Air Quality Forecasting Using Machine Learning: A Global Perspective with Relevance to Low-Resource Settings

Review of Integrative Business & Economics Research

Christian Mulomba Mukendi*
Department of Advanced Convergence, Handong Global University, Pohang, South Korea

Choi Hyebong
School of Global Entrepreneurship and Information Communication Technology, Handong Global University, Pohang, South Korea

ABSTRACT

This research leverages a comprehensive global weather database to validate the potential of short-term air quality predictions using minimal data. This approach is particularly promising for resource-limited environments, which are often more vulnerable to such hazards compared to developed nations. The study employs machine learning methodologies and incorporates meteorological, air pollutant, and Air Quality Index (AQI) features from 197 capital cities. Our findings underscore the efficacy of the Random Forest algorithm in generating reliable predictions, especially when applied to classification rather than regression. This approach enhances the generalizability of the model on unseen data by 42%, considering a cross-validation score of 0.38 for regression and 0.89 for classification. To instill confidence in these predictions, different methods for explainable machine learning were considered. This research highlights the potential for resource-limited countries to independently project short-term air quality while waiting for larger datasets to enhance their predictions. Implementing this approach could promote public health and reduce dependence on external entities. In conclusion, this study serves as a guiding light, paving the way towards accessible and explainable air quality forecasting**.**

Keywords: Machine learning forecasting, resource-constrained settings, short-term air quality projection, cost estimation.

Received 6 December 2023 | Revised 2 March 2024 | Accepted 8 April 2024.

1. INTRODUCTION

Air pollution, which consists of harmful chemicals or particles in the air, poses a significant risk to the health of humans, animals, and plants, making it a complex issue to tackle. As reported by *Nationale Geographic* (n.d.) - Jillian Mackenzie and Jeff Turintine (2023), air pollution is now the world's fourth-largest risk factor for early death, causing approximately 4.5 million deaths in 2019 due to exposure to outdoor air pollution and nearly 2.2 million deaths from indoor air pollution. Thus, environment awareness plays a significant role nowadays more than ever before Daniel Yudistya Wardhana (2022). The issue related to air pollution is particularly prevalent in large cities where emissions from various sources are concentrated. Moreover, climate change exacerbates the production of allergenic air pollutants, necessitating urgent action. Current mainstream research in this field is primarily focused on understanding the health effects of air pollutants in the short and long term, especially on vulnerable populations. There is also a strong emphasis on the use of technology

ISSN: 2304-1013 (Online); 2304-1269 (CDROM); 2414-6722 (Print)

and big data to innovate in health science and enhance our understanding of the impact of air pollution. Monitoring air quality through observations and instrumentation, as well as modeling air quality, is considered crucial for making accurate projections, informing policy decisions, and guiding public health interventions and communication strategies.

These strategies are being developed to effectively convey information about air pollution risks and the necessary interventions. In terms of technology use, since their development has transformed almost all aspects of humans activities (Purbasari et al., 2023), machine learning is seen as a game-changer. By leveraging large datasets, it provides valuable insights from the wealth of information available, aiding in the development of robust responses to this hazard. For instance, the research of Méndez et al. (2023) reviewed machine learning algorithms applied in forecasting air quality from 2011 up to 2021 giving more insight on the features considered and the effectiveness of algorithms considered. Research of Hasnain et al. (2022) provided the result for the prediction of both short and long term of air quality in the Jiangsu province in China based on Prophet forecasting in forecasting the concentration of air pollutants. The estimation of PM2.5 levels in air was conducted in Garg and Jindal (2021) where ARIMA, Facebook Prophet, 1D CNN and LSTM was compared. Their results showing the good performance of LSTM in terms of mean absolute percentage error. In Kumar and Pande (2023), several machine learning algorithms were compared to predict air quality in India showing the good performance of the XGBoost compared to the naïve Bayesian and support vector machine. In Maduri et al. (2023), the LightGBM, GBM and Random Forest were used to predict air quality using physical parameters and showing how they outperform deep learning algorithms in predicting the level of contaminations in the nearby area. According to Yang et al. (2022), meteorological features wield significant influence in forecasting air quality when integrated with air pollutant features. Utilizing explainable machine learning, specifically the Shapley Additive Explanation method, the analysis reveals that enhancements in air quality are not solely achieved through the incorporation of meteorological features. Instead, the synergy between meteorological features and certain pollutant features proves pivotal, highlighting the importance of their interactive effects in achieving improved air quality. Current trend and challenges in the prediction of air quality is discussed in Sokhi et al. (2022) where the use of different source of information is considered as relevant to integrate the results of predictions which is of a high importance for policy makers.

While a variety of promising solutions are being offered, countries with limited resources often struggle to analyze and implement their own tools to anticipate hazardous air quality even though they are more exposed to these hazards compared to developed countries (Méndez et al., 2023). There are several tools available globally that can provide such information, but in some regions, certain information is not accessible due to these limitations, making these countries more susceptible to this risk. Moreover, the use of machine learning often requires extensive datasets to be effective. However, countries with limited resources may lack the necessary resources or time to develop robust solutions to this ever-increasing hazard. This underscores the importance of the current study, which aims to provide a straightforward yet effective method for achieving very short-term air quality projections using two months of data. By leveraging the unique information source provided by the world weather repository, a reliable projection of air quality using the air_quality_gb-defra-index[1] was accomplished, and the results were generalized to various countries. To bolster confidence in the results, an explainable machine learning approach was employed, incorporating the use of Local Interpretable Model-agnostic Explanation (LIME) (Zhu et al.,

[1] https://uk-air.defra.gov.uk/air-pollution/daqi

ISSN: 2304-1013 (Online); 2304-1269 (CDROM); 2414-6722 (Print)

2023), Explain like I am 5 (Eli5)Gezici and Tarhan (2022), and Partial Dependent Plots (PDPs) (Nduwayezu et al., 2023), thereby validating the results as authentic and worthy of consideration. The remainder of this paper is organized as follows: Section two is dedicated to the methodology, while section three discusses the results. The conclusion is presented last.

