EasyChair Preprint

№ 15683

Machine Learning for Air Quality Prediction

Arafat Amrullaev and Remudin Reshid Mekuria

July 3, 2025

Machine Learning for Air Quality Prediction

Arafat Amrullaev
Ala-Too International University
Bishkek, Kyrgyzstan

Remudin Reshid Mekuria
Ala-Too International University
Bishkek, Kyrgyzstan

ABSTRACT

With air pollution being one of the major problems for urban sustainability and human health in Bishkek, accurate prediction of air quality is crucial. This study aims to predict air quality using machine learning techniques with meteorological and pollution concentration data. Using historical dataset from 2019 to 2023, we employed the CatBoostRegressor model to predict the Air Quality Index, prioritizing features such as humidity, pressure, and historical values of pollutants. Our model demonstrated exceptional performance, achieving a lower value of RMSE and R^2 of 0.98 on validation dataset. Our findings support the potential of machine learning in environmental monitoring and suggest improvements in air quality awareness programs.

KEYWORDS

Machine Learning, Air Quality Prediction, Meteorological Data, AQI, Data Analysis

CLASSIFIER

UDC 504.3 - Air Pollution, 004.8 - Artificial Intelligence

1 INTRODUCTION

Growing air quality challenges related to rapid urbanization, industrial activities, and vehicle emissions face Bishkek, the capital city of Kyrgyzstan. Specific geographical location-nestling in the mountains-the city further aggravates the situation by trapping pollutants with minimal air circulation. Local environmental studies suggest that in Bishkek, air pollution has reached alarming proportions, especially during the winter months, when increased heating fuels emit high concentrations of particulate matter and other harmful pollutants. These conditions are very dangerous for public health since, in the short run, there is an increase in respiratory and cardiovascular diseases, but they also have a wider implication for the environment and quality of life in the longer term [1].

Traditionally, air quality monitoring in Bishkek relies on a small number of fixed monitoring stations that, in turn, have insufficient spatial and temporal coverage of air quality data. The emerging machine learning techniques create new hopes for improvement in air quality prediction and management. Large inputs coming from sources like meteorological parameters and the historical air quality indices are taken into consideration by ML algorithms that unravel complex relationships and patterns to enhance the accuracy of prediction.

The highlight of the key points studied in this paper includes the review of existing machine learning models for prediction of AQI within Bishkek using CatBoostRegressor. By incorporating current meteorological data and historical pollutant levels, the study is bound to develop robust predictive models that shall inform public health initiatives and environmental policies relevant to the local context. The central hypothesis is that incorporating diversified data will add considerable value in the prediction of AQI in Bishkek for the formulation of efficient strategies in air quality management.

This research will review the prior work conducted within similar urban areas through a comprehensive literature review and identify any shortcomings in the existing methods which are pertinent to Bishkek. The findings from this study will contribute not only to the scholarly knowledge of air quality processes in Bishkek but also to the actionable knowledge for both policy makers and the community for mitigating the acute problem of air pollution.

2 LITERATURE REVIEW

Understanding and predicting air quality have become essential for mitigating its adverse impacts, particularly in heavily polluted cities like Bishkek. This has led to increased interest in leveraging advanced computational techniques, such as machine learning, for air quality monitoring and forecasting.

Machine learning has proven effective across a wide range of predictive modeling tasks, including domains beyond environmental monitoring. For example, Al Khan et al. applied ML algorithms such as Random Forest, LSTM, and LASSO to predict software defects using historical development data and software metrics [2]. Their results highlight the value of combining time-based features with supervised learning to forecast failures, which parallels our approach in forecasting air quality. Although the domains differ, both studies demonstrate how temporal patterns and feature engineering contribute to improved model performance in practical, real-world systems.

