TELEMETRY-DRIVEN HAZARDOUS GAS PREDICTION IN ATMOSPHERIC MONITORING

Project Reference No.: *48S_MCA_0058*

College	*:*	*R.V. College of Engineering, Bengaluru*
Branch	*:*	*Master of Computer Applications*
Guide(s)	*:*	*Dr. Divya T.L.* *Dr. Andhe Dharani*
Student(s)	*:*	*Mr. Dheeraj S* *Mr. Chandrashekar P G*

Keywords:

Environmental Telemetry, Air Quality Prediction, Random Forest & XGBoost, Ensemble Learning

Introduction/Background (15-20 lines):

This project is designed to contribute to global efforts in mitigating the adverse effects of air pollution by leveraging environmental sensor telemetry data for the real-time prediction of hazardous gas concentrations. In alignment with the United Nations Sustainable Development Goal (SDG) 3 on Good Health and Well-Being, and SDG 11 on Sustainable Cities and Communities, this initiative addresses the growing concerns of air quality degradation and its direct implications on public health and the environment.

The system integrates telemetry data capturing critical air pollutants, such as carbon monoxide (CO), sulfur dioxide (SO_2), nitrogen dioxide (NO_2), ozone (O_3), particulate matter $PM_{2.5}$, and PM_{10}. Using advanced machine learning models, including Random Forest, XGBoost, and an ensemble Voting Classifier, the project seeks to accurately classify and predict pollutant levels. These models, enhanced by ensemble learning techniques, are designed to improve prediction reliability, offering a robust tool for early warning and preventive action.

The project is in alignment with the World Health Organization (WHO) guidelines on air quality and aims to provide a data-driven framework for proactive environmental monitoring. This framework supports decision-makers and environmental agencies in managing air quality and mitigating risks associated with exposure to hazardous air pollutants.

Objectives:

- To utilize advanced preprocessing techniques such as Box-Cox transformations, IQR-based outlier treatment, and StandardScaler to normalize and prepare the dataset
- To handle class imbalance using SMOTE and Cluster Centroids to improve model generalization, ensuring reliable predictions across both majority and minority classes
- To build and evaluate Random Forest and XGBoost models, fine-tuned with GridSearchCV, and combined them in a soft-voting ensemble to enhance predictive performance
- To optimize classification thresholds based on F1-score to balance precision and recall, thereby improving the model's ability to detect hazardous air quality levels
- To deploy the trained ensemble model using Joblib, ensuring readiness for real-time prediction scenarios

- To develop a mobile application for seamless and widespread user accessibility to the algorithm

Methodology:

Phase 1: Data Preparation & Feature Engineering

- Collect daily air quality data (Dec 2018 – Nov 2024) from the World Air Quality Index (Jayanagar 5th Block, Bengaluru).
- Select six key pollutants: CO, SO_2, NO_2, O_3, $PM_{2.5}$, PM_{10}.
- Handle missing values using mean imputation (SimpleImputer).
- Apply Box-Cox transformation for normalization and IQR clipping for outlier treatment.
- Standardize features using StandardScaler.

- Conduct EDA via heatmaps, boxplots, and histograms; analyzed correlations (e.g., $PM_{2.5}$–PM_{10}).
- Engineer features (pollutant ratios) and addressed class imbalance using SMOTE and ClusterCentroids.

