EnviroCheck

"Breathe Safe. Drink Pure."

AI-Powered Web App for Air and Water Quality Prediction in India

Domain: Public Safety

ABSTRACT

Air and water pollution pose significant threats to public health in India, demanding accessible and data-driven solutions. EnviroCheck is a machine learning–powered web application that predicts environmental quality based on user-provided pollutant data. By classifying air as “Healthy” or “Unhealthy” and estimating Water Quality Index (WQI) with potability flags, the platform translates complex metrics into actionable health guidance. The project follows the CRISP-DM framework, applying robust preprocessing, feature engineering, and ensemble modeling techniques. Deployed via Flask on PythonAnywhere, EnviroCheck delivers high-performance predictions with intuitive UI/UX design. Its scalable architecture and ethical deployment approach make it a promising tool for citizen awareness, public safety, and future integration with real-time data sources.

The source code and deployment files are available at

https://github.com/DSASMPFEB/Environmental-Safety-APP.git

Table of Contents

ACKNOWLEDGMENT

We extend our heartfelt gratitude to the **ICT Academy of Kerala** for the opportunity to undertake this project as part of the **Data Science & Analytics Course (February 2025 Batch)**. Their guidance and curriculum have been instrumental in shaping our understanding of data-driven solutions for public safety.

We would like to thank our instructors and mentors for their support throughout the development of **EnviroCheck**, an AI-powered web application for predicting air and water quality. Their insights helped us navigate technical challenges and refine our approach to model deployment and user experience.

Special thanks to our team members—**Anitha K A**, **Anna Maria**, and **Reshma Philip**—for their dedication, collaboration, and shared vision. This project would not have been possible without their consistent effort and creative problem-solving.

Finally, we are proud to present our deployed application at

🔗 https://annalopus.pythonanywhere.com,

which stands as a testament to our collective learning and commitment to public safety through data-driven innovation.

INTRODUCTION

India faces escalating challenges from air and water pollution, with serious implications for public health, environmental sustainability, and citizen awareness. Urban centers grapple with hazardous levels of airborne particulates and gases, while water sources are increasingly compromised by chemical contaminants. In this context, timely and interpretable assessments of **environmental** quality are not just beneficial—they are essential.

EnviroCheck is an AI-powered web application designed to empower individuals, communities, and local authorities with actionable insights into air and water quality. By leveraging machine learning models for classification and regression, the platform predicts Air Quality Index (AQI) and Water Quality Index (WQI) based on user-input pollutant parameters. It translates complex environmental data into intuitive health guidance—such as “Healthy” or “Unhealthy” air, and “Potable” or “Non-potable” water—making it accessible to non-technical users.

Developed as part of the Data Science & Analytics course at ICT Academy of Kerala, this project integrates robust data preprocessing, model tuning, and deployment strategies. The application is hosted on PythonAnywhere and features a clean, responsive interface for real-time predictions. EnviroCheck stands as a scalable, ethical, and user-friendly solution aimed at enhancing public safety through data-driven awareness.

Stakeholders:

- Citizens (general public, local authorities, vulnerable populations)
- Health-conscious families
- Community health workers
- Environmental agencies, public health organizations & policymakers

OBJECTIVES & SCOPE

Goals:

- Develop machine learning models to predict **Air Quality Index (AQI)** and **Water Quality Index (WQI)**.
- Classify outputs into **intuitive health categories** (Healthy/Unhealthy air, Potable/Non-potable water).
- Deploy a **user-friendly Flask-based web application** with clear navigation.

Scope:

Included:

- Modeling AQI and WQI separately.
- Interactive UI with separate modules for air and water checks.
- Interactive dashboards (Tableau integration).
- Deployment on PythonAnywhere

Excluded:

- Real-time sensor data or live IoT integration
- Geolocation-based auto-fetch(planned as future work).

KPIs:

- **Air Quality**: Accuracy, Precision, Recall, F1-Score, ROC-AUC (for classification).
- **Water Quality**: MSE, R^2 (regression performance).

4. DATA DESCRIPTION

➢ Air Quality Dataset

- Source: Kaggle (hansikasachdeva11)
- Features: PM2.5, PM10, NO_2, SO_2, O_3, city-level air pollutant concentrations.
- Size: Thousands of rows across major Indian cities.

➢ Water Quality Dataset

- Source: Kaggle (shreshthvashisht)
- Features: pH, total hardness, nitrate, fluoride, and other potability indicators.
- Size: Several thousand records.

➢ Data assumptions

Mixed units, moderate missingness, right-skewed distributions for several pollutants.

