Infosys Springboard Virtual Internship 6.0 Completion Report

Batch Number-

Start date - oct 13

Names:

Internship Duration: 8 Weeks

1.Project Title

AIR QUALITY INDEX PREDICTION

2. Project Objective

The main objective of this project is to develop a Machine Learning–based Air Quality Prediction system capable of forecasting the Air Quality Index (AQI) using various pollutant parameters such as PM2.5, PM10, NO_2, SO_2, CO, and O_3. The project aims to analyze real-time and historical air quality data, build an accurate regression model, and classify the predicted AQI into meaningful categories (Good, Satisfactory, Moderate, Poor, Very Poor, Severe) for easy understanding.

3. Project description in detail

The project involves developing an end-to-end Machine Learning system to predict the Air Quality Index (AQI) based on major atmospheric pollutants. The workflow begins with collecting a data-set containing pollutant concentrations such as PM2.5, PM10, NO_2, SO_2, CO, and O_3. To make the system interactive and user-friendly, a Gradio-based web interface was developed. The interface allows users to select a city, enter pollutant levels, and obtain the predicted AQI. The model also classifies the AQI value into standard air quality categories—Good, Satisfactory, Moderate, Poor, Very Poor, and Severe—based on CPCB guidelines. This provides meaningful insight to users about the severity of pollution.

The main focus of the project was accuracy, user experience, and practical relevance, helping users make informed decisions related to environmental awareness and health safety.

4. Timeline Overview

Week	Activities Planned	Activities Completed
Week 1	Understanding the problem statement, collecting the dataset, developing environment	Dataset collected, environment configured.
Week 2	Data processing for model training.	Completed preprocessing, correlation analysis.
Week 3	Implementing and testing ML regression models.	Trained multiple models and finalized best-performing model.
Week 4	Model tuning and saving the final model using	improved accuracy, saved the trained model.

	Joblib for deployment.	
Week 5	Designing a user interface using Gradio for AQI prediction.	Developed Gradio UI with pollutant inputs.
Week 6	Adding AQI category classification and city selection dropdown.	Successfully integrated AQI categories.
Week 7	Testing the complete system, fixing UI/logic issues.	Conducted system testing, improved UI responsiveness.
Week 8	Final documentation, report preparation.	Completed project report, documentation.

5a. Key Milestones

Milestone	Description	Date Achieved
Project Kickoff	Initial planning and brief explanation on entire project	Oct 19
Prototype/First Draft	Cleaning dataset, preparing features.	Oct 28
Mid-Term Review	Integrated knowledge base and feedback system	Nov 10
Final Submission	Completed UI deployment	Nov 22
Presentation	Final demo to mentor and peers	Dec 2

5b. Project execution details

The project was carried out in multiple stages. First, the air quality dataset was collected and cleaned, followed by preprocessing and selecting key pollutant features. Various machine learning regression models were trained and evaluated, and the best-performing model was chosen for prediction.

Next, a Gradio-based user interface was developed, allowing users to enter pollutant values and select a city. AQI prediction and category classification were integrated into the interface. Finally, the system was tested, refined, and documented for final submission.

6. Snapshots / Screenshots

In [1]:
```
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler
from sklearn.ensemble import RandomForestRegressor, GradientBoostingRegressor, ExtraTreesRegressor, StackingRegressor
from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_squared_error
```

In [2]:
```
from google.colab import files
uploaded = files.upload()
```

Upload widget is only available when the cell has been executed in the current browser session. Please rerun this cell to enable.
Saving AQI-and-Lat-Long-of-Countries.csv to AQI-and-Lat-Long-of-Countries.csv

In [3]:
```
import pandas as pd

data = pd.read_csv("AQI-and-Lat-Long-of-Countries.csv")
data.head()
```

Out[3]:

	AQI Value	CO AQI Value	Ozone AQI Value	NO2 AQI Value	PM2.5 AQI Value	lat	lng
0	51	1	36	0	51	44.7444	44.2031
1	41	1	5	1	41	-5.2900	-44.4900
2	41	1	5	1	41	-11.2958	-41.9869
3	66	1	39	2	66	37.1667	15.1833
4	34	1	34	0	20	53.0167	20.8833

In [4]:
```
X = data.drop(["AQI Value"], axis=1)
y = data["AQI Value"]

X.head(), y.head()
```

Out[4]:
```
(   CO AQI Value  Ozone AQI Value  NO2 AQI Value  PM2.5 AQI Value      lat  \
 0             1               36              0               51  44.7444
 1             1                5              1               41  -5.2900

         lng
0  44.2031
1 -44.4900
2 -41.9869
3  15.1833
4  20.8833  ,
0    51
1    41
2    41
3    66
4    34
Name: AQI Value, dtype: int64)
```

