Comparative Analysis of Machine Learning based Models for Air Quality Prediction

Pullabhotla Vijay, Sukanta Nayak
School of Computer Science and Engineering,
Department of Mathematics, School of Advanced Sciences,
VIT-AP University, Amaravati, Andhra Pradesh, 522241, India

Abstract

Air pollution is one of the major environmental concerns, and it has a significant impact on health. Accurate Air Quality Index (AQI) prediction is essential for effective policy decisions and public awareness. In this study, we compare various machine learning models, including XGBoost (as used in previous research), Random Forest, LightGBM, CatBoost, and TabTransformer, to determine the most effective model for AQI prediction in Gujarat, India. Our analysis reveals that while traditional models like XGBoost perform well, newer deep learning-based approaches such as TabTransformer show potential but require further optimization. The study compares the accuracy and computational efficiency of different ML models. The performance of models was evaluated using methods like RMSE, R^2, and MAE scores, and a detailed feature importance analysis is conducted.

Introduction

In rapidly industrializing regions like Gujarat, India, deteriorating air quality has been linked to increased respiratory diseases, cardiovascular conditions, and reduced life expectancy. Factors such as vehicular emissions, industrial activities, agricultural burning, and meteorological influences contribute to fluctuating Air Quality Index (AQI) levels.

This study builds upon prior research that primarily relied on Random Forest and XGBoost models for AQI prediction. While these models demonstrated high predictive accuracy, our research evaluates additional machine learning approaches including LightGBM, CatBoost, and TabTransformer. We aim to determine the most effective AQI prediction algorithm through comparative analysis and hyperparameter tuning. This study contributes to the development of data-driven environmental policies in Gujarat and similar regions.

Our research is particularly relevant for policymakers and environmental agencies looking to develop proactive measures against air pollution. By identifying the most effective AQI prediction models, authorities can optimize pollution control strategies, implement early warning systems, and improve public health interventions. This study advances machine learning applications in environmental science and contributes to data-driven policymaking aimed at mitigating air pollution in Gujarat and similar industrialized regions worldwide.

Methods and Materials

The dataset includes AQI measurements from multiple cities in Gujarat. Features include:
Pollutants: PM2.5, PM10, NOx, SO2, O3, SPM
Meteorological: MAX_TEMP
The data was sourced from CPCB and analyzed for correlations. PM10, PM2.5, NOx were found most influential.
Models Evaluated are XGBoost, Random Forest, LightGBM,CatBoost ,TabTransformer

- Research Question: Which machine learning model delivers the highest accuracy in predicting AQI for industrial regions like Gujarat?
- Background: Prior models like XGBoost and Random Forest have shown promise in AQI prediction but need comparison with newer approaches.
- Hypothesis: CatBoost and other boosting algorithms will outperform traditional and deep learning models in structured AQI datasets.
- Procedure: Collected AQI and meteorological data from CPCB, preprocessed it, and trained five different ML models.
- Case Study: Focused on major industrial cities in Gujarat such as Ahmedabad, Vadodara, and Ankleshwar.
- Data & Analysis: Analyzed pollutant levels (PM2.5, PM10, NOx) and temperature using EDA, followed by feature importance and residual analysis.
- Summary: CatBoost emerged as the most reliable and accurate model, followed by XGBoost and Random Forest.

Results

- This study evaluates five models for AQI prediction—XGBoost, Random Forest, LightGBM, CatBoost, and TabTransformer. CatBoost outperformed all models with an RMSE of 2.05 and an R^2 of 0.99, showcasing its superior handling of categorical features.
- XGBoost followed closely (RMSE: 2.19, R^2: 0.99), while Random Forest showed stable performance. LightGBM was efficient but had higher RMSE, indicating underfitting.
- TabTransformer performed the weakest due to limited data and tuning issues.
- Feature importance analysis revealed PM10, PM2.5, and NOx as the most influential predictors.
- These findings validate CatBoost as the most accurate model for structured AQI forecasting.