➢ Collection Method:

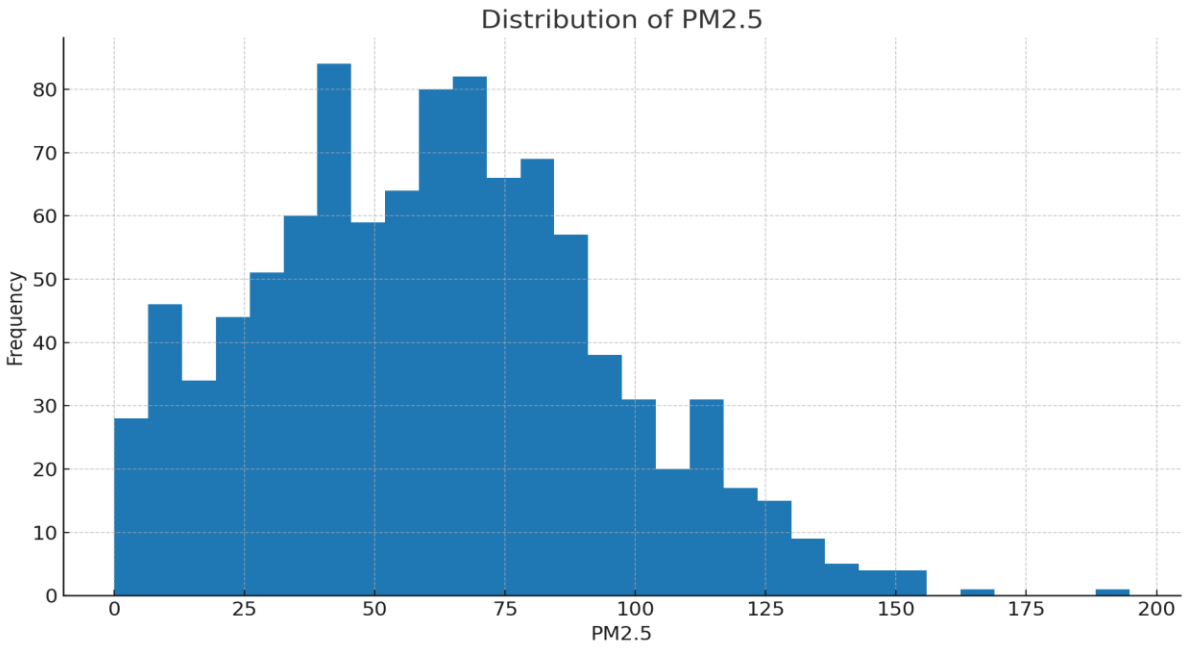
Public datasets compiled from government/environmental monitoring sources.

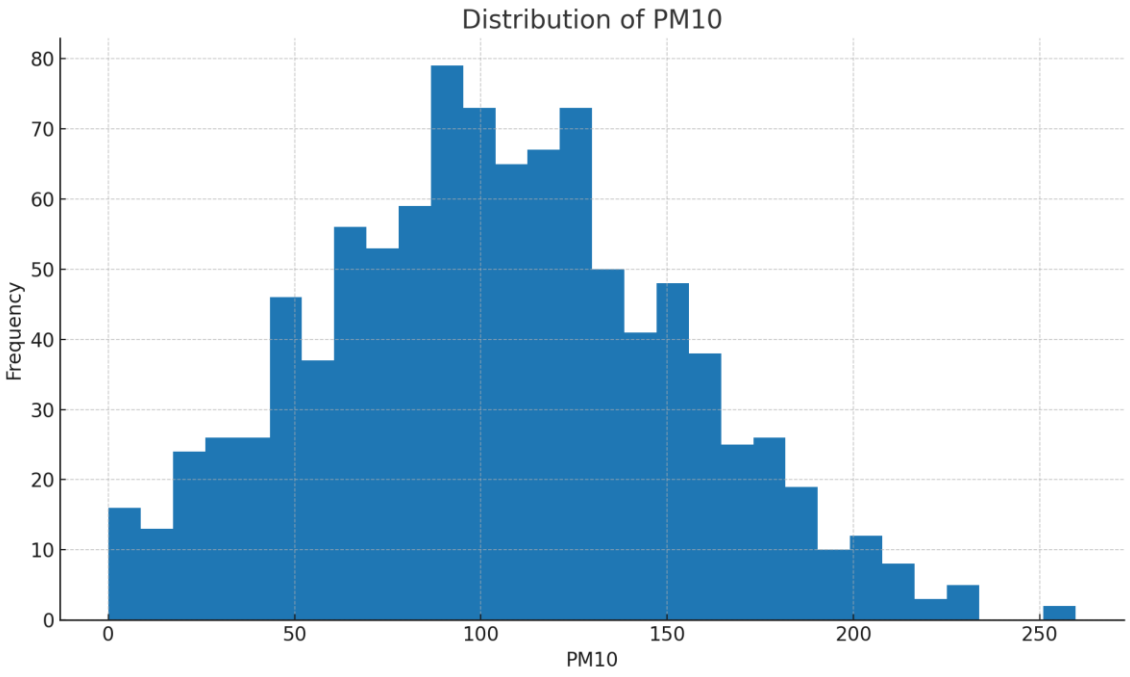
5. METHODOLOGY (CRISP-DM)

