MSW MANAGEMENT -Multidisciplinary, Scientific Work and Management Journal
ISSN: 1053-7899
Vol. 34 Issue 1, Jan-June 2024, Pages: 207-219

A Comparative Evaluation of Machine Learning Approaches for estimating Air Quality

Geeta Arneja
Department of Computer Science, Shyama Prasad Mukherji College for Women, University of Delhi, New Delhi, India

Sonalika Arneja
Beedie School of Business, Simon Frasor University, Vancouver, Canada

Abstract

Air quality is critically important for the purpose of preserving a fresh environment, preventing ailments, and ensuring good health. It describes the extent of air pollution or cleanliness, which is determined by the concentrations of hazardous compounds such as ozone, nitrogen dioxide, carbon monoxide and dust particles etc. Low air quality can result in serious illnesses like cardiovascular disease and problems with respiration, as well as premature deaths. Data analytics and machine learning methods have been employed to analyse the real time air quality data. However, real-world datasets provide a number of unique constraints, such as uneven data distributions, and an inadequate number of labelled samples. In this paper, we presented a comparison between machine learning (ML) models; nearest neighbour, support vector machines (SVM),, decision tree, logistic regression, naïve bayes and Adaboost to predict air quality category using a real time dataset available on Kaggle, while addressing different issues like data imbalance problem and fewer samples with labels. The current work handles these irregularities and generalize well while retaining high accuracy and efficiency. The findings reveal that the proposed model outperforms other models with a performance accuracy of 97%.

Keywords: Air Quality, Machine Learning, Data Balancing, SMOTE, Adaboost.

1. Introduction

Air pollution entails serious threats to health and obstacles to sustainability, making it an enormous global problem that influences both urban and rural areas. World Health Organisation (WHO) stated outside pollution as the second largest global cause of noncommunicable diseases (NCDs) and around 6.7 million deaths in 2019 were attributed to air pollution (World Health Organization, 2024). The government has established numerous regulations to reduce pollution from vehicles, industry, and other sources by keeping a check on the quality of the air. Despite this, maintaining air quality can be challenging at times the rules are not pursued precisely. .

Air quality is determined by the concentration of contaminants in the environment, including ozone, carbon monoxide and particulate matter (Liang et al., 2020; Ke et al., 2022). Its relevance stems from its immediate influence on human well-being, weather stability and environmental sustainability. The Air Quality Index (AQI) and individual pollutant concentrations are prime elements for monitoring and controlling the presence of harmful elements in air.AQI, an integrated metric used for evaluating air purity, includes several important contaminants monitored (Gupta et al., 2023). These are “ground-level ozone (O3), carbon monoxide (CO), sulphur dioxide (SO2), carbon dioxide (CO2), nitrogen

MSW MANAGEMENT -Multidisciplinary, Scientific Work and Management Journal
ISSN: 1053-7899
Vol. 34 Issue 1, Jan-June 2024, Pages: 207-219

dioxide (NO2) and particulate matter (PM10 and PM2.5)" (Gupta et al., 2023). The AQI measures present or expected air quality, with higher values indicating increased exposure hazards for the population. Higher values of pollutants such as PM2.5, NO2, and CO have caused an increase in breathing problems in workers and their families (Leong, Kelani, & Ahmad, 2020). These health problems are mostly caused by prolonged exposure to tiny particulate pollution.

Forecasting the concentrations of contaminants entails the use of a variety of machine learning models and datasets(Harishkumar, Yogesh, & Gad, 2020; Liang et al., 2020;Gladkova& Saychenko, 2022) Ensuring timely alerts depend on accurate prediction of air quality metrics. This supports improved public health and environmental decision-making by helping to estimate pollution levels and trends in air quality. The competency of machine learning techniques to adapt to substantial and diverse datasets makes them apt for producing precise forecasts of air quality.