In a recent study by Sadriddin 2024 et. al, Random Forest Regression (RFR) and Long Short-Term Memory (LSTM) models were used to predict air quality index in Bishkek, where Random Forest Regression showed higher accuracy and lower Root Mean Square Error (RMSE) compared to LSTM [3].They collected historical AQI data and meteorological data, including air temperature and relative humidity from 5 stations in Bishkek. This study also showed the importance of lagging variables. The inclusion of lagging variables significantly improved the performance of the RFR model compared to the model without lagging variables. They were able to achieve 89% accuracy and RMSE of 27.3 with the RFR model.

In the study by E. Isaev et al., various models such as ANN, RFR, XgbR, KNN, DTR were tested [4]. Among these, the combination of RFR with Hyperparameter Tuning (HPT) and XgbR with HPT demonstrated superior performance in forecasting air quality. Based on the findings, an operational air quality forecasting system was developed using the RFR model with HPT. This system was integrated into the standard practices of Kyrgyzhydromet, the national meteorological service of Kyrgyzstan, to enhance air quality monitoring and prediction efforts in Bishkek.

Automated methods are becoming ever more noticeable in specialized fields. As an illustration, the authors Rakimbekuulu et al. suggested a code generation framework aimed at facilitating ablation techniques, showcasing how automated software development can tackle challenges specific to certain domains [5]. Even though their emphasis is on another domain, the basic idea of automating intricate decision-making processes via structured input is in close alignment with our application of machine learning for air quality prediction. Both methods underscore the promise of fusing algorithmic reasoning with expertise in a specific area to improve the effectiveness of decision-making and lessen the need for manual work.

3 METHODOLOGY

Data collection

To train the model, historical AQI data and meteorological data were collected. The source for historical AQI data was AirNow, whose website has open access to hourly data from 2019 to 2024 [6]. Meteorological data was taken from “timeanddate” website, which has open access to data for every 3 hours [7]. To collect data from the “timeanddate” website, a parser was used that scrapes data from the website from 2019 to 2024.

Data Processing

Before using our dataset to train the model, several data preprocessing steps were done to ensure data quality and compatibility with machine learning algorithms. Important columns such as NowCast Conc., AQI, AQI Category, Raw Conc. were retained. Columns that do not have an impact on prediction were excluded from the dataset. Missing Values were filled using Interpolation, which estimates missing data points by leveraging the surrounding data values, ensuring smoothness and consistency within the dataset. Time features such as year, month, day, hour were extracted from the Timestamp column. TimeSeriesSplit is used to split the data into sequential training and testing sets while maintaining time order. This method prevents future data leakage.

Building a model

The target variable for prediction was the AQI value, that represents the continuous numerical air quality index. For our model we employed CatBoostRegressor, an ensemble-based gradient boosting algorithm, due to its ability to handle regression tasks efficiently and robust performance on tabular data. Since our data is time-based, we used TimeSeriesSplit with five splits to ensure chronological order was preserved between training and validation sets. This approach allowed robust evaluation by simulating real-world scenarios where future data is predicted based on past observations. To perform hyperparameter optimization, the Optuna framework was used, a powerful framework for automated hyperparameter search. The objective function minimized the root mean squared error (RMSE) by exploring hyperparameter combinations, including: iterations (Number of boosting), depth (Maximum depth of the tree), learning_rate (Step size shrinkage to prevent overfitting), l2_leaf_reg (Regularization coefficient to reduce model complexity), random_strength (Strength of randomization), bagging_temperature (Fraction of randomly selected features for each iteration), border_count

(Number of splits for numeric features). The final CatBoostRegressor model was trained on the entire training dataset using the best hyperparameters identified during the tuning process. A final split from TimeSeriesSplit was reserved to evaluate the model's performance before testing.