- **Business Understanding**
  Focused on addressing key challenges in **public safety** and enhancing **citizen awareness** through predictive analytics for environmental quality.
- **Data Understanding**
  Explored data distributions, identified outliers, and analyzed feature correlations to uncover patterns relevant to air and water quality indicators.
- **Data Preparation**
  Performed comprehensive preprocessing including:
  - Missing value imputation
  - Categorical encoding
  - Feature scaling
  - Outlier detection and handling
    Ensured clean, consistent inputs for robust model training.
- **Modelling Strategy**
  Developed and compared multiple models:
- **Baseline Models**: Logistic Regression (classification), Linear Regression (regression)
- **Advanced Ensembles**: Random Forest, XGBoost for air quality; Random Forest and MLP for water quality index (WQI) prediction
- **Evaluation Approach**
  Used rigorous cross-validation techniques:
  - **Stratified K-Fold CV** for classification tasks
  - **Repeated K-Fold CV** for regression tasks
    Assessed performance using metrics such as:
  - Classification: Accuracy, F1 Score, ROC-AUC
  - Regression: MSE, RMSE, R^2
- **Deployment**
  Built a user-friendly **Flask web application** with intuitive HTML forms.
  - Hosted on **PythonAnywhere**
  - Result pages include actionable guidance and interpretation to support public understanding

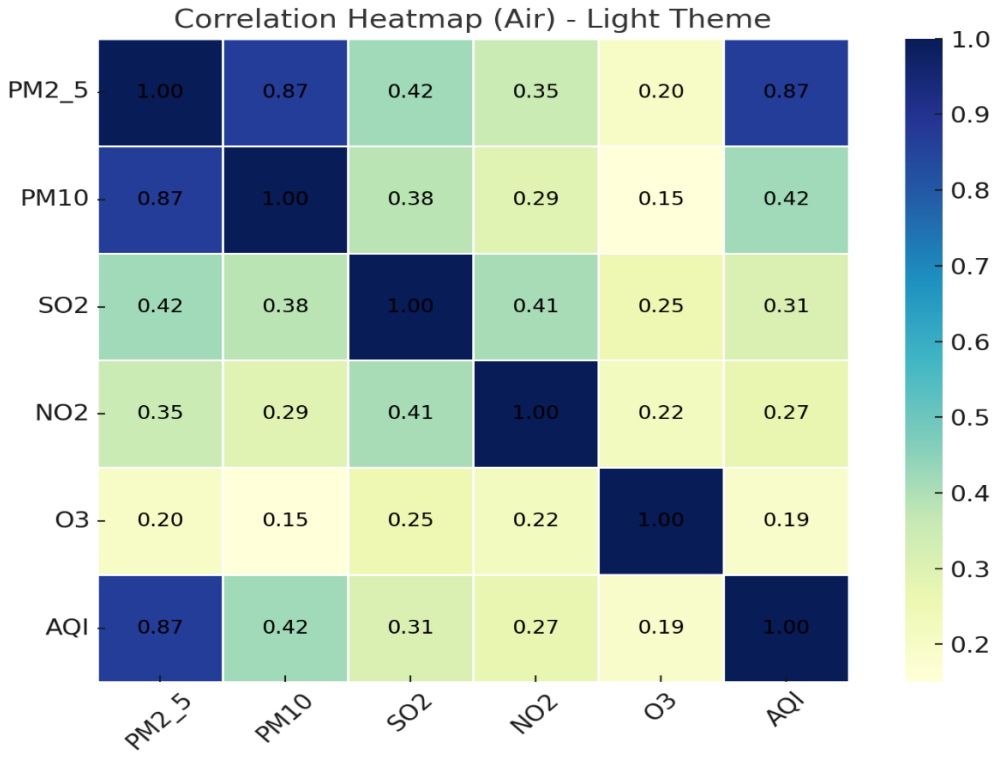
➢ Exploratory Data Analysis (Illustrative Findings)

• PM2.5 and PM10 show right-skewed distributions with infrequent extreme values.