2. METHODOLOGY

2.1. Dataset

The research utilized the World Weather Repository NIDULA ELGIRIYEWITHANA (2023), a real-time dataset publicly accessible which offers over 40 environmental and weather-related features for approximately 197 capital cities worldwide. Data recording commenced on August 29, 2023, and continues to be regularly updated. The air quality index (AQI) to predict is the air_quality_gb-defra-index, developed in the United Kingdom. This index provides a range of values for air quality, from 0 to 10, with 0 indicating low air pollution and 10 indicating very high air pollution. Table 1 provide the definition of each variable.

Table 1: Description of variables in the dataset

Features names	Definition	Units
Country	Name of the country	
Location name	Name of the capital city where the temperature was recorded	
Temperature_celsius	Temperature in the area	Degree Celsius
Wind_mph	Wind speed	Miles per hour
Wind_degree	Wind direction	Degree
Wind_direction	Wind direction as a 16-point compass	
Pressure_mb	Level of pressure	Millibars
Precipitation_mm	Idem	Inches
Humidity	Level of humidity in the atmosphere	Percentage
Cloud	Cloud cover	Percentage
Feels_like_celsius	Human feeling of the tempereature	Celsius
Visibility	Human distance of visibility	Kilometers
UV_index	Ultra violet index from the sun	
gust_mph	Wind gust	Miles per hour
Air_quality_Carbon_Monoxide,	Measurement of Carbon Monoxide in air	
Air_quality_Nitrogen_Monoxide	Measurement of Nitrogen Monoxide in air	
Air_quality_Ozone	Measurement of ground level ozone in air	
Air_quality_sulphur_dioxide	Measurement of sulphur dioxide in air	
Air_quality_PM2.5	Concentration of fine inhalable particles with a diameter of 2.5 micrometers or less in the air	
Air_quality_PM10	Concentration of fine inhalable particles with a diameter of 10 micrometers or less in the air	

ISSN: 2304-1013 (Online); 2304-1269 (CDROM); 2414-6722 (Print)

The dataset had missing information (not missing values) for some countries, representing one percent of the entire dataset. To ensure robustness and reliability, the dataset was used as is. Three types of features were considered for prediction: meteorological (Temperature_celsius, Wind_mph, Wind_degree, Wind_direction, Pressure_mb, Precipitation_mm, Humidity, Cloud, Feels_like_celsius, Visibility, UV_index and gust_mph) 12 in total, the air quality index, one (1), and air pollutant Air_quality_Carbon_Monoxide, Air_quality_Nitrogen_Monoxide, Air_quality_Ozone, Air_quality_suylphur_dioxide, Air_quality_PM2.5, Air_quality_PM10, 6 in total, bringing the number of features to 19. Among the meteorological features, Feels_like_celsius (Rajat Lunawat, 2022) was added to evaluate the impact of subjectivity in the model performance.

2.2. Exploratory data analysis

Two overlapping clusters of countries for the period under consideration was observed, the first consisting of 166 countries, while the second comprises 197 countries. The distinction between them resides in the fact that both the meteorological and AQI indices are higher in the second cluster compared to the first, which represent days with more extreme weather conditions compared to the first cluster. These conditions include higher temperatures, stronger winds, more precipitation, etc. Conversely, the lower levels of air pollutant features in the second cluster indicate better air quality compared to the first cluster. This suggests that the second cluster represents days with cleaner air. Consequently, all capital cities have experienced varying degrees of poor air quality, even those that have demonstrated a very good AQI during this period. In this study, these cities are categorized as 'differences', totaling 31 in number. Figure 1 and Table 2 illustrate this.

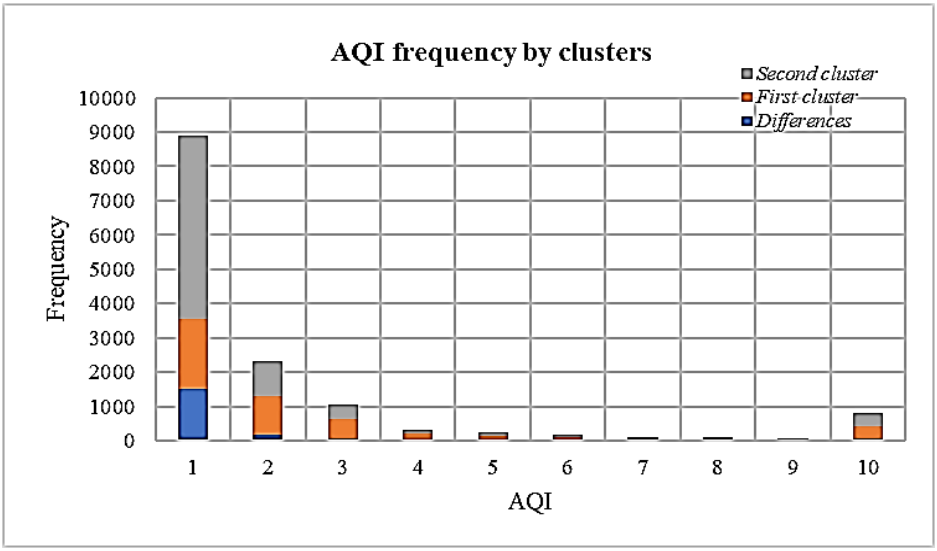


Figure 1: AQI by cluster

Table 2: AQI Frequency by cluster

AQI	1	2	3	4	5	6	7	8	9	10
Cluster 1	2018	1144	554	188	148	101	73	66	44	353
Cluster 1	5338	1025	424	125	107	61	36	50	38	382
Differences	1528	172	68	20	14	9	3	4	4	73

ISSN: 2304-1013 (Online); 2304-1269 (CDROM); 2414-6722 (Print)

This result confirms the existing interaction among the features considered. However, a correlation analysis illustrates the non-existent correlation among them (Figure 2)