In current research, we have performed an analysis of machine learning methods for evaluating their performance on a Kaggle dataset to estimate category of quality of air. Our proposed model surpassed other models by achieving highest level of accuracy. The originality of our work is in applying a three -phase strategy; feature selection for extracting significant features followed by applying the "Synthetic Minority Over-sampling Technique(SMOTE)" to neutralize real-world data imbalance, thereafter using hyperparameter tuning to attain optimal outcomes. We also tackled the rare cases of very few labelled samples by manual oversampling. Finally, the performance of the suggested model is confirmed through a comprehensive experimental investigation utilizing air pollution data, evaluating its accuracy.

Further, Section II examines prior studies in view of relevant works. Section III describes the procedures utilized in the methodology. Section IV contains the empirical analysis and discussion of the findings. Conclusion has been given in Section V.

2. Related Works

Assessing air pollution becomes essential for monitoring pollution levels in different regions. Machine learning can be utilized for estimating and forecasting pollution levels in air. An automatic machine learning system for air quality prediction estimates levels of "six major pollutants: SO2, NO2, O3, CO, PM2.5 and PM10" (Ke et al., 2022). The particulate matter PM2.5, a key pollutant that accounts for air quality has been predicted using machine learning models based on Taiwan dataset and findings proved gradient boosting model as best in this case (Harishkumar, Yogesh, & Gad, 2020). Liang et al. (2020) demonstrates that stacking ensemble and AdaBoost perform well in AQI forecasting, with regional performance differences, using 11 years of Taiwanese data.

Furthermore, a novel method for measuring the AQI in Nanjing areas without monitoring stations is introduced. It uses a Back Propagation neural network for spatial forecasting and an improved k-Nearest Neighbor method for time-based forecasts, with an average accuracy of 73.8%(Zhao et al., 2020). An ensemble machine learning method shows how the cloud model manages the uncertainty and unpredictability in estimates of air quality by forecasting AQI using Sparse Spectrum GPR and evaluating air quality (Hardini et al., 2023).

MSW MANAGEMENT -Multidisciplinary, Scientific Work and Management Journal
ISSN: 1053-7899
Vol. 34 Issue 1, Jan-June 2024, Pages: 207-219

In another study, a new ensemble model has been presented for multivariate time series forecasting: the Geographically Weighted Predictor (GWP) technique. To forecast hourly air pollution and AQI in places without measuring tools, it combines "Random Forest, eXtreme Gradient Boosting, and Deep Neural Networks" (Phruksahiran, 2021). According to Wang et al. (2020), Gas Recurrent Neural Networks (GRNNs) are more effective at regulating sensor drift, although being more sensitive to variations in humidity. Thus, when it comes to air quality prediction, GRNNs outperform Multi-Layer Perceptrons (MLPs) and Support Vector Regression (SVR).

A hybrid approach that combines the powers of "autoregressive moving average (ARMA), long short-term memory (LSTM), and wavelet decomposition (WD)" for foreseeing air quality is presented by Fan et al., (2021) which is more accurate as compare to standard conventional models. Another model integrates "input variable selection (IVS), machine learning (ML), and regression techniques" to estimate the level of pollutants and achieves "greater R^2 and a small root mean squared error (RMSE)" (Udristioiu, Mghouchi, &Yildizhan, 2023). Further studies must use real-time data to improve predicting accuracy and effectiveness along with considering computational issues

3. Proposed Methodology

The current research provides an effective approach for predicting air quality category in cases where high class imbalance with few labelled samples is present. To deal with the skewness caused, we applied manual oversampling for rare cases where labelled samples were less than 100. Furthermore, "Synthetic Minority Over-sampling Technique (SMOTE)" has been applied to balance dataset to improve the performance of machine learning models that might otherwise have very large observations belonging to majority class. The balanced dataset thus undergoes classification task to predict air quality category. In current study, we proposed an Adaboost classifier followed by hyperparameter tuning on a balanced dataset that has been evaluated against other classifiers. A complete architecture used in the proposed work has been shown in Figure 1.