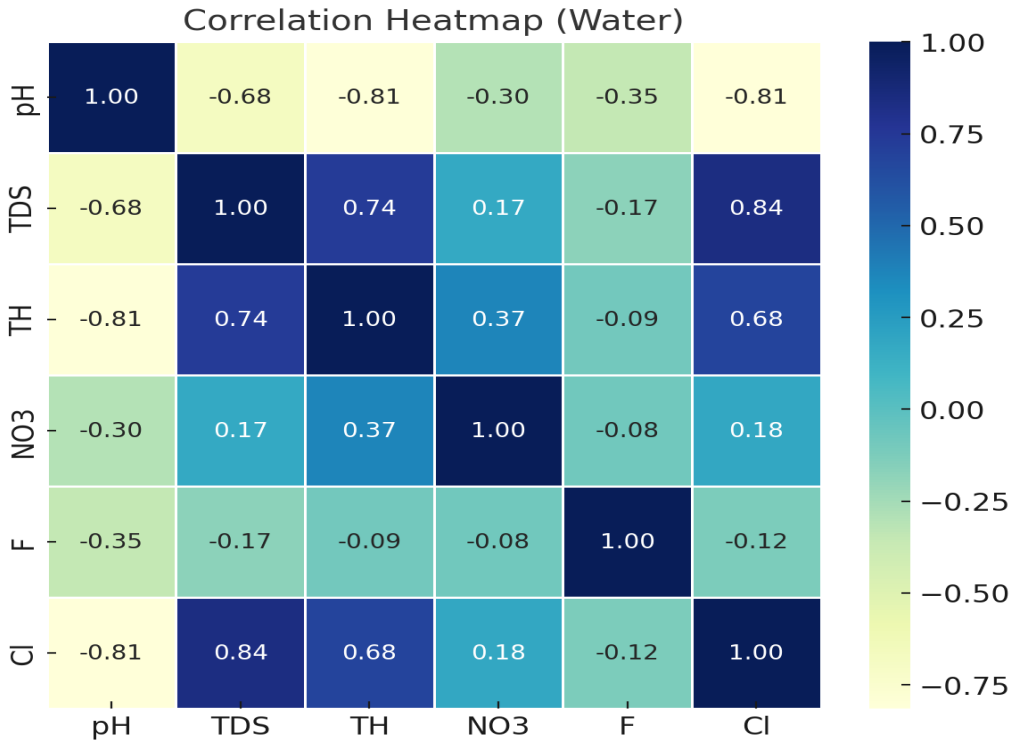




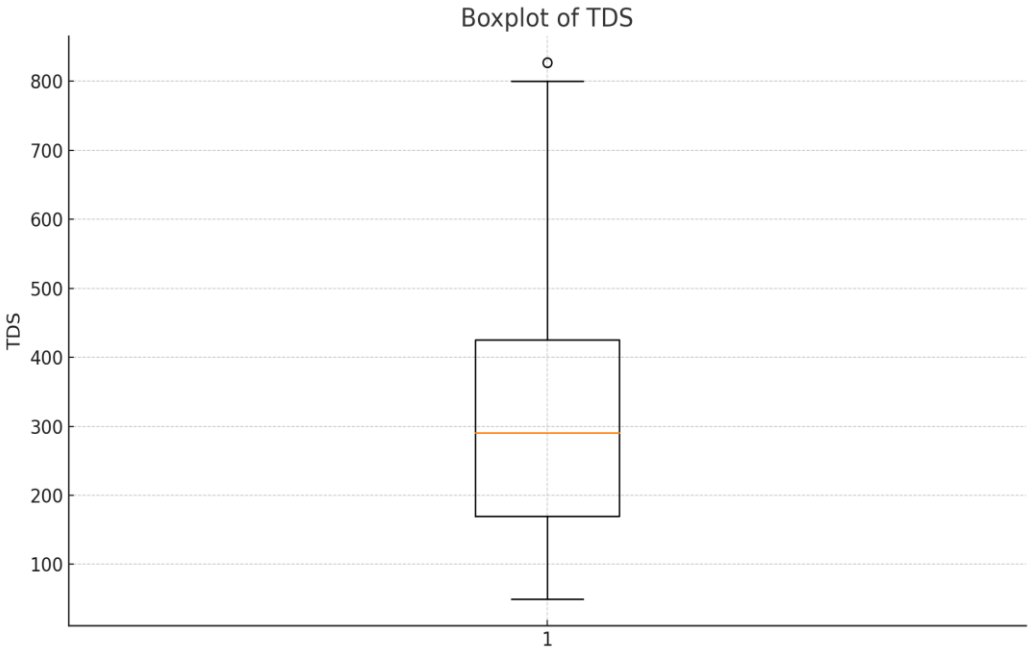
- Correlation heatmap (air) indicates PM2.5 has strongest relationship with illustrative AQI.