4 RESULTS

The CatBoostRegressor model was trained on and evaluated through the AQI dataset, which consists of the span from 2019 to until 2023, as well as new dataset. We focus on explaining the results through two different essential graphs. Figure 1 shows the predicted vs. actual AQI values for both train and test sets over time. The model is effective at correctly predicting the trend of AQI in both datasets, suggesting that it successfully captures some temporal variations in this data. Figure 2 is how we evaluated the model on new data. Here is a graph for 6 months of 2024 prediction vs actual AQI that shows most matching of predicted and actual here which clearly states that the model is very stable with unseen data. Performance was evaluated using the following metrics: Root Mean Squared Error (RMSE), R-squared (R^2). The RMSE for the training dataset was 2.09 and 0.67 on the test set. The RMSE on unseen data was 6.03. The results from Figure 1 show that for both training and test datasets, the CatBoostRegressor model had low RMSE values which indicated a good generalization of how enough underlying patterns were learned in AQI data to accurately predict. The peaks and troughs in AQI — the temporal dynamics which is typical of many real-world data, were also captured well by our model thereby highlighting robustness to time series. Figure 2 further supports the model's generalization ability. On unseen data, the predictions remain accurate, with minimal deviations from the actual values. This highlights the robustness of the CatBoost model in predicting AQI even when exposed to data from different temporal or environmental conditions. Figure 3 shows the learning curve of the model, with training and validation RMSE plotted against the number of iterations. The curves indicate a steady decline in error, stabilizing after approximately 800 iterations. The close proximity of the training and validation RMSE curves suggests that the model avoids overfitting and generalizes well to unseen data. Nevertheless, there are still some inconsistencies in certain times when AQI fluctuates greatly (i.e. sharp peaks). This could be due to intrinsic noise in the data or lack of other influencing factors (e.g. meteorological) not present in the model. Furthermore, this could additionally contribute to accurate forecasting.

Figure 1: Comparison of actual and predicted AQI over time for the training and testing datasets using the CatBoostregressor

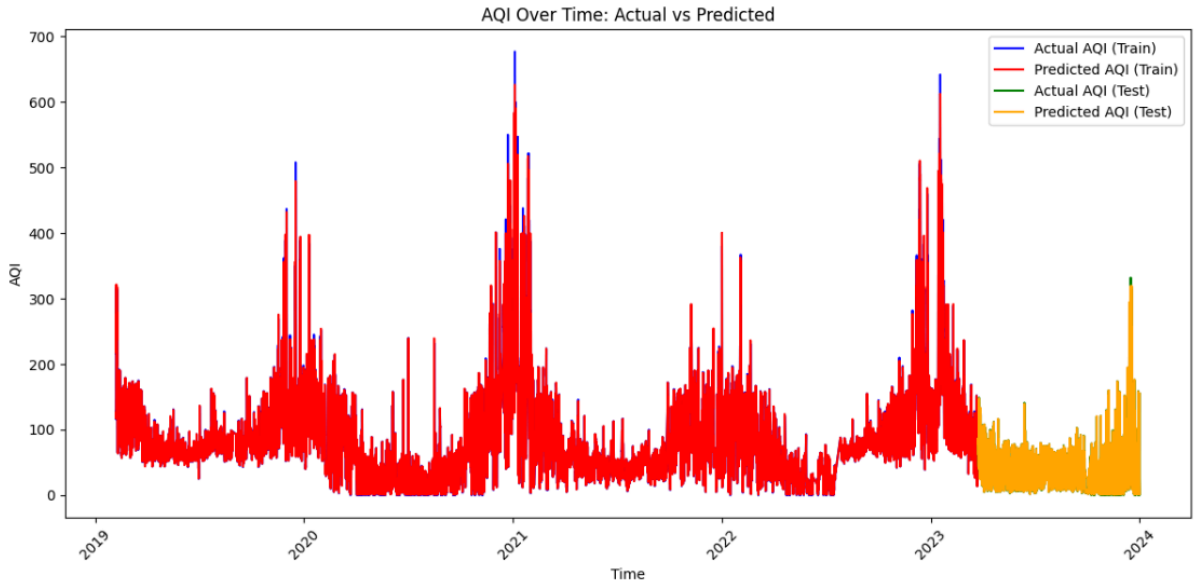

Figure 2: Evaluation of CatBoostRegressor predictions against actual AQI values on new data, showcasing the model's generalization ability.