3.1. Dataset

In this proposed model, we worked on a Kaggle dataset (Kaggle, 2023), the 2023 air quality statistics for Core-Based Statistical Areas (CBSAs) which include specific pollutant measures that can be used to assess regional air quality changes and guide public health and policy decisions. The dataset used has been summarized in Table 1. It has 10923 observations that each provide the AQI values and categories for Core-Based Statistical Area (CBSA) is in the United States. These observations include city and state data along with measurements of "AQI, CO, PM2.5, PM10, NO_2 and ozone" for year 2023.Six different categories have been defined on the basis of AQI values.

MSW MANAGEMENT -Multidisciplinary, Scientific Work and Management Journal
ISSN: 1053-7899
Vol. 34 Issue 1, Jan-June 2024, Pages: 207-219

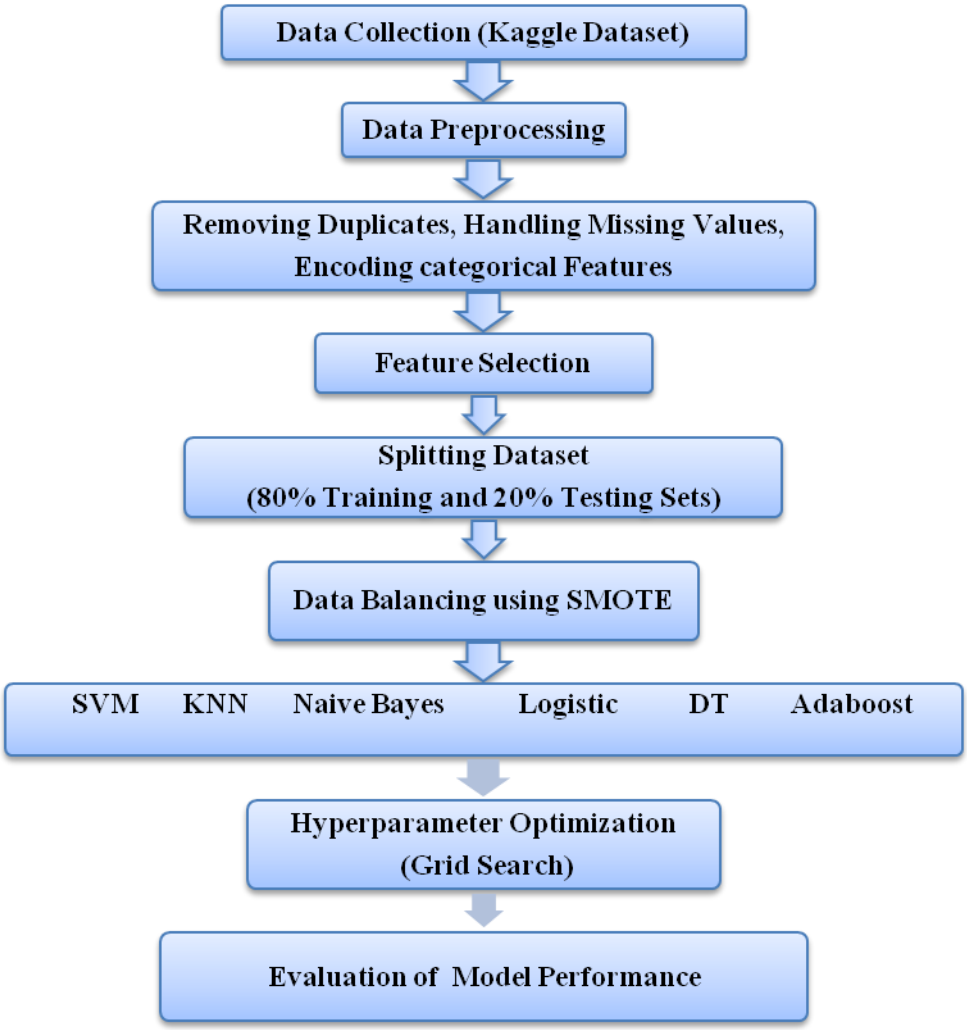

Figure 1 The Overall Architecture of AQ Category Prediction

A score of "Good" (0–50) indicates that everyone can breathe well. "Moderate" (51–100) is considered acceptable. According to "Unhealthy for Sensitive Groups" (101–150), aged and vulnerable individuals may be affected badly. "Unhealthy" (151-200) rating is quite harmful so health risks may arise. An excessive risk to one's health is indicated by the "Very Unhealthy" (201-300) rating, which suggests staying indoors. Lastly, "Hazardous" (301–500) denotes extremely high pollution levels that are extremely risky to health; individuals should heed health advisories and stay indoors.