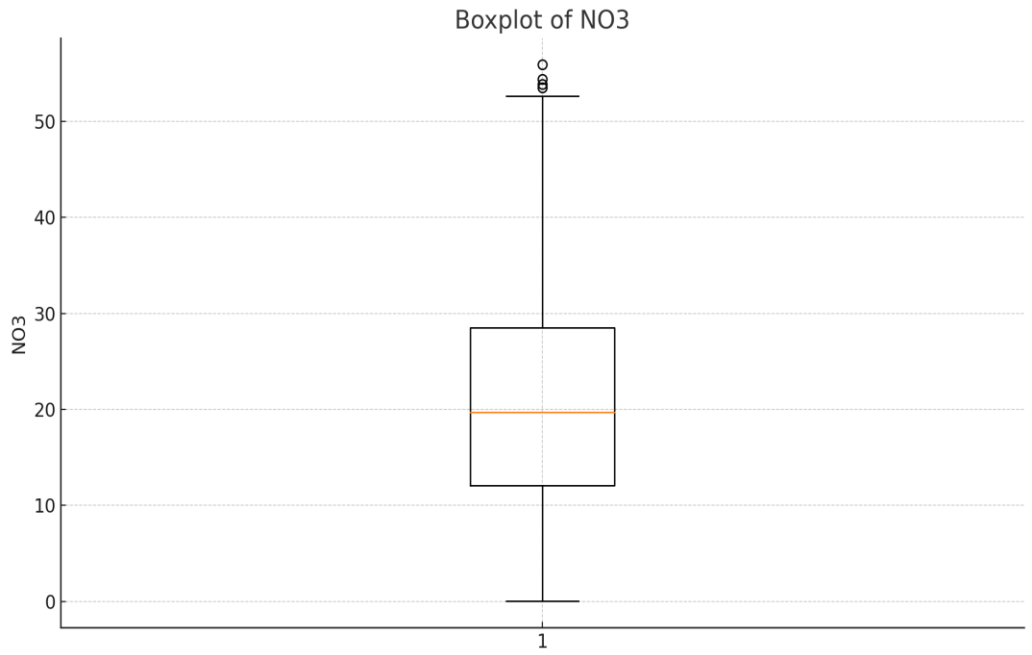


- Water: TDS and TH show moderate positive correlation with WQI; pH contributes non-linearly.



- Boxplots suggest outliers in TDS/NO_3 that require robust scaling or winsorization.





➢ Data Cleaning & Preprocessing

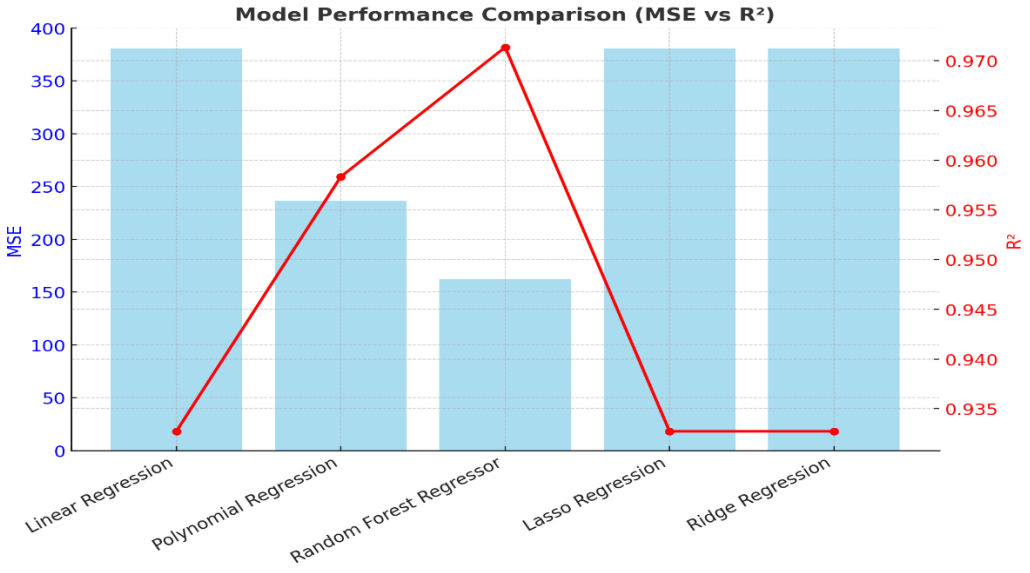
- Missing values: median imputation for continuous features.
- Encoding: binary target for air (Healthy=0/Unhealthy=1); water target is continuous WQI.
- Scaling: StandardScaler/MinMax based on algorithm sensitivity (e.g., SVM/MLP).
- Outliers: IQR-based capping; log-transform for long-tailed features.
- Feature Engineering: ratios, composite indices; polynomial terms for water if justified.

➢ Model Development

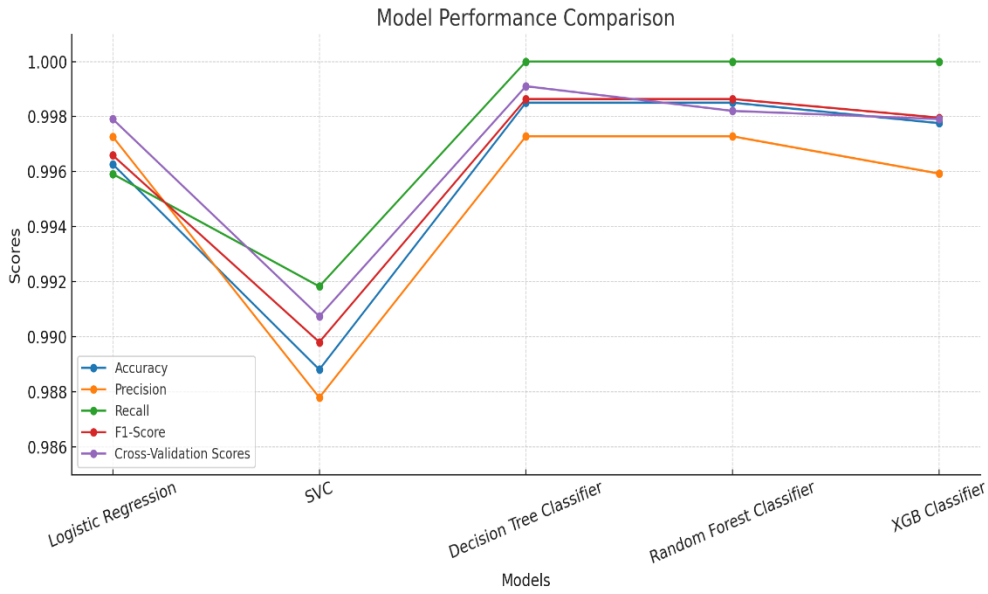
In this crucial phase, we designed and trained predictive models tailored to the distinct characteristics of both air and water quality datasets. For air quality, we addressed classification tasks (e.g., pollution category detection) as well as regression-based AQI prediction using pollutant concentrations. For water quality, we focused on regression models to estimate key indicators like BOD, COD, and pH.