Phase 2: Model Development & Evaluation

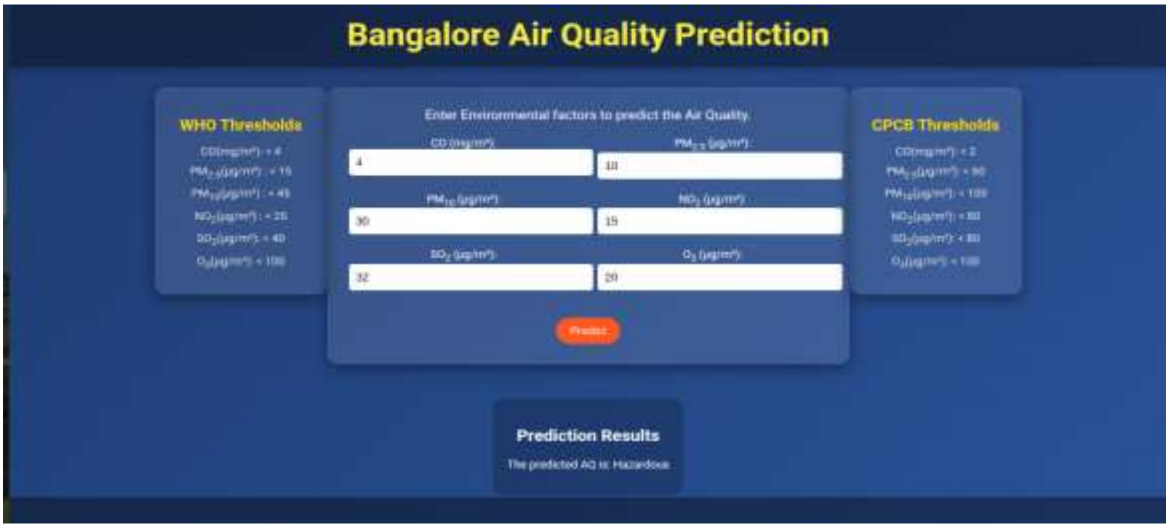
- Train Random Forest and XGBoost classifiers.
- Tune XGBoost using GridSearchCV; combined models via soft voting (VotingClassifier).
- Optimize classification threshold (~0.64) based on F1-score.
- Evaluate using Accuracy, F1-score, ROC-AUC, and Confusion Matrix.
- Validate using Stratified K-Fold Cross-Validation.
- Assess feature importance for model interpretability.

Phase 3: Deployment

- Save the trained ensemble model using Joblib.
- Deploy model via FastAPI for real-time predictions.
- Develop a mobile application for user accessibility to the algorithm

Results & Conclusions (15-20 lines):

- The machine learning pipeline successfully classified air quality into *Safe* or *Hazardous* based on real-world telemetry data from six pollutants (CO, SO_2, NO_2, O_3, $PM_{2.5}$, PM_{10}).