MSW MANAGEMENT -Multidisciplinary, Scientific Work and Management Journal
ISSN: 1053-7899
Vol. 34 Issue 1, Jan-June 2024, Pages: 207-219

Table 1 2023 Air Quality Dataset for CBSAs

Dataset Summary	
Time Window	Year 2023
Total Number of observations	10923
Total number of features	15
Number of observations (AQ Category-wise) 1. "Good (0-50)" 2. "Moderate (51-100)" 3. "Unhealthy for sensitive groups (101-150)" 4. "Unhealthy (151-200)" 5. "Very Unhealthy (201-300)" 6. "Hazardous (301-500)"	 6321 3956 129 502 14 1

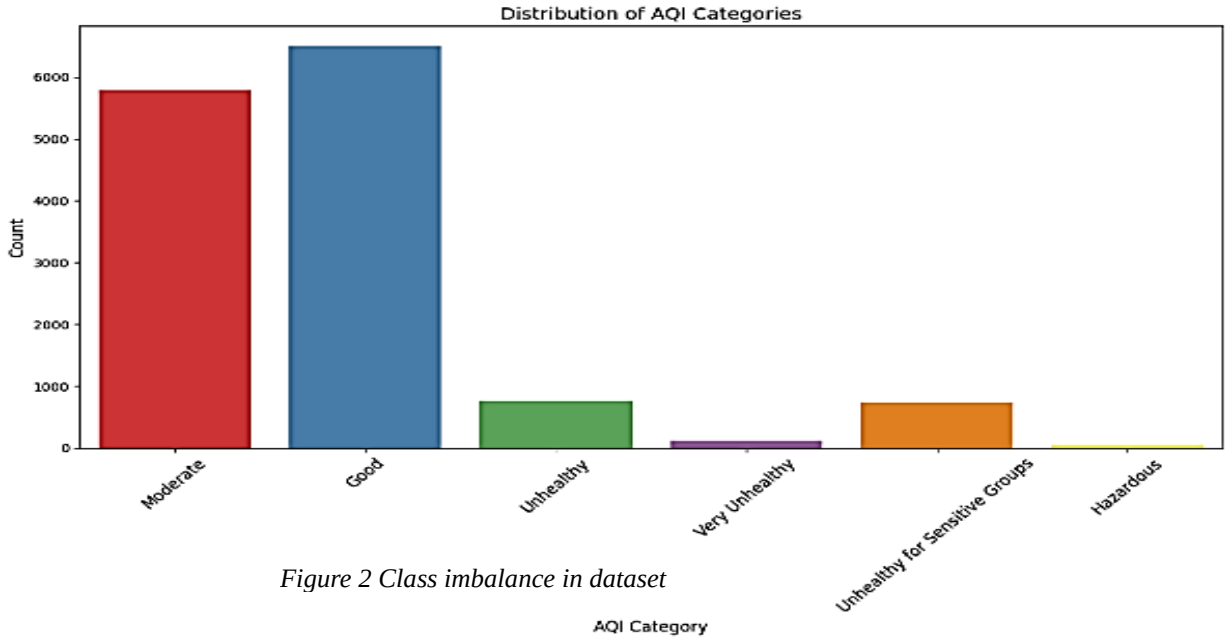

Figure 2 Class imbalance in dataset

A high data imbalance can be viewed in Figure 2. The dataset contains two rare AQ categories; "Very Unhealthy" and "Hazardous" having very few observations in contrast to other categories causing exceptional skewness. Apart from this, other classes also show a high data imbalance. This mismatch emphasizes the importance of specialized preprocessing strategies to accommodate the disparity in data distribution. Figure 3 shows the date-wise trend of overall AQI values.