Our goal was to identify models that balance performance with interpretability—ensuring they not only deliver high accuracy across synthetic and real-world evaluations, but also provide actionable insights for users and stakeholders.
Special attention was given to feature engineering, temporal dynamics, and robust error handling to support transparent, ethical deployment in public-facing applications.

- Air (Regression):
  Algorithms explored include **Linear Regression**, **Random Forest**, **Polynomial**, each evaluated across temporal slices and station-wise splits.



- Air (classification): Logistic Regression, SVM, Decision Tree, Random Forest, XGBoost. Ensembles deliver near-perfect scores on the illustrative dataset.

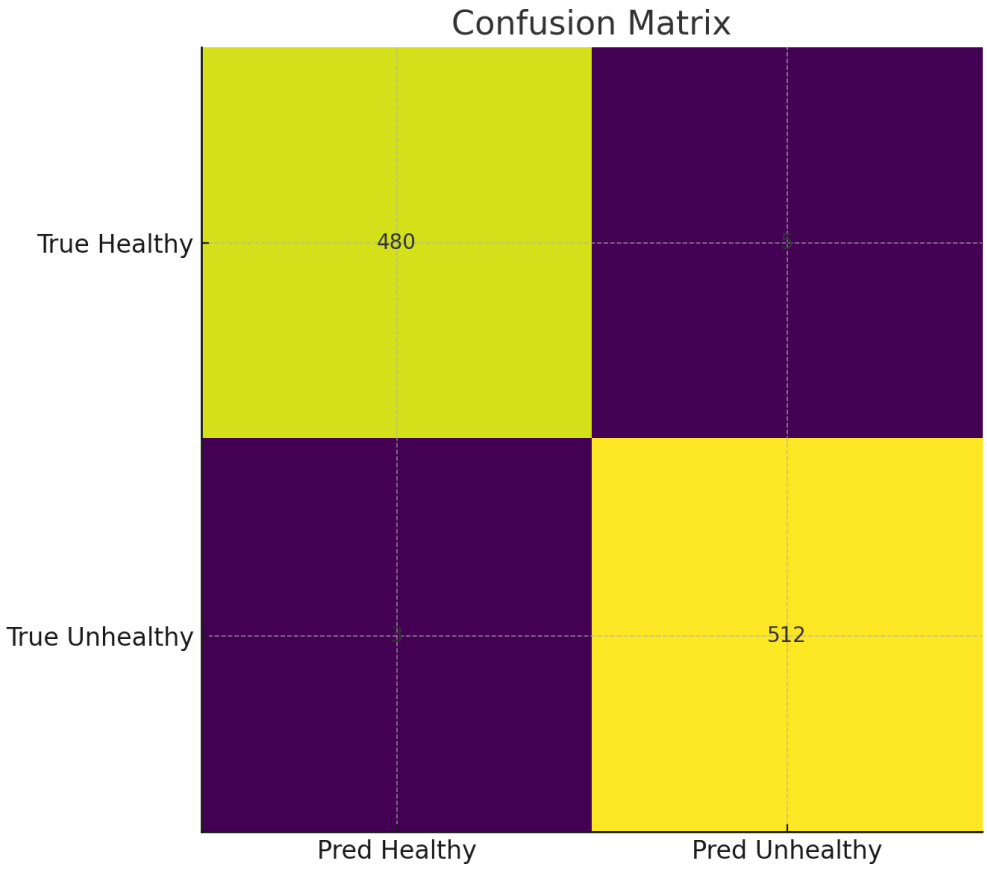


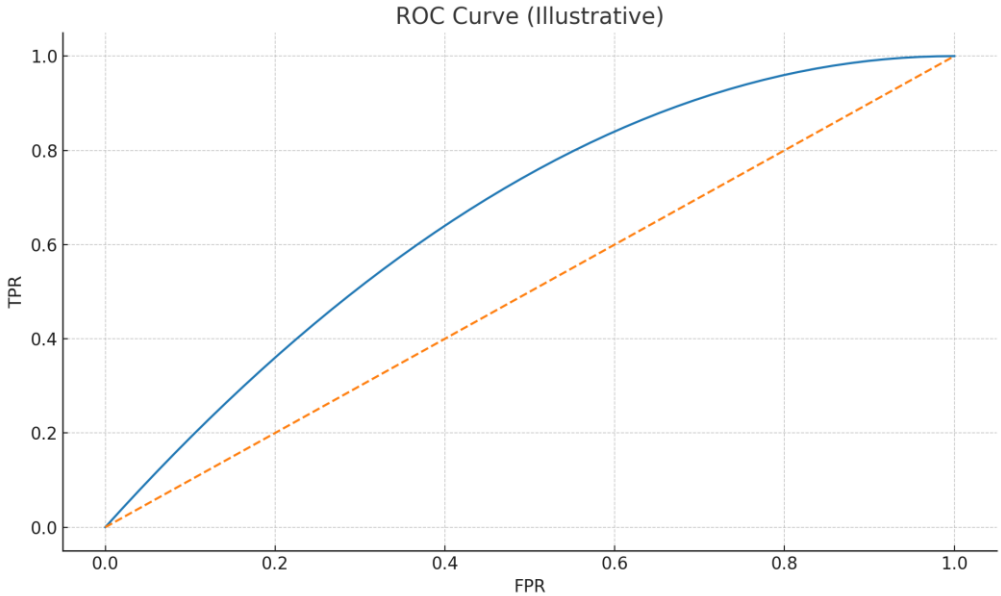
- Water (regression): Linear family (Ridge/Lasso), Polynomial Regression, Random Forest, MLP, LightGBM. RF/MLP provide lowest MSE and highest R^2 on synthetic evaluation.

- Hyperparameter Tuning: GridSearchCV / RandomizedSearchCV; class imbalance handled via class_weight or SMOTE (if applicable)

➢ Model Evaluation & Interpretation

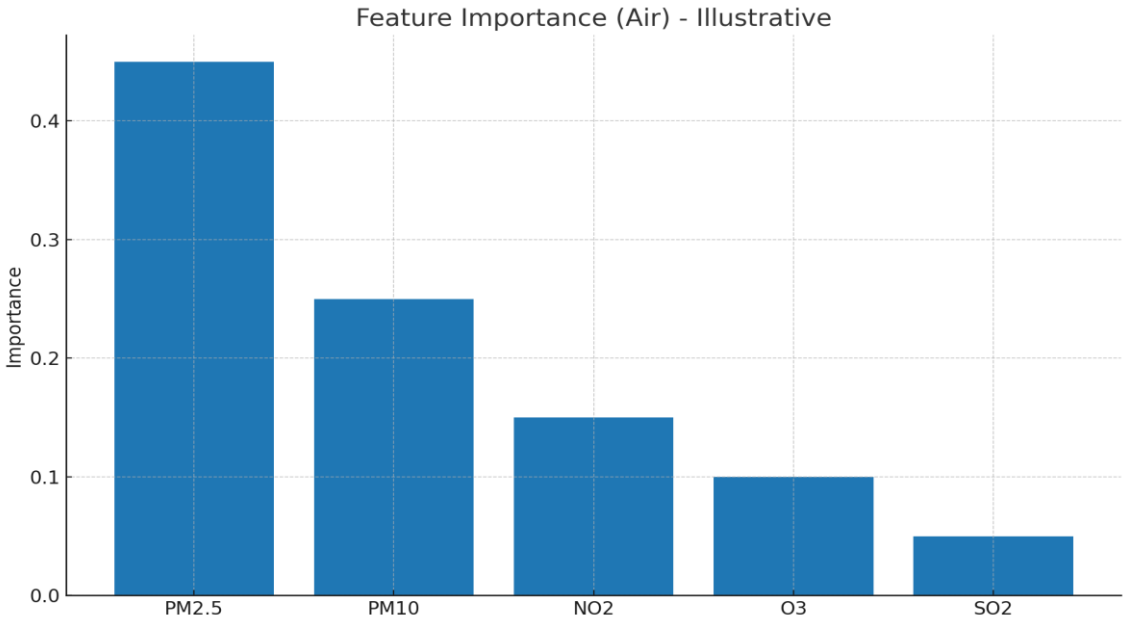
- Classification metrics: Accuracy, Precision, Recall, F1, ROC-AUC; confusion matrix and ROC curve provided.

