Handling

Handled NaN values appropriately using methods such as such as interpolation or mean imputation.

Dataset Normalization

Normalized the datasets using techniques like MinMaxScaler or StandardScaler to prepare data for modeling.

Code Snapshot

Data Cleaning & Preparation

Code Snapshot

05

Understanding Code

Understanding the Data

AQI Trends Analysis

Analyzed the trends of AQI over different time periods to gauge fluctuations.

Correlation Heatmap

Generated a correlation heatmap to to understand relationship strengths strengths among various pollutants. pollutants.

Polluted Cities

Identified the most polluted cities based on available data.

Understanding the Data

Visualizations

Created visualizations such as:

- Line plots depicting AQI trends over time.
- Heatmaps illustrating correlation between pollutants.
- Bar charts representing average AQI levels categorized by city.

06

Model Selection - Random forest Regression

Why ?

Ensemble of Decision Trees

A Random Forest is is made up of many many Decision Trees. Trees. Each tree is trained on a random random subset of the data. This is called bagging (Bootstrap Aggregating). The idea is to reduce overfitting and increase accuracy

Feature Randomness

When a tree is split split at each node, a a random subset of of features is chosen chosen rather than than using all features. This helps helps ensure the trees are less correlated with each each other, which improves the model's generalization

Voting or Averaging

For classification, each tree votes for a class, and the majority vote is the final prediction.

For regression, the final output is the average of all tree predictions.

.

07

Model Evaluation

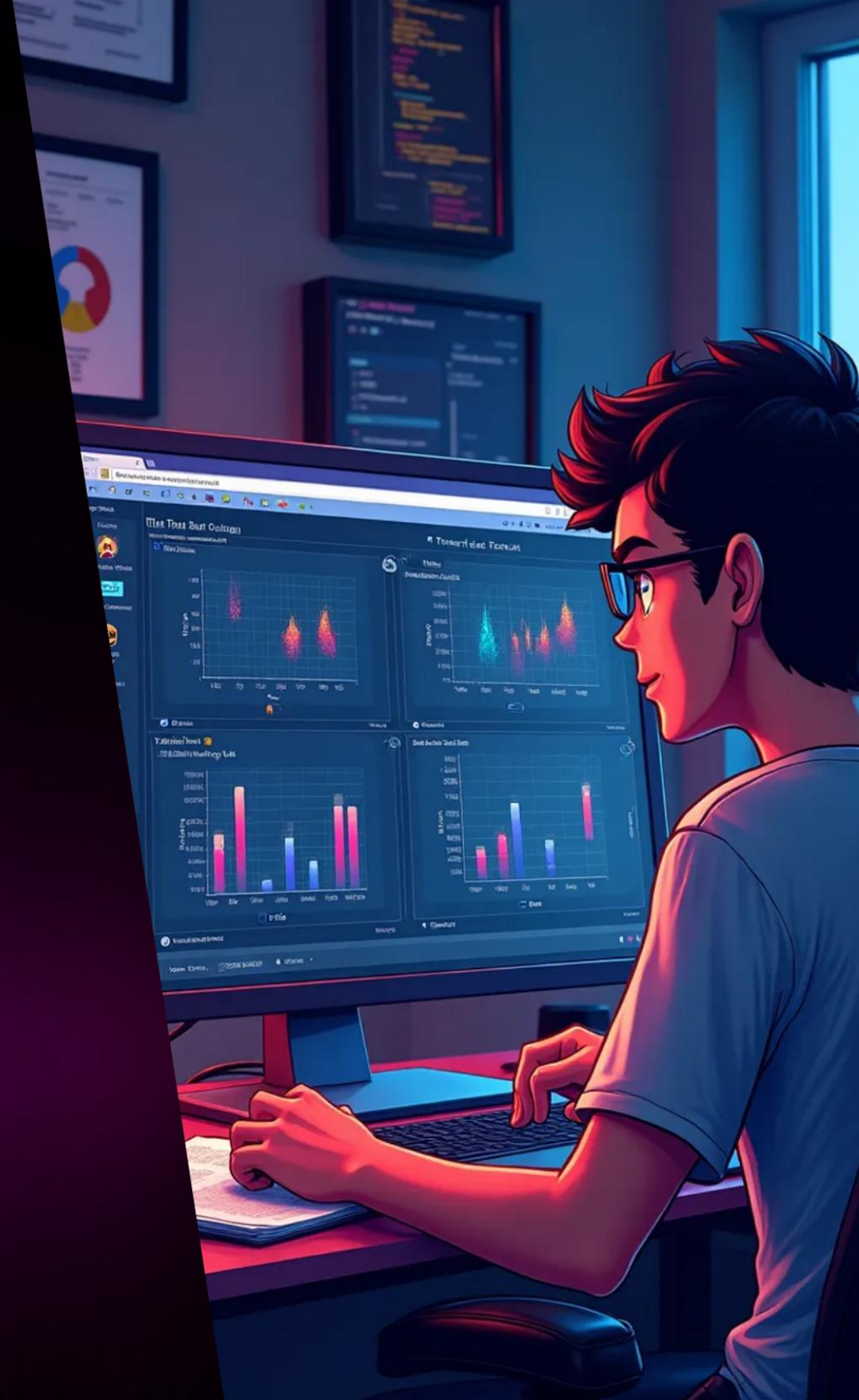
Model Performance.