In [5]:
```
from sklearn.model_selection import train_test_split

# Step 4: Train-test split
X_train, X_test, y_train, y_test = train_test_split(
    X, y, test_size=0.2, random_state=42
)

X_train.shape, X_test.shape
```

Out[5]: ((13356, 6), (3339, 6))

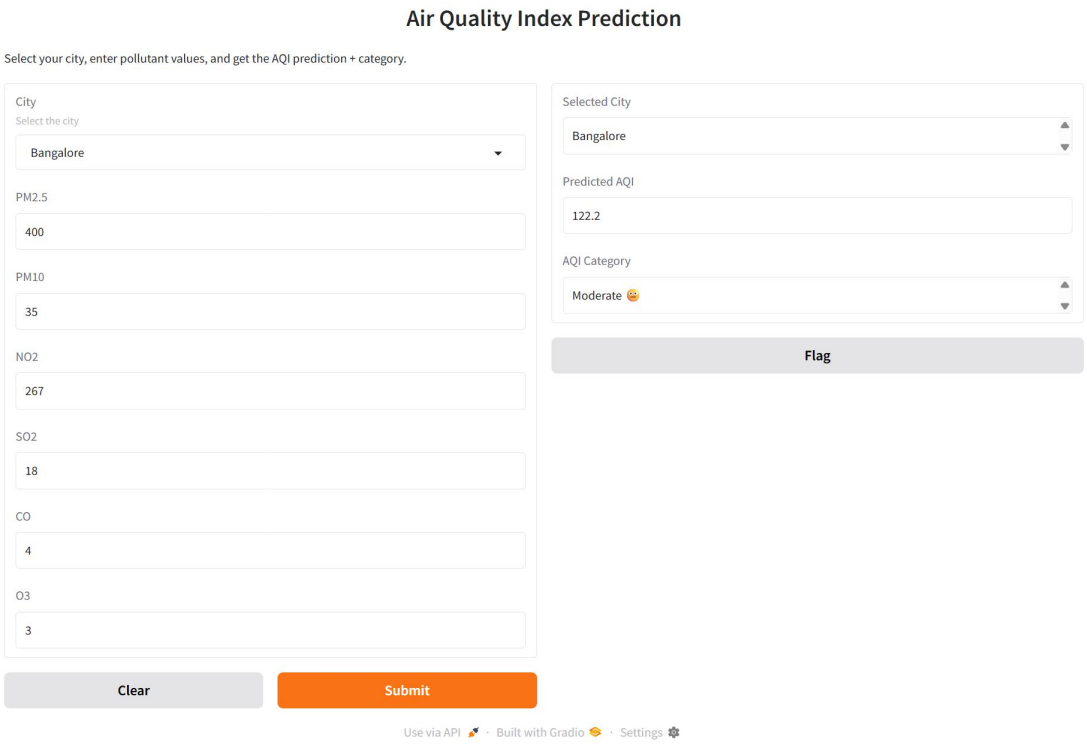
In [6]:
```
from sklearn.ensemble import RandomForestRegressor, GradientBoostingRegressor
from sklearn.linear_model import LinearRegression

# Step 5: Train base models
model_rf = RandomForestRegressor(random_state=42)
model_gb = GradientBoostingRegressor(random_state=42)
model_lr = LinearRegression()

model_rf.fit(X_train, y_train)
model_gb.fit(X_train, y_train)
model_lr.fit(X_train, y_train)

print("Base models trained successfully!")
```

Base models trained successfully!



7. Challenges Faced

- Cleaning and preprocessing inconsistent air quality data.
- Selecting the most accurate regression model.
- Hyperparameter tuning to improve performance.
- Mapping predicted AQI values to correct categories.

8. Learnings & Skills Acquired

- Learned data cleaning and preprocessing techniques.
- Gained experience in training and evaluating ML models.
- Improved understanding of AQI and pollutant impacts.
- Built a Gradio interface and integrated it with the model.
- Enhanced skills in debugging, tuning, and deployment.

9. Testimonials from team

"The internship enhanced our practical skills in ML and pyhton."

"Building a Air quality index prediction was both challenging and rewarding."

"We gained confidence in deploying real-world ML applications."

10. Conclusion

This internship provided a valuable experience in designing, developing, and deploying an ML model on air quality index prediction. It enhanced both technical and soft skills, aligning well with the students' academic and career goals in ML and python.

11. Acknowledgements

We sincerely thank **Infosys Springboard**, our mentors, and the academic coordinators for their continuous guidance, support, and encouragement throughout this internship.