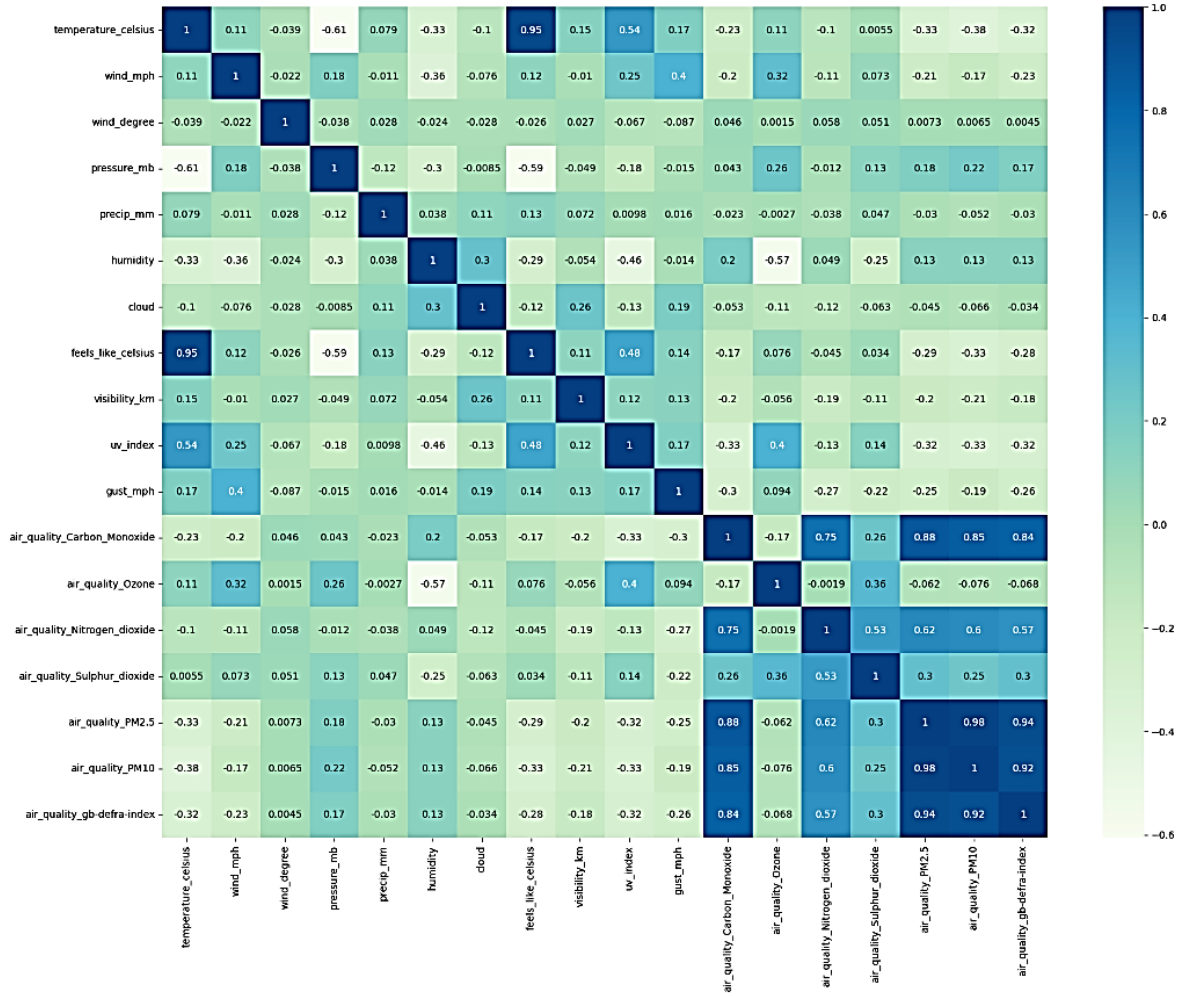


Figure 2: Correlation table

There is a strong correlation among some air pollutant features to the AQI air_quality_PM2(0.94), air_quality_PM10 (0.92) and air_quality_carbon_monoxyde (0.84), while others are lowly correlated [air_quality_ozone (-0.068), air_quality_Nitrogen_dioxyde (0.57) and air_quality_sulfur_dioxyde (0.3)], but not with AQI and meteorological feature.

2.3. Machine learning algorithms

Several regressors and Classifiers were considered based on their good performance in similar cases. These regressors were utilized: *Linear Regression* - Huang (2023), *Ridge*Singh et al. (2023), *Decision Tree Regressor* - Luo et al. (2021), *Random Forest Regressor* - El Mrabet et al. (2022), *XGBoost Regressor* - Patel et al. (2022), *Light GBM Regressor* - Khawaja et al. (2023), *and Support Vector Regressor* - Khawaja et al. (2023). For classification, the classifiers used were: *Logistic Regression* - Wichitaksorn et al. (2023), *Random Forest Classifier* - Schonlau and Zou (2020), *Decision Tree Classifier* - Charbuty and Abdulazeez (2021), *KNeighbors Classifier* - Alkaaf et al. (2020), *XGBoost Classifier* - Swathi and Kodukula (2022), *Light GBM Classifier* - Naim et al. (2022) and Alam et al. (2020). This makes a total of 14 algorithms.

ISSN: 2304-1013 (Online); 2304-1269 (CDROM); 2414-6722 (Print)

2.4. Metrics

Our approach to evaluating each algorithm involved two rounds of metric assessments. The initial round was designed to assess the algorithm's performance on future or unseen data for projection purposes. Following the selection of the most effective algorithm, the second round was conducted to evaluate its performance on the training data. For regression tasks, the first round of metrics included the Mean Squared Error (Hodson et al., 2021), R squared, whose a higher value explains better generalization from the model (Eugenio and Guhao Jr, 2023; Karch, 2020), Cross-Validation Score (Yates et al., 2023) using 5 folds, and Residuals (Zhang et al., 2018). The second round considered the normalized mean squared error (Handel, 2018) (nRMSE), with a threshold below 10 percent for each country to be deemed as a satisfactory prediction. For classification tasks, the first-round metrics included the Cross-Validation Score (Yates et al., 2023) using 5 folds, accuracy, precision, recall, and F1 score (AMAN KHARWAL, n.d.). The second round utilized the Chicco and Jurman (2020) Classification Report (AMAN KHARWAL, n.d.), and Confusion Matrix (Heydarian et al., 2022). Given that the primary focus of our work is on projection, the Cross-Validation Score (Yates et al., 2023) serves as a particularly useful metric in determining which algorithm is likely to perform better on unseen data. This comprehensive evaluation process ensures a robust assessment of each model's performance.