Metrics:

Mean Squared Error (MSE): e.g., 110.45

R^2 Score: e.g., 0.89 (means 89% of variance in AQI is explained by the model) model)

✅ High R^2 indicates good model fit.



CODE:-

```
from google.colab import files

import zipfile

import io

import pandas as pd

import matplotlib.pyplot as plt

import seaborn as sns

from sklearn.model_selection import train_test_split

from sklearn.ensemble import RandomForestRegressor

from sklearn.preprocessing import StandardScaler

from sklearn.metrics import mean_squared_error, r2_score

# Step 1: Upload the ZIP file

print("Upload the ZIP file containing the AQI dataset...")

uploaded = files.upload()

# Step 2: Extract ZIP contents

for file_name in uploaded.keys():
```

Model Performance

Visualization

A predicted vs actual AQI scatter plot was created to visualize the model's performance.

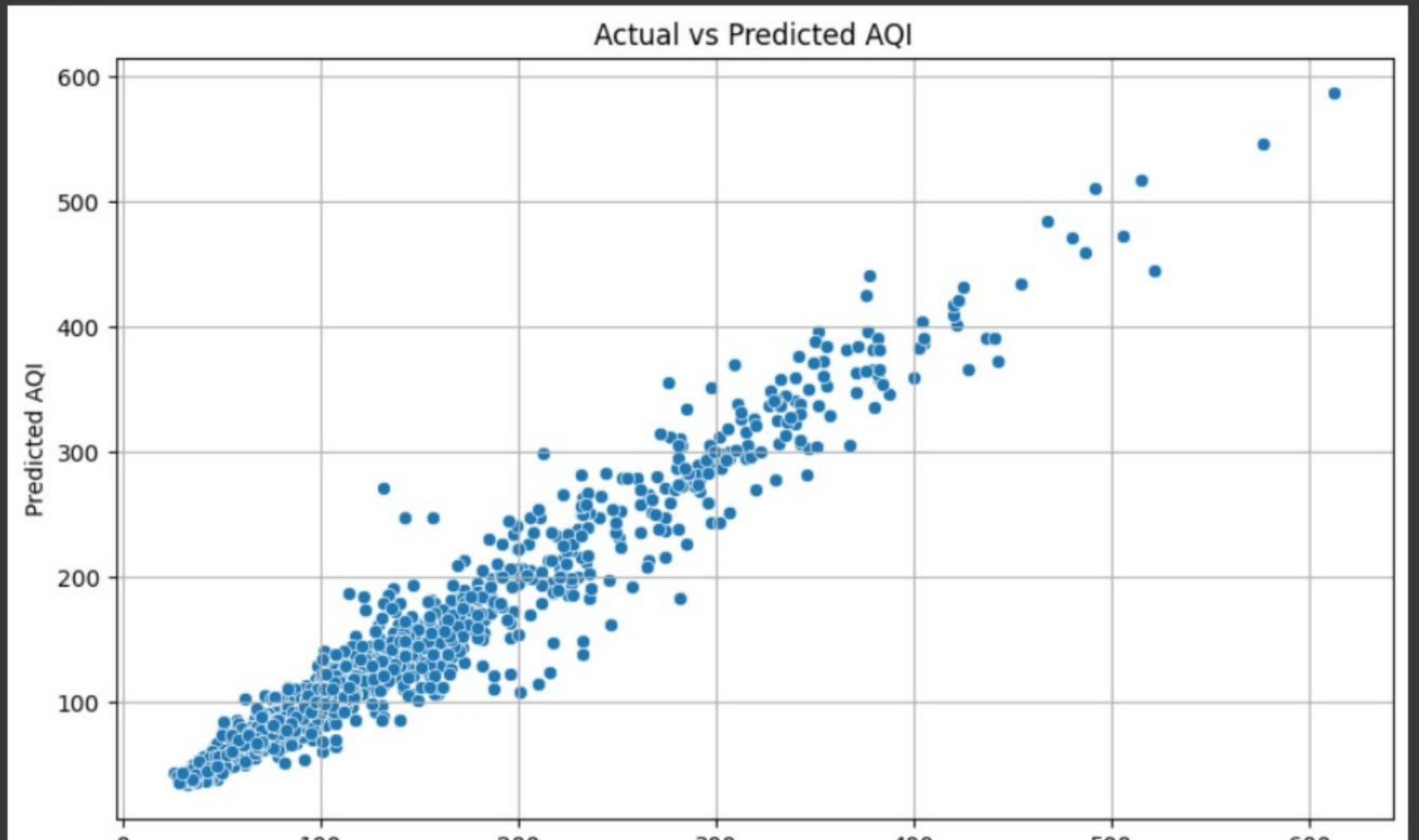
08

Visualization of Results

Choose Files archive (2).zip

- **archive (2).zip**(application/x-zip-compressed) - 76469579 bytes, last modified: 5/27/2025 - 100% done

```
Saving archive (2).zip to archive (2) (1).zip
MSE: 429.13
R²: 0.95
```



Regional AQI Visualization

1 Mapped AQI values across cities to identify regional pollution hotspots.

2 This helps in visualizing the distribution of air quality on a quality on a geographical scale.

Regional AQI Visualization

Tools Used

•Utilized visualization libraries such as matplotlib, seaborn, seaborn, or plotly.