MSW MANAGEMENT -Multidisciplinary, Scientific Work and Management Journal
ISSN: 1053-7899
Vol. 34 Issue 1, Jan-June 2024, Pages: 207-219

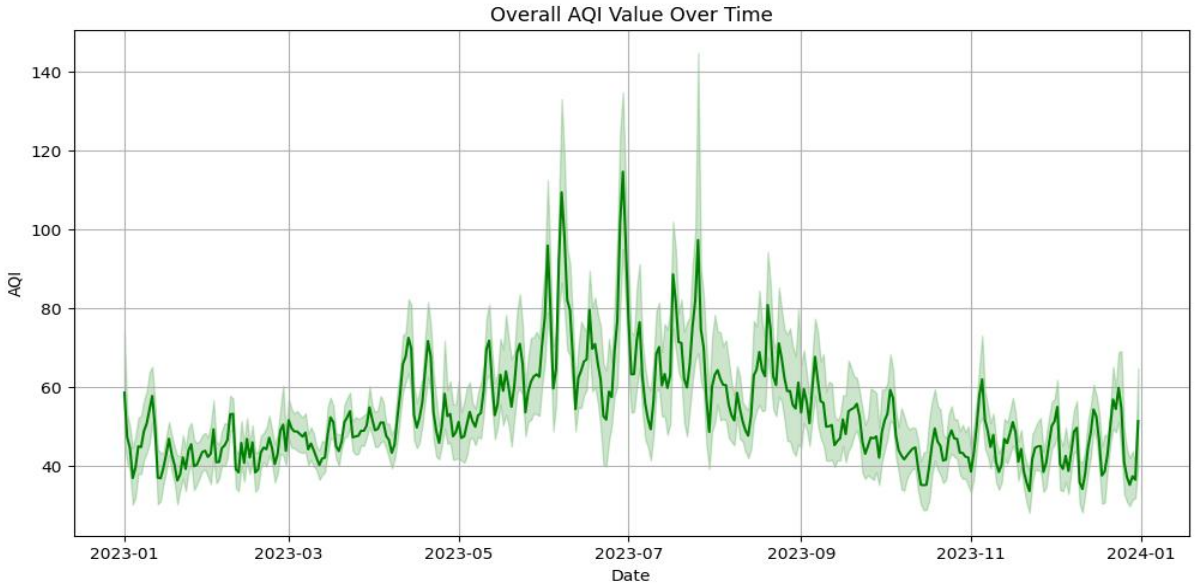

Figure 2 Overall AQI in core-based statistical area in U.S (2023).

3.2. Data Preprocessing

Data preprocessing standardizes the structure of data in the first phase, making it ready for training. This ensures the data is appropriate for the suggested model by filling in missing values, and normalizing the data. The dataset contains missing values for Pm2.5, PM10, CO, NO_2 in some observations. In this phase, missing values have been filled using imputation. Furthermore, categorical variables have been transformed into numeric format using Label Encoders as many machine learning algorithms require numerical input.

To guarantee equal impact of every attribute to the model, numerical features have been transformed by "standard scaling (Z-score normalization)" (Hastie, Tibshirani, & Friedman, 2023). The scaling formula as in Eq. (1), that is, "

$$Xs = \frac{X-\mu}{\sigma} \quad (1)$$

where:

- X_s is value of scaled feature
- X is value of initial feature
- μ is mean of all values of feature.
- σ is standard deviation of the feature."

is commonly used to standardize data (Hastie, Tibshirani, & Friedman, 2023).