2.5. Explainable machine learning

To enhance trust in prediction provided by the algorithm, three popular but powerful interpretability tools were employed, namely, the LIME for instance-based interpretation, ELI5 to visualize the contribution in terms of weights of each feature to the prediction, thereby offering a comprehensive view of the model's performance and lastly, PDPs to visualize the pattern of contribution of each variable to the prediction of the target.

2.6. Research design

The dataset was prepared for one-day projection by grouping information by country either for regression and or classification task. This preparation excluded the last information of each group (information of 2023-10-30) to be used as scenario for projection of the next day (2023-10-31). For the classification approach, the Air Quality Index (AQI) was grouped according to the categories present in the dataset before grouping information by country and preparing data for a time series classification. This process ensures that the data is appropriately structured for both regression and classification tasks. The prepared data was subsequently utilized to train the considered regressors and classifiers, with each model's performance being evaluated accordingly. The model that demonstrated a good cross-validation score and performed well on other metrics was selected to generate a scenario for projecting the next day's air quality. To assess the model's performance, a country from among the low-resource countries was chosen, and the information for the last day was withheld. The remaining information was then fed into the model to predict its value. Given that the data used for prediction, minus the last day's information, will yield a result for the hidden day, this hidden day's information is later used as a scenario to predict the next non-existent day. Despite the inherent uncertainty of climate, this approach allows for a certain level of confidence to be built in the model's performance. Lastly, the explainable machine learning components were implemented to enhance confidence in the prediction results. This was done before comparing the two approaches (regression and classification) to determine which one provides the highest level of confidence for rapid implementation. Thanks to this

ISSN: 2304-1013 (Online); 2304-1269 (CDROM); 2414-6722 (Print)

design, one can leverage these models and place trust in the outcomes they provide. The result of analysis and projections are available on my github[2]

3. RESULTS

3.1. Model selection

Table 3 and 4 provide the result for the best model for each approach.

Table 3: Results for regression

Regressor	Mean CVS	MSE	R2 score	Mean Residuals
Linear Regression	0.39	0.05	0.41	-1.2 e-15
Ridge	0.39	0.05	0.41	-1.2 e-15
Decision Tree	-0.25	4.5 e-35	1.00	-3.44 e-19
Random Forest	**0.38**	**0.0067**	**0.91**	**-0.004**
XGBoost	0.31	0.01	0.88	0.00011
LGBM	0.39	0.02	0.65	-0.0051
SVR	0.37	0.04	0.52	-0.0051

Table 4: Results for classification

Classifier	Mean CVS	Accuracy	Precision	Recall	F1
Logistic Regression	0.88	0.88	0.83	0.88	0.85
KNeighbors	0.87	0.90	0.88	0.90	0.89
Decision Tree	0.81	1.00	1.00	1.00	1.00
Random Forest	**0.89**	**1.00**	**1.00**	**1.00**	**1.00**
XGBoost	0.87	0.99	0.99	0.99	0.99
LGBM	0.88	0.99	0.99	0.99	0.99
SVC	0.89	0.89	0.86	0.89	0.86

The two Tables illustrate the superior performance of the Random Forest algorithm. On the regression approach, the LGBM model provided the best cross-validation score (0.39). However, in terms of Mean Squared Error (MSE) (0.02) and coefficient of variation (0.65), it was unable to surpass the performance of the Random Forest model (0.0067 and 0.91 respectively). The Decision Tree model, on the other hand, was found to overfit the data. In the classification task, the Support Vector Classifier (SVC) and the Random Forest model both achieved the highest cross-validation score (0.89). However, considering other metrics, the Random Forest model outperformed the SVC, making it the most suitable model for both cases despite a slight risk of overfitting, as indicated by the residuals (-0.004).

3.2. Model selection

In the second round of evaluation using the best model (Random Forest), the mean nRMSE on the trained data was 0.089, and the mean residuals were 0.03. The number of capital cities with a nRMSE above the threshold of 10% was 73, with values ranging between 11 and 32 percent. This represents 37% of the total capital cities. For the classification task, despite the presence of imbalanced classes, the Matthews Correlation Coefficient was 1.0, and the classification report (a) and confusion metrics (b) were perfect, as illustrated in the group of Figure 3.

[2] https://github.com/Dechrist2021/Mulomba.git

ISSN: 2304-1013 (Online); 2304-1269 (CDROM); 2414-6722 (Print)

Classification report:

	precision	recall	f1-score	support
0	1.00	1.00	1.00	219
1	1.00	1.00	1.00	10167
2	1.00	1.00	1.00	708
3	1.00	1.00	1.00	787
accuracy			1.00	11881
macro avg	1.00	1.00	1.00	11881
weighted avg	1.00	1.00	1.00	11881

(a)

Confusion matrix:

```
[[  219     0     0     0]
 [    0 10167     0     0]
 [    0     0   708     0]
 [    0     0     0   787]]
```

(b)

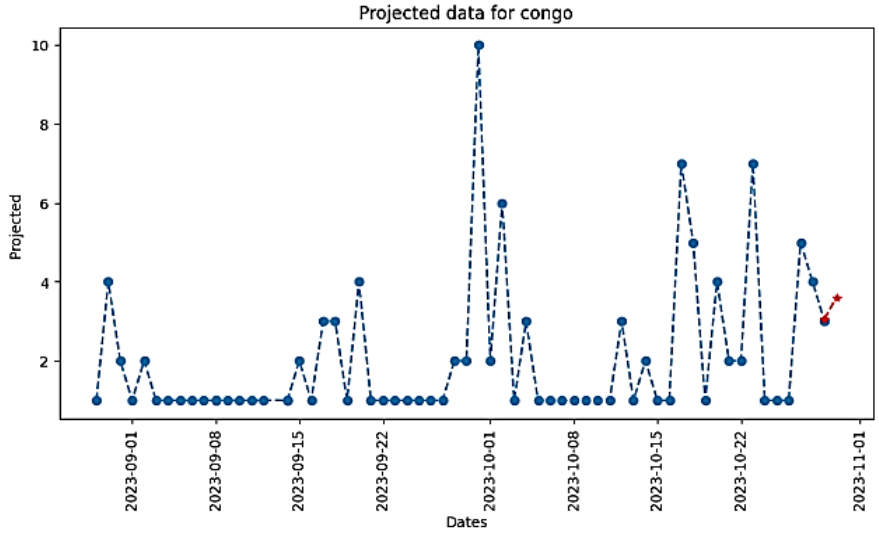
Figure 3: Classification report and confusion matrix