Fig. 1: comparison - RMSE

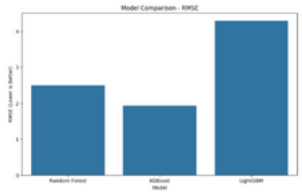


Fig 2: Light-GBM Feature Importance

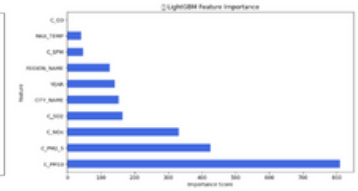


Fig 3: Residual Distribution

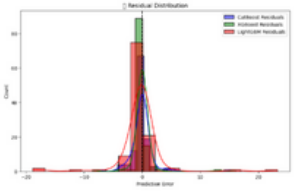


Chart 1. Actual Vs Predicted AQI After Fine Tuning.

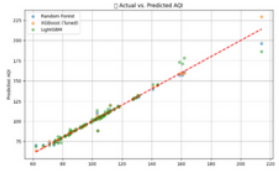
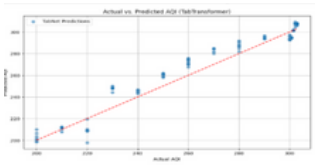


Table 1. Comparison Of Multiple ML Models

Model	RMSE ↓	R^2 Score ↑	MAE ↓
XGBoost	2.190	0.990	0.535
LightGBM	3.971	0.968	1.458
CatBoost	2.054	0.991	0.967
TabTransformer	9.051	0.934	2.342
Random Forest	2.542	0.986	0.701

Chart 2. Actual Vs Predicted AQI (Tab-Transformer Model)



Discussion

This study presents a comparative analysis of machine learning models for AQI prediction. Among all, CatBoost demonstrated the highest accuracy with RMSE = 2.05 and R^2 = 0.99, making it the most effective model. Its ability to handle categorical data without extensive preprocessing and its use of ordered boosting techniques provided an advantage in reducing overfitting and enhancing generalization. XGBoost followed closely with slightly lower accuracy but maintained strong interpretability and speed. LightGBM, although efficient, showed signs of underfitting on structured data. Deep learning-based TabTransformer struggled with performance due to data limitations and hyperparameter sensitivity.

Conclusions

The comparative results of this research establish CatBoost as the most accurate model for AQI prediction among the five evaluated models. Its robust performance, combined with minimal preprocessing needs and resistance to overfitting, make it ideal for environmental modeling. While XGBoost remains a strong alternative with competitive accuracy and excellent interpretability, LightGBM offers better computational efficiency at the cost of slightly lower prediction performance. TabTransformer, despite its potential, requires larger datasets and advanced tuning to be effective. These findings are crucial for policymakers, enabling more accurate and timely AQI forecasts for public health intervention. The study concludes that ensemble learning methods—especially boosting models—are best suited for structured AQI datasets and recommends further integration of real-time data sources and deep learning enhancements for future improvements.

Future Directions

Future research should explore hybrid models that integrate CatBoost or XGBoost with deep learning for long-term AQI forecasting. Incorporating real-time pollution tracking, satellite data, and IoT sensors will enhance accuracy and spatial coverage. Emphasis should also be placed on developing climate-responsive models and cloud-based platforms for scalable, real-time AQI monitoring to assist governments and environmental researchers worldwide.

Contact Information

Pullabhotla Vijay
VIT-AP, Amaravati
vijay.21bce7006@vitapstudent.ac.in
+91 8978100562

References

- Breiman, L. (2001). "Random forests." Machine Learning, 45(1), 5-32. https://doi.org/10.1023/A:1010933404324
Discusses the Random Forest algorithm and its applications in predictive modeling.
- Cultivating clean skies: unveiling the tapestry of air quality in Gujarat through innovative machine learning analysis. https://www.bidacv.com/article_208453.html

Acknowledgements

I am immensely grateful to Dr. Sukanta Nayak for his constant support, expert guidance, and encouragement during this research. I would also like to thank VIT-AP University for offering the essential resources and infrastructure that made this work possible.