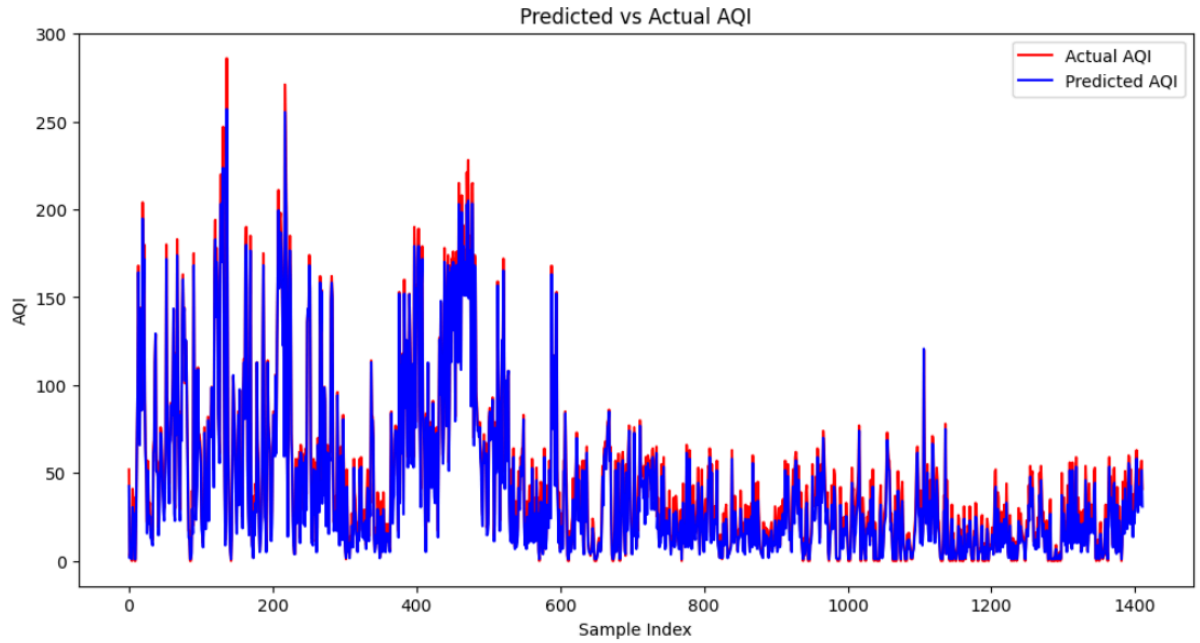

Figure 3: Learning Curve

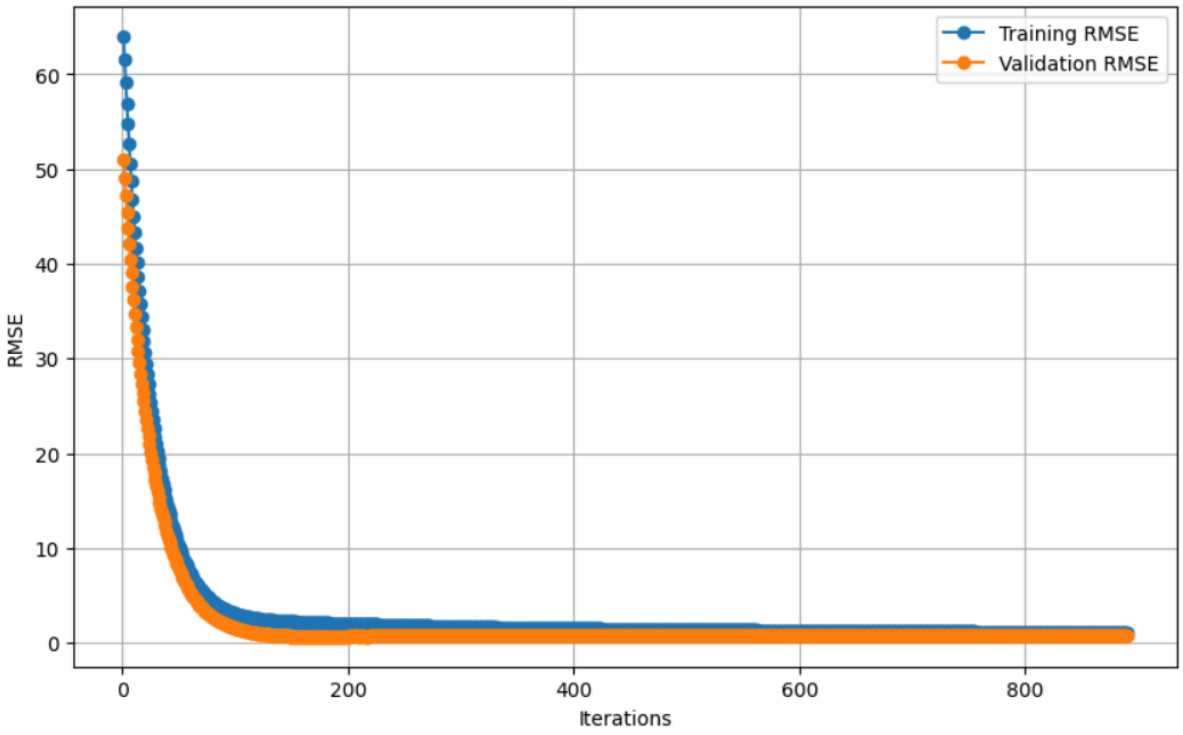

Table 1: Results

	RMSE	**R-2 Score**
Training Dataset	2.09	0.99
Testing Dataset	0.67	0.99
Cross-Validation	8.17	0.98
New Dataset	6.03	0.98

CONCLUSION

The CatBoost algorithm, which predicts AQI on the basis of historical time series data was used in this research. The results in Table 1 showed that this model was very successful in learning patterns for AQI data and generalizing to an unseen scenario, the high level of closeness between predicted values vs. Actual during testing and on a new set of data not included at training time.

This paper showcases the capabilities of CatBoostRegressor for air quality prediction tasks. In the future, we can also improve our method by including more meteorological variables and different combinations of feature engineering. Furthermore, comparing it with other machine

learning models could contribute to understanding better in which ways CatBoostRegressor is adequate.

REFERENCES