A 42% improvement in generalizability was observed when using classification over regression. The cross-validation score being respectively for regression and classification 0.38 | 0.89. This result suggests that the classification approach using Random Forest model is more suitable for this case. Actually, The, through its multiple decision trees constructed during training using a process known as bootstrap aggregating or bagging, this model was able to better predict the mode of the class (which is the class itself) or the mean prediction for regression, of the individual trees. This was achieved while maintaining a good balance between bias and variance, which is crucial to prevent overfitting or underfitting. Furthermore, Traditionally, regression models have been employed to predict continuous air quality values. However, the study demonstrates that classification models can outperform regression models in this context, achieving a better result. This improvement in performance suggests that classification models could be a more effective approach for short-term air quality forecasting in resource-constrained settings. Indeed, the use of classification models offers several advantages over regression models. First, classification models are generally simpler to interpret, which can be beneficial when working with limited data. Second, classification models are less sensitive to outliers and noise, which can be common in air quality data. Finally, classification models can be more computationally efficient, which can be important when working with limited resources.

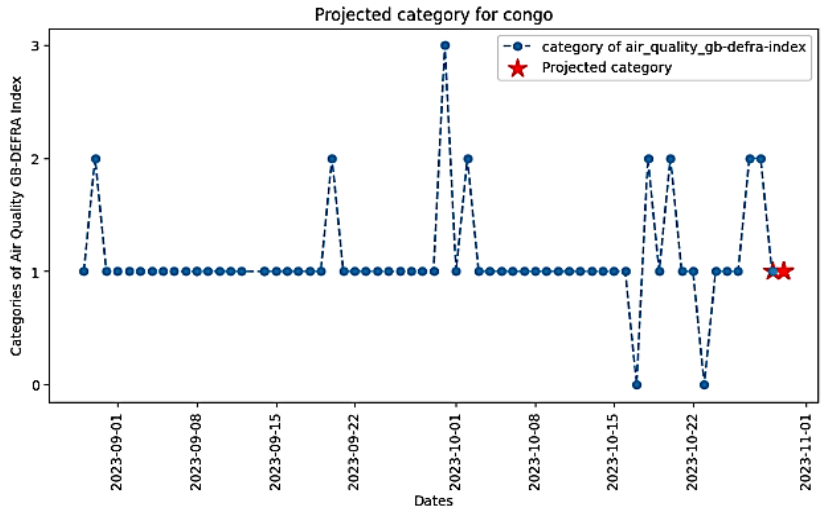
3.3. Case study

The Democratic Republic of Congo which is among the low resource country was considered. According to the dataset, for each category, the number of instances observed were: 0 = 2, 1 = 53, 2 = 7, 3 = 1. It is evident that the Air Quality Index (AQI) in Kinshasa is typically moderate, although there was an instance when it reached a very high level. By utilizing the data from October 28 and 29, 2023, to predict subsequent values, the models accurately predicted the AQI for October 29, 2023. This confirms that the model is well-trained and capable of providing a projection for October 30, 2023. The same methodology was applied for classification. The time series plot of these results, provided in the Figure 4, offers a better visual of the situation.

ISSN: 2304-1013 (Online); 2304-1269 (CDROM); 2414-6722 (Print)



(a)



(b)

Figure 4: Result of projection using regression and classification

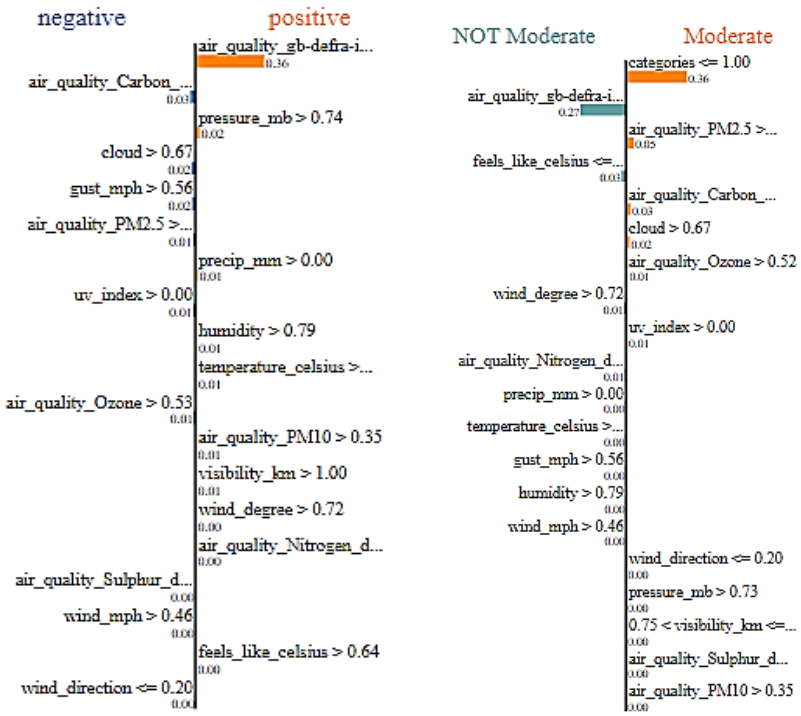
The projected values for classification (b) are 1 for both dates. For regression (a), the value is 3.28 for October 29, 2023, which matches the actual value for that date, and 3.61 for October 30, 2023, which is the projected value. Therefore, both results fall within the same category of moderate AQI. This indicates that the model could accurately predict the AQI category for these dates.

3.4.Model explainability

3.4.1. LIME

Applied on regression, a positive contribution of AQI, pressure_mb, precipitation_mm, humidity, temperature, air_quality_PM10, visibility_km, wind degree, air_quality_Nitrogen_dioxyde, feels_like_celsius was observed while the rest contributed negatively. On a classification approach, a positive contribution of categories, air_quality_PM2.5, air_quality_Monoxyde, cloud, air_quality_ozone, uv_index, wind direction, pressure_mb, visibility_km, air_quality_sulphure_dioxide and air_quality_PM10 was observed while the rest contributed negatively (see Figure 5).