Visualization Comparison

•Created a choropleth or bubble map of India highlighting highlighting the AQI levels across various locations.

09

Challenges & Improvements

Limitations & Future Work

Addressed issues regarding data quality and missing value entries which can which can skew results.

Noted city-wise imbalanced data that could affect model accuracy.

Suggested incorporating additional weather data to enhance model performance.

Future enhancements could include the use of deep learning techniques such techniques such as Long Short-Term Memory (LSTM) networks for time series time series forecasting.

10

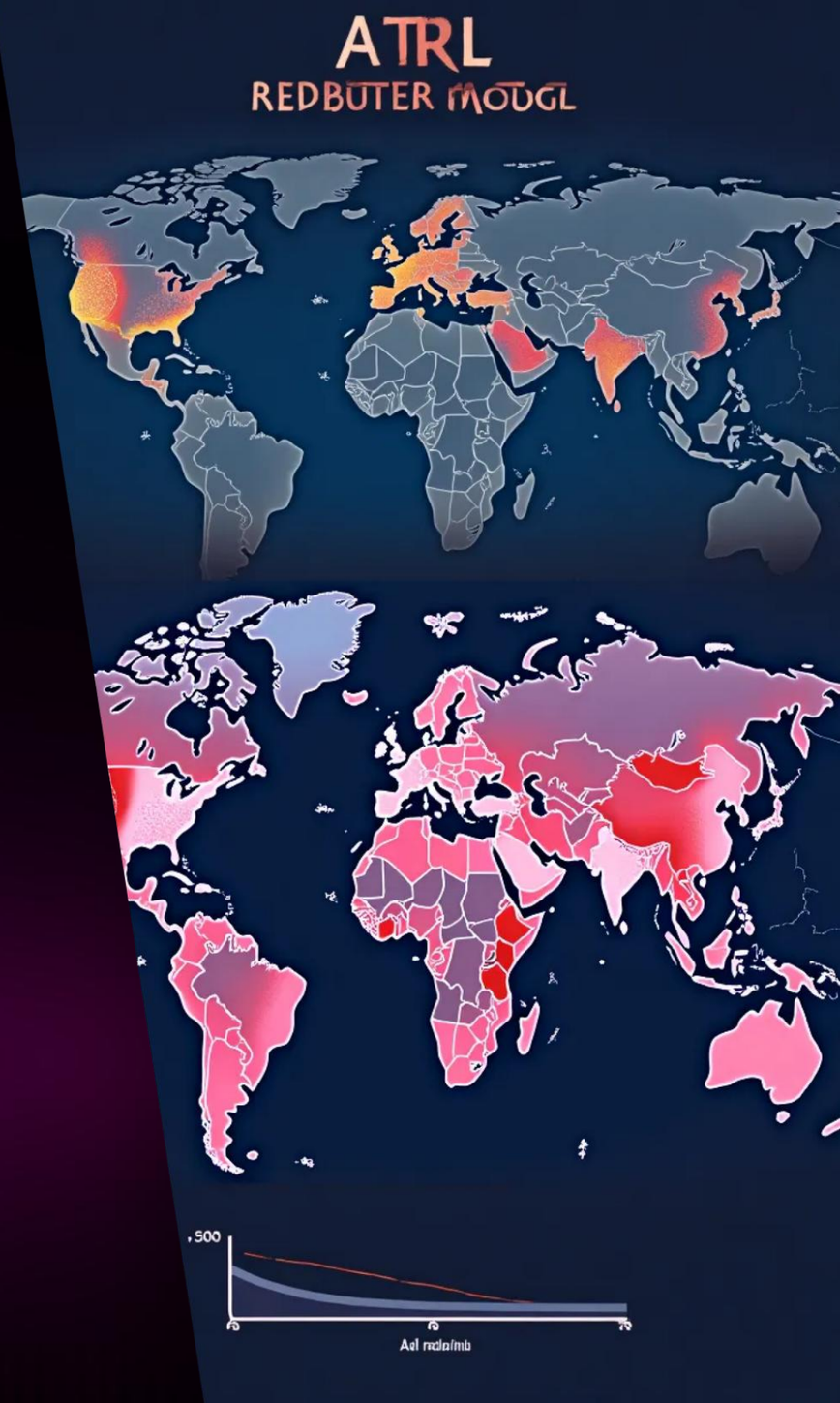
Conclusion & Q/A

Conclusion

The Random Forest Regression (RFR) model effectively predicted AQI levels vs actual aqi level

Notably, pollution levels exhibit considerable variation across different regions, which is critical for policy-making.

The predictive models developed can serve as valuable tools in guiding pollution control strategies and public health initiatives.



Thank You