[1] Dzushupov, K. O., Buban, J. M., Aidaraliev, A. A., Ahmadi, A., Chahal, P., Ibrahim, M., ... & Kouwenhoven, M. B. N. (2022). Air pollution in Bishkek, Kyrgyzstan: Driving factors and state response. *Public Health Challenges*, *1*(4), e22.

[2] Khan, A., Mekuria, R. R., & Isaev, R. (2023, April). Applying machine learning analysis for software quality test. In *2023 International Conference on Code Quality (ICCQ)* (pp. 1-15). IEEE.

[3] Sadriddin, Z., Mekuria, R., & Gaso, M. (2024). Machine Learning Models for Advanced Air Quality Prediction. In *Proceedings of the International Conference on Computer Systems and Technologies 2024* (pp. 51–56). Association for Computing Machinery.

[4] Isaev, E., Ajikeev, B., Shamyrkanov, U., Kalnur, K. U., Maisalbek, K., & Sidle, R. C. (2022). Impact of climate change and air pollution forecasting using machine learning techniques in Bishkek. *Aerosol and Air Quality Research*, *22*(3), 210336.

[5] Rakimbekuulu, S., Shambetaliev, K., Esenalieva, G., & Khan, A. (2024, November). Code Generation for Ablation Technique. In *2024 IEEE East-West Design & Test Symposium (EWDTS)* (pp. 1-7). IEEE.

[6] *AirNow.gov*. (n.d.). https://www.airnow.gov/

[7] timeanddate.com. (2024, December 3). *Timeanddate.com*. https://www.timeanddate.com/