ISSN: 2304-1013 (Online); 2304-1269 (CDROM); 2414-6722 (Print)



(a) (b)

Figure 5: LIME for regression (a) and classification (b)

3.4.2. Eli5

Weight	Feature
0.4011 ± 0.0238	air_quality_gb-defra-index
0.0710 ± 0.0160	air_quality_PM2.5
0.0441 ± 0.0127	gust_mph
0.0403 ± 0.0134	air_quality_Ozone
0.0401 ± 0.0142	air_quality_Carbon_Monoxide
0.0399 ± 0.0118	feels_like_celsius
0.0398 ± 0.0126	air_quality_Sulphur_dioxide
0.0390 ± 0.0121	pressure_mb
0.0381 ± 0.0131	air_quality_PM10
0.0376 ± 0.0139	air_quality_Nitrogen_dioxide
0.0367 ± 0.0117	humidity
0.0338 ± 0.0109	wind_degree
0.0308 ± 0.0104	temperature_celsius
0.0300 ± 0.0096	wind_mph
0.0245 ± 0.0087	wind_direction
0.0227 ± 0.0090	cloud
0.0126 ± 0.0061	precip_mm
0.0092 ± 0.0063	uv_index
0.0090 ± 0.0055	visibility_km

(a)

Weight	Feature
0.1596 ± 0.1620	categories
0.1347 ± 0.1427	air_quality_gb-defra-index
0.0644 ± 0.0472	air_quality_Ozone
0.0575 ± 0.0221	air_quality_PM2.5
0.0559 ± 0.0286	air_quality_PM10
0.0488 ± 0.0136	feels_like_celsius
0.0468 ± 0.0149	air_quality_Carbon_Monoxide
0.0454 ± 0.0154	air_quality_Nitrogen_dioxide
0.0437 ± 0.0117	humidity
0.0435 ± 0.0132	gust_mph
0.0426 ± 0.0134	wind_degree
0.0416 ± 0.0130	pressure_mb
0.0411 ± 0.0124	air_quality_Sulphur_dioxide
0.0411 ± 0.0131	temperature_celsius
0.0357 ± 0.0131	wind_mph
0.0323 ± 0.0113	wind_direction
0.0288 ± 0.0156	cloud
0.0157 ± 0.0107	visibility_km
0.0117 ± 0.0067	precip_mm
0.0091 ± 0.0061	uv_index

(b)

Figure 6: Eli5 for regression (a) and classification (b)

The best model considered on regression provided a strong contribution of the AQI followed by air pollutant features while the visibility on the other hand contributed the least. Among the meteorological features, gust_mph contributed the most followed by feels_like_celsius. Applied on classification, the category contributed the most and the uv_index on the other

ISSN: 2304-1013 (Online); 2304-1269 (CDROM); 2414-6722 (Print)

hand contributed the least. The pollutants features contributed more after the category and among the meteorological features, the feels_like_celsius contributed the most (see Figure 6).

3.4.3. PDPs

On the regression approach, the temperature, wind_mph, wind_degree, wind direction, pressure_mb, precipitation_mm, humidity, feels_like_celsius indicate that both low and high values of the feature lead to high values of the predicted outcome. The AQI, air_quality_PM10, precipitation shows a rising trend meaning the positive correlation between them to the target. For the remaining variable, the trend is not well defined, sometimes decreasing, increasing and varying in different directions, meaning a complex and non-linear relationship to the target. On classification, each class depicted a different dependence. For classes I and 3, the temperature, wind_mph, wind_degree, humidity, feels_like_celsius, gust_mph, air_quality_PM10 shows both low and high values of the feature lead to a high probability of a certain class, while medium values of the feature lead to a low probability of that class. The remaining features start high and then decrease, remaining low over time, suggests that higher values of the feature are associated with a lower probability of predicting a certain class, indicating a possible negative effect of the feature to the predicted class. These features show a U-shaped relationship with the predicted class suggesting that extreme conditions (either low or high) of these weather factors are associated with the occurrence of the predicted class. The remaining features indicate a possible negative effect of these features on the predicted class. In other words, as these features increase, the likelihood of the predicted class decreases.

For the class 2, the 'temperature_celsius', 'wind_mph', 'wind_degree', 'wind_direction', 'pressure_mb', 'humidity', 'feels_like_celsius', 'gust_mph' suggests that both low and high values of the feature lead to a low probability of a certain class, while medium values of the feature lead to a high probability of that class. The other features depict a positive correlation to the target. The inverted U-shaped relationship with the predicted class suggests that moderate conditions of these weather factors are associated with the occurrence of the predicted class. For the other features, they show an increasing trend over time indicating a possible positive effect of these features on the predicted class. In other words, as these features increase, the likelihood of the predicted class increases. Finally for the class 3, the 'temperature_celsius', 'wind_degree', 'humidity', indicate a possible negative effect on the predicted class. 'gust_mph', 'air_quality_Carbon_Monoxide' suggests that as the value of the feature increases, the probability of a certain class (as predicted by the model) decreases. This mean that the features have a negative correlation with the predicted class. The higher values of the feature make the predicted class less likely. Others depict a nonlinear relationship to the target class meaning to have a complex relationship to the target. the influence of each feature on the target variable can fluctuate, depending on the specific aim of the prediction being either regression or classification case. Despite the utilization of diverse methodologies, it was discerned that the Air Quality Index (AQI) feature predominantly impacts the prediction for the subsequent day. This is followed by the pollutant index and meteorological features, with the 'feels like' temperature demonstrating a particularly significant impact in comparison to other meteorological features. In the context of classification, the category assumes a substantial role in forecasting, succeeded by the AQI feature and pollutant features. Notably, the 'feels like' temperature once again exhibits a considerable contribution, surpassing other meteorological features. This highlights the critical role of the perceived temperature in both regression and classification tasks within this context. The integration of

ISSN: 2304-1013 (Online); 2304-1269 (CDROM); 2414-6722 (Print)

the three types of features considered in this study, which includes the 'feels like' temperature within the meteorological features, underscores the significant role of subjective environmental conditions in the forecasting of AQI. This insight could prove instrumental in enhancing the accuracy and reliability of future air quality forecasts not only useful for understanding predictions but also for validating the model and ensuring its proper functionality.