MSW MANAGEMENT -Multidisciplinary, Scientific Work and Management Journal
ISSN: 1053-7899
Vol. 34 Issue 1, Jan-June 2024, Pages: 207-219

3.3. Feature Selection and Data Splitting

Feature selection is a critical phase in machine learning which requires identifying the most significant characteristics from the dataset to improve model efficiency and accuracy. It streamlines the model by minimizing the total number of input features which improves accuracy, prevents overfitting, and accelerates training process (Chen, Dewi, Huang, & Caraka, 2020). As in our dataset, the most significant features that contribute to Air quality are, PM2.5, PM10, ozone, NO_2 and CO as shown in Figure 4.Rest of the features have not been used in the training process. Feature selection increases model correctness, improves model interpretability, and minimizes computing requirements thus making it an important stage in the data preparation process.

An essential aspect of machine learning that promises efficient model training is dataset splitting. Making subsets of the data for testing and training is the primary aim of this approach. The ratios used to separate training and testing data include 80/20, 70/30, and 90/10, depending on the nature of the problem. We have utilized 80/20 ratio of the dataset for training/testing purpose in our study.

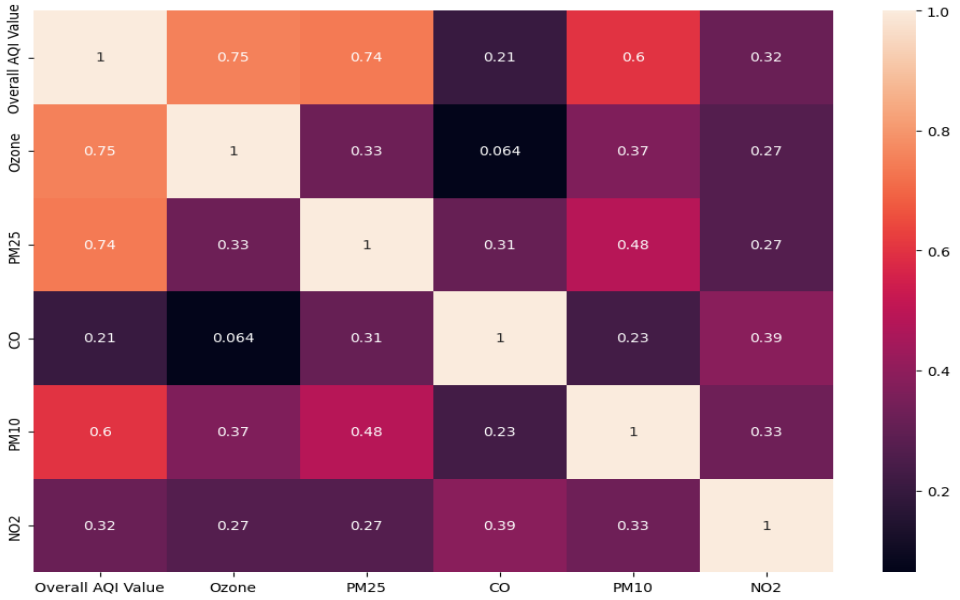

Figure 3 Correlation between Significant Features and Air Quality

3.4. Classification using Machine learning models

We examined six different algorithms—"Support Vector Machine (SVM), K-Nearest Neighbours (KNN), Naive Bayes (NB), Decision Tree (DT), Logistic Regression (LR) and Adaboost" in our assessment of several machine learning models for classification task. To evaluate each model's effectiveness and performance, it was tested in a variety of scenarios, including those with and without SMOTE and hyperparameter adjustment.

MSW MANAGEMENT -Multidisciplinary, Scientific Work and Management Journal
ISSN: 1053-7899
Vol. 34 Issue 1, Jan-June 2024, Pages: 207-219

3.5. Data Balancing Using SMOTE

SMOTE (Synthetic Minority Over-sampling Technique) is an approach that induces synthetic data samples for the minority class in disparate datasets in order to boost machine learning model efficacy (Pradipta et al., 2021).In the case of multiclass identification, SMOTE can be modified to properly handle multiple classes. SMOTE generates synthetic samples for all deficient classes in the dataset, rather than just one. In present work we accomplished data balancing by SMOTE, which generates new samples by applying interpolation to between existing instances in the class (Mduma, 2023). In multiclass contexts, SMOTE assures that each minority class is balanced against the majority classes, which improves overall model performance and reduces class bias.