Bangalore Air Quality Prediction

WHO Thresholds

$CO(mg/m^3)$: < 4

$PM_{2.5}(\mu g/m^3)$: < 15

$PM_{10}(\mu g/m^3)$: < 45

$NO_2(\mu g/m^3)$: < 25

$SO_2(\mu g/m^3)$: < 40

$O_3(\mu g/m^3)$: < 100

Enter Environmental factors to predict the Air Quality

$CO\ (mg/m^3)$: 4

$PM_{2.5}\ (\mu g/m^3)$: 10

$PM_{10}\ (\mu g/m^3)$: 30

$NO_2\ (\mu g/m^3)$: 15

$SO_2\ (\mu g/m^3)$: 32

$O_3\ (\mu g/m^3)$: 20

CPCB Thresholds

$CO(mg/m^3)$: < 2

$PM_{2.5}(\mu g/m^3)$: < 60

$PM_{10}(\mu g/m^3)$: < 100

$NO_2(\mu g/m^3)$: < 80

$SO_2(\mu g/m^3)$: < 80

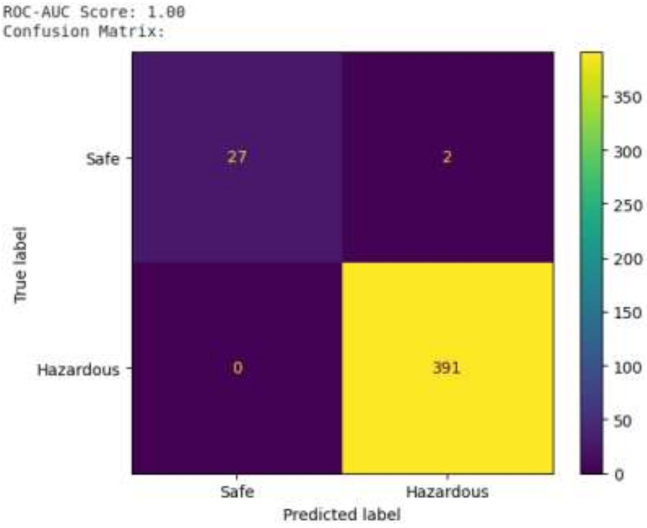
$O_3(\mu g/m^3)$: < 100

Prediction Results

The predicted AQ is: Hazardous

Fig 1: Successful Classification of Air Quality Level

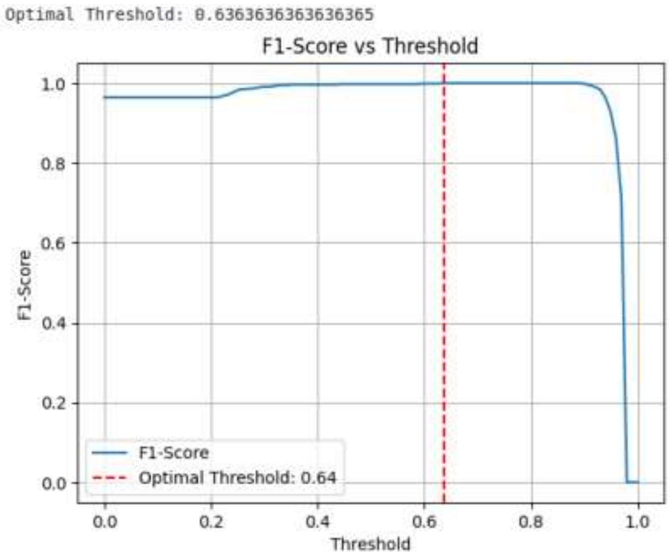
- Data preprocessing significantly improved model performance by addressing skewness (via Box-Cox), handling missing values, and removing outliers using IQR.
- Feature engineering revealed that $PM_{2.5}$, PM_{10}, and NO_2 were the most influential pollutants in determining hazardous air quality conditions.
- Class imbalance was effectively addressed using SMOTE and ClusterCentroids, ensuring better representation of minority (hazardous) cases during training.
- Ensemble learning (Voting Classifier combining Random Forest and XGBoost) achieved 99.5% accuracy and a ROC-AUC score of 1.00.



```
ROC-AUC Score: 1.00
Confusion Matrix:
```

Fig 2: Confusion Matrix and ROC-AUC Curve of Ensemble Model

- Threshold optimization (at ~0.64) based on F1-score helped achieve a balance between sensitivity and precision in hazard classification.



```
Optimal Threshold: 0.6363636363636365
```

Fig 3: Optimal Threshold Computation

- The model supports real-time prediction, aligns with WHO standards, and contributes to SDG Goals 3 and 11.

Project Outcome & Industry Relevance:

- **Environmental Health:** It helps predict harmful pollutant levels, aiding in public health interventions and policies.
- **Air Quality Monitoring:** Can be integrated into systems for real-time pollution forecasts.
- **Public Health:** Enables authorities to issue health warnings to vulnerable populations.
- **Smart Cities:** Supports dynamic traffic and emission control to improve urban air quality.
- **Industries:** Assists industries in monitoring and complying with environmental regulations.

Working Model vs. Simulation/Study:

The project is primarily a simulation. It involves the development of a software-based system, utilizing a web frontend and a FastAPI backend, to predict hazardous air pollutant levels using machine learning algorithms, without the use of physical hardware or real-time sensors.

Project Outcomes and Learnings:

Key Outcomes:

- Accurate Pollutant Prediction: Successfully predicted hazardous air pollutant levels using machine learning.
- Web Application: Developed a functional web app with frontend (HTML, CSS, JS) and backend (FastAPI).
- Decision Support: Provided insights for public health, regulatory authorities, and industries.

Key Learnings:

- Model Tuning: Gained experience in selecting and fine-tuning machine learning models.
- Full-Stack Integration: Learned how to integrate frontend and backend for real-time predictions.
- Data Handling: Improved skills in preprocessing and managing real-world data.
- Web Deployment: Developed expertise in web deployment and user interface design.

Future Scope:

- **Real-Time Integration**: Connect with live air quality sensors for dynamic predictions.
- **Mobile App**: Expand to a mobile app for wider accessibility.
- **Geospatial Analysis**: Include location-specific predictions and pollution hotspots.
- **Advanced Models**: Explore deep learning for improved accuracy.

- **Predictive Alerts**: Develop an automated notification system based on air quality forecasts.