4. CONCLUSION

The prediction of air quality has become a topic of high interest in recent times, considering its significant impact on society. This study provides a robust approach that could be utilized by countries with limited resources to develop their own projection tools. By combining limited data with the mature technology of machine learning, reliable projections can be made. To enhance trust in this approach, an explainable machine learning method was proposed, providing convincing evidence of the reliability of the obtained results. While these results are promising, there are some limitations to this study. The locations considered are only the capital cities. Although this gives a broad idea of the level of pollution, as there are often more people and activities in capital cities, it does not represent the pollution level of the entire country. In some cases, industrial regions could be more pollutant than the capital. Therefore, these results should be considered as representing the level of pollution only for the specified locations. Furthermore, in the set of features, meteorological, AQI, and pollutant features have been considered based on existing research. However, to deepen our understanding of the topic, it could be relevant to consider economic factors and human activity factors. These factors could be based on the time of exposure to the sun and the moon, as some activities with the potential to pollute air quality are strongly connected to these phases. Our results unfold the acknowledged capability of machine learning to provide reliable projections even with limited data but having a good level of granularity. Despite the limitations, this study marks a significant step forward in the use of machine learning for air quality prediction, particularly in resource-limited settings. Future research could build upon these findings by incorporating more diverse data and refining the machine learning models used.

ACKNOWLEDGEMENTS

The author is thankful to his thesis advisor for his invaluable support and to the reviewers for the insight they provided through their review.

REFERENCES

[1] Alam, Shamshe; Sonbhadra, Sanjay Kumar; Agarwal, Sonali; Nagabhushan, P. (2020): *One-Class Support Vector Classifiers: A Survey*, in: Knowledge-Based Systems 196, 105754.
[2] Alkaaf, Howida Abuabker; Ali, Aida; Shamsuddin, Siti Mariyam; Hassan, Shafaatunnur (2020): *Exploring Permissions in Android Applications Using Ensemble-Based Extra Tree Feature Selection*, in: Indonesian Journal of Electrical Engineering and Computer Science 19: 1, 543.
[3] AMAN KHARWAL (n.d.): *Classification Report in Machine Learning*.
[4] Charbuty, Bahzad; Abdulazeez, Adnan (2021): *Classification Based on Decision Tree Algorithm for Machine Learning*, in: Journal of Applied Science and Technology Trends 2: 01, 20–28.

ISSN: 2304-1013 (Online); 2304-1269 (CDROM); 2414-6722 (Print)

[5] Chicco, Davide; Jurman, Giuseppe (2020): *The Advantages of the Matthews Correlation Coefficient (MCC) over F1 Score and Accuracy in Binary Classification Evaluation*, in: BMC Genomics 21: 1, 6.
[6] Daniel Yudistya Wardhana (2022): *Environmental Awareness, Sustainable Consumption and Green Behavior Amongst University Students*, in: Review of Integrative Business and Economics Research 11: 1, 1–11.
[7] El Mrabet, Zakaria; Sugunaraj, Niroop; Ranganathan, Prakash; Abhyankar, Shrirang (2022): *Random Forest Regressor-Based Approach for Detecting Fault Location and Duration in Power Systems*, in: Sensors 22: 2, 458.
[8] Eugenio S; Guhao Jr (2023): *Prediction Models on Work Engagement among Public School Teachers: A Hierarchical Regression Analysis*, in: Review of Integrative Business and Economics Research 12: 4, 17–29.
[9] Garg, Satvik; Jindal, Himanshu (2021): *Evaluation of Time Series Forecasting Models for Estimation of PM2.5 Levels in Air*, in: 2021 6th International Conference for Convergence in Technology (I2CT), 1–8.
[10] Gezici, Bahar; Tarhan, Ayca Kolukisa (2022): *Explainable AI for Software Defect Prediction with Gradient Boosting Classifier*, in: 2022 7th International Conference on Computer Science and Engineering (UBMK), 2022 7th International Conference on Computer Science and Engineering (UBMK), Diyarbakir, Turkey, 1–6.
[11] Handel, Peter (2018): *Understanding Normalized Mean Squared Error in Power Amplifier Linearization*, in: IEEE Microwave and Wireless Components Letters 28: 11, 1047–1049.
[12] Hasnain, Ahmad; Sheng, Yehua; Hashmi, Muhammad Zaffar; Bhatti, Uzair Aslam; Hussain, Aamir; Hameed, Mazhar; Marjan, Shah; Bazai, Sibghat Ullah; Hossain, Mohammad Amzad; Sahabuddin, Md; Wagan, Raja Asif; Zha, Yong (2022): *Time Series Analysis and Forecasting of Air Pollutants Based on Prophet Forecasting Model in Jiangsu Province, China*, in: Frontiers in Environmental Science 10, 945628.
[13] Heydarian, Mohammadreza; Doyle, Thomas E.; Samavi, Reza (2022): *MLCM: Multi-Label Confusion Matrix*, in: IEEE Access 10, 19083–19095.
[14] Hodson, Timothy O.; Over, Thomas M.; Foks, Sydney S. (2021): *Mean Squared Error, Deconstructed*, in: Journal of Advances in Modeling Earth Systems 13: 12, e2021MS002681.
[15] Huang, Sijia (2023): *Linear Regression Analysis*, in: International Encyclopedia of Education(Fourth Edition), 548–557.
[16] Jillian Mackenzie; Jeff Turintine (2023): *Air Pollution: Everything You Need to Know*.
[17] Karch, Julian (2020): *Improving on Adjusted R-Squared* Van Ravenzwaaij, Don; Van Ravenzwaaij, Don (Eds.), in: Collabra: Psychology 6: 1, 45.
[18] Khawaja, Yara; Shankar, Nathan; Qiqieh, Issa; Alzubi, Jafar; Alzubi, Omar; Nallakaruppan, M. K.; Padmanaban, Sanjeevikumar (2023): *Battery Management Solutions for Li-Ion Batteries Based on Artificial Intelligence*, in: Ain Shams Engineering Journal, 102213.
[19] Kumar, K.; Pande, B. P. (2023): *Air Pollution Prediction with Machine Learning: A Case Study of Indian Cities*, in: International Journal of Environmental Science and Technology 20: 5, 5333–5348.
[20] Luo, Haoran; Cheng, Fan; Yu, Heng; Yi, Yuqi (2021): *SDTR: Soft Decision Tree Regressor for Tabular Data*, in: IEEE Access 9, 55999–56011.
[21] Maduri, Praveen Kumar; Dhiman, Preeti; Chaturvedi, Chinmay; Rai, Abhishek (2023): *Air Pollution Index Prediction: A Machine Learning Approach*, in: Yadav, Sanjay; Haleem, Abid; Arora, P. K.; Kumar, Harish (Eds.): Proceedings of Second International Conference in Mechanical and Energy Technology, volume 290, Singapore, 37–51.