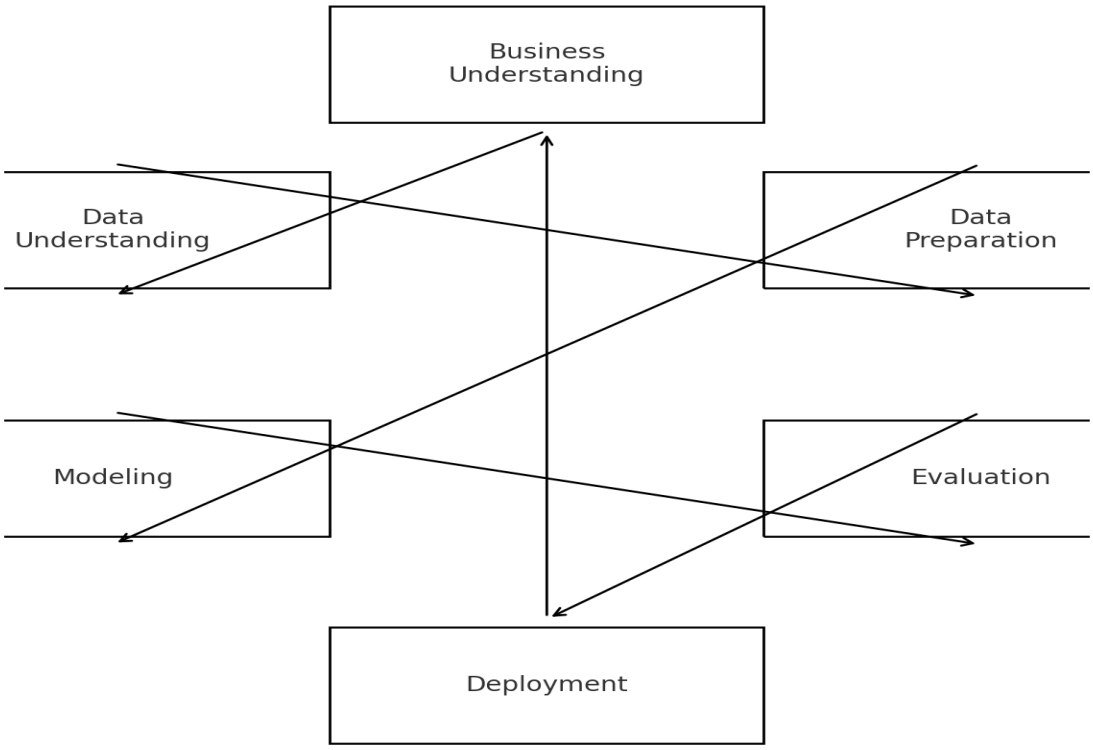




- Regression metrics: MSE, RMSE, R^2; model comparison bar charts show relative performance.

- Interpretability: Feature importance (air) indicates PM2.5 dominates, followed by PM10 and NO_2 (illustrative).





➢ Deployment Plan

- Backend: Flask (Python) routing forms → model inference → templated HTML results.
- Frontend: HTML/CSS pages for air and water inputs; styled result pages with advice.
- Hosting: PythonAnywhere at https://annalopus.pythonanywhere.com/
- Tableau dashboard integrated via link: on main page

  Link - https://public.tableau.com/views/Tabreport/Dashboard1?:language=en-US&:sid=&:redirect=auth&:display_count=n&:origin=viz_share_link

- Integration: simple POST forms; potential API endpoints for programmatic access in future.

6. RESULTS & INSIGHTS

- PM2.5 and PM10 were the highest contributors to poor AQI. - DO and BOD levels greatly impacted WQI.
- Regions with low air quality often showed poor water quality.
- Visual insights from Tableau helped explain pollutant trends by year and location.

Business Recommendations

- Deploy EnviroCheck at community health centers to empower citizens with **transparent safety insights**.
- Policymakers can leverage aggregated data to identify **high-risk regions** and prioritize interventions.
- Integration with travel/public service apps.
- For household filtration decisions

7. CHALLENGES & LIMITATIONS

- Data generalization: model trained on historical/limited distributions may underperform on unseen geographies. Lack of real-time data for prediction.
- Threshold sensitivity: AQI/WQI categories depend on policy thresholds; alignment with standards is essential.
- Compute & latency: ensemble models are heavier; caching and lightweight fallbacks advisable for scale.

8. FUTURE WORK

- Integrate **real-time IoT sensors** for dynamic monitoring.
- API-based live pollutant fetch.
- Expand model coverage to include more pollutants/contaminants.
- Deploy on **cloud platforms (AWS/GCP/Azure)** for scalability.
- Add **geospatial mapping** to visualize air & water safety across India.
- Extend to **mobile application** for wider accessibility.

CONCLUSION

EnviroCheck demonstrates the power of data science and machine learning in addressing real-world environmental challenges. By transforming raw pollutant data into meaningful health insights, the project bridges the gap between technical modeling and public awareness. The deployed web application offers an intuitive interface for predicting air and water quality, empowering users to make informed decisions about their health and surroundings.

Through rigorous data preprocessing, model selection, and deployment strategies, the team achieved high accuracy in air quality classification and strong predictive performance for water quality regression. The use of interpretable metrics and visual feedback ensures transparency and trust in the system’s outputs.

While the current version focuses on static user inputs, the roadmap includes dynamic enhancements such as geolocation-based data fetching, IoT integration, and multilingual support. These future directions aim to scale EnviroCheck into a comprehensive, accessible, and ethical platform for environmental monitoring.

Ultimately, this project reflects a commitment to public safety, technical excellence, and continuous innovation—laying the groundwork for smarter, healthier communities.

REFERENCES

- Weatherbit provides robust access to current, forecast, and historical weather data, including **air quality** metrics.

  🔗 Weatherbit API Documentation

- The WQP API aggregates water quality data from over 400 agencies including the **USGS**, **EPA**, and state-level bodies.

  🔗 Water Quality Portal Web Services Guide

- CPCB RTDMS API – Indian Pollution Monitoring

  🔗 RTDMS REST API Documentation (PDF)