3.6. Hyperparameter Tuning (HT)

Hyperparameter tweaking is used in machine learning to boost performance of models (Yang & Shami, 2020). It involves choosing which combination of parameters best regulates the training process. In current study, after balancing the dataset we performed hyperparameter optimization using Grid search. It is a powerful technique that tests a predetermined set of hyperparameters. With this method, a predetermined set of hyperparameter variables are rigorously investigated. Grid search examines every possibility in detail by creating a grid with possible hyperparameter combinations (Zahedi et al., 2021). It works particularly well in situations when the hyperparameter space is possible and compact. Grid search is frequently combined with cross-validation to increase reliability and reduce overfitting. GridSearchCV employ "5-fold cross-validation" by default.

4. Results and Discussion

This section evaluates the proposed model's performance using realistic Air quality dataset from Kaggle (Table 1). A comprehensive analysis of accuracy performance of several ML algorithms such as "SVM, KNN, Naive Bayes (NB), Logistic Regression, Decision Tree (DT), and AdaBoost" reveals interesting patterns. This evaluation is carried out in a Python 3.9 environment, with all models assessed against the Kaggle dataset (Kaggle,2023) to assess their performance characteristics. We use accuracy as performance metrics in our current work. Accuracy is measured as a ratio of accurately categorized things to the overall number of items. As stated by Erickson and Kitamura (2021), the formula for calculating accuracy is given as "

$$\text{Accuracy} = \frac{TP+TN}{TP+FP+TN+FN}$$

where:

TP (True Positives) is the number of truly reported positive cases.

TN (True Negatives) is the number of truly reported negative cases.

FP (False Positives) is the number of falsely reported positive cases.

FN (False Negatives) is the number of falsely reported negative cases."

MSW MANAGEMENT -Multidisciplinary, Scientific Work and Management Journal
ISSN: 1053-7899
Vol. 34 Issue 1, Jan-June 2024, Pages: 207-219

The findings of our experiment have been shown in Table 2,The outcomes of the SVM and KNN models is comparable, with SVM marginally outperforming KNN when hyperparameter adjustment is used. With hyperparameter optimization, both models reach a 96% accuracy rate; however, when SMOTE is included. Regardless of whether SMOTE or hyperparameter tuning is used, the Decision Tree (DT) model continuously obtains the greatest accuracy of 96% across all scenarios, proving its stability and durability in classification tasks. Naive Bayes (NB), on the other hand, exhibits lesser accuracy, ranging from 85% to 87%, suggesting that it might have trouble processing the complicated data or that it might not be the ideal option for this classification task.

AdaBoost achieves the maximum accuracy of 97% when both SMOTE and hyperparameter tuning are used, suggesting its greater capacity to handle unbalanced data and benefit from fine-tuning. The best parameters for adaboost have been found to be “algorithm: SAMME; max_depth: 3; learning_rate: 1; n_estimators: 200 ” using decision tree as estimator. This shows that AdaBoost's performance is significantly improved by these strategies, making it the most robust model in this comparison. All the experimental results have been shown in Figure 4-7 graphically.

Table 2 Comparative analysis of ML models for predicting Air Quality

Model	Accuracy (Without Smote + Without hyperparameter tuning)	Accuracy (Without Smote + With hyperparameter Tuning)	Accuracy (With Smote + Without hyperparameter tuning)	Accuracy (with smote + with hyperparameter tuning
SVM	95	96	92	95
KNN	95	96	93	94
NB	87	87	85	86
Logistic	89	89	86	85
DT	96	96	96	96
Adaboost	88	90	93	97

MSW MANAGEMENT -Multidisciplinary, Scientific Work and Management Journal
ISSN: 1053-7899
Vol. 34 Issue 1, Jan-June 2024, Pages: 207-219