ISSN: 2304-1013 (Online); 2304-1269 (CDROM); 2414-6722 (Print)

[22] Méndez, Manuel; Merayo, Mercedes G.; Núñez, Manuel (2023): *Machine Learning Algorithms to Forecast Air Quality: A Survey*, in: Artificial Intelligence Review 56: 9, 10031–10066.
[23] Naim, Iram; Singh, Aanya Raj; Sen, Anjali; Sharma, Anurag; Mishra, Devesh (2022): *Healthcare CHATBOT for Diabetic Patients Using Classification*, in: Kumar, Rajesh; Ahn, Chang Wook; Sharma, Tarun K.; Verma, Om Prakash; Agarwal, Anand (Eds.): Soft Computing: Theories and Applications, volume 425, Singapore, 427–437.
[24] *Nationale Geographic* (n.d.):in: Air Pollution.
[25] Nduwayezu, Gilbert; Zhao, Pengxiang; Kagoyire, Clarisse; Eklund, Lina; Bizimana, Jean Pierre; Pilesjo, Petter; Mansourian, Ali (2023): *Understanding the Spatial Non-Stationarity in the Relationships between Malaria Incidence and Environmental Risk Factors Using Geographically Weighted Random Forest: A Case Study in Rwanda.*, in: Geospatial Health 18: 1.
[26] NIDULA ELGIRIYEWITHANA (2023): *World Weather Repository (Daily Updating)*.
[27] Patel, Shobhit K.; Surve, Jaymit; Katkar, Vijay; Parmar, Juveriya; Al-Zahrani, Fahad Ahmed; Ahmed, Kawsar; Bui, Francis Minhthang (2022): *Encoding and Tuning of THz Metasurface-Based Refractive Index Sensor With Behavior Prediction Using XGBoost Regressor*, in: IEEE Access 10, 24797–24814.
[28] Purbasari, R; Munajat, E; Fauzan, F (2023): *Digital Innovation in the Digital Innovation Ecosystem: A Digital Collaboration Networks Approach. Review of Integrative Business and Economics Research*, in: Review of Integrative Business and Economics Research 12: 3, 200–216.
[29] Rajat Lunawat (2022): *What Is 'Feels like' Temperature?*, in: Met office.
[30] Schonlau, Matthias; Zou, Rosie Yuyan (2020): *The Random Forest Algorithm for Statistical Learning*, in: The Stata Journal: Promoting communications on statistics and Stata 20: 1, 3–29.
[31] Singh, Pooja; Shamseldin, Asaad Y.; Melville, Bruce W.; Wotherspoon, Liam (2023): *Development of Statistical Downscaling Model Based on Volterra Series Realization, Principal Components and Ridge Regression*, in: Modeling Earth Systems and Environment 9: 3, 3361–3380.
[32] Sokhi, Ranjeet S.; Moussiopoulos, Nicolas; Baklanov, Alexander; Bartzis, John; Coll, Isabelle; Finardi, Sandro; Friedrich, Rainer; Geels, Camilla; Grönholm, Tiia; Halenka, Tomas; Ketzel, Matthias; Maragkidou, Androniki; Matthias, Volker; Moldanova, Jana; Ntziachristos, Leonidas; Schäfer, Klaus; Suppan, Peter; Tsegas, George; Carmichael, Greg; Franco, Vicente; Hanna, Steve; Jalkanen, Jukka-Pekka; Velders, Guus J. M.; Kukkonen, Jaakko (2022): *Advances in Air Quality Research – Current and Emerging Challenges*, in: Atmospheric Chemistry and Physics 22: 7, 4615–4703.
[33] Swathi, Kommana; Kodukula, Subrahmanyam (2022): *XGBoost Classifier with Hyperband Optimization for Cancer Prediction Based on Geneselection by Using Machine Learning Techniques*, in: Revue d'Intelligence Artificielle 36: 5, 665–670.
[34] Wichitaksorn, Nuttanan; Kang, Yingyue; Zhang, Faqiang (2023): *Random Feature Selection Using Random Subspace Logistic Regression*, in: Expert Systems with Applications 217, 119535.
[35] Yang, Yuting; Mei, Gang; Izzo, Stefano (2022): *Revealing Influence of Meteorological Conditions on Air Quality Prediction Using Explainable Deep Learning*, in: IEEE Access 10, 50755–50773.
[36] Yates, Luke A.; Aandahl, Zach; Richards, Shane A.; Brook, Barry W. (2023): *Cross Validation for Model Selection: A Review with Examples from Ecology*, in: Ecological Monographs 93: 1, e1557.

ISSN: 2304-1013 (Online); 2304-1269 (CDROM); 2414-6722 (Print)

[37] Zhang, Ke; Sun, Miao; Han, Tony X.; Yuan, Xingfang; Guo, Liru; Liu, Tao (2018): *Residual Networks of Residual Networks: Multilevel Residual Networks*, in: IEEE Transactions on Circuits and Systems for Video Technology 28: 6, 1303–1314.
[38] Zhu, Xu; Chu, Qingyong; Song, Xinchang; Hu, Ping; Peng, Lu (2023): *Explainable Prediction of Loan Default Based on Machine Learning Models*, in: Data Science and Management 6: 3, 123–133.

ISSN: 2304-1013 (Online); 2304-1269 (CDROM); 2414-6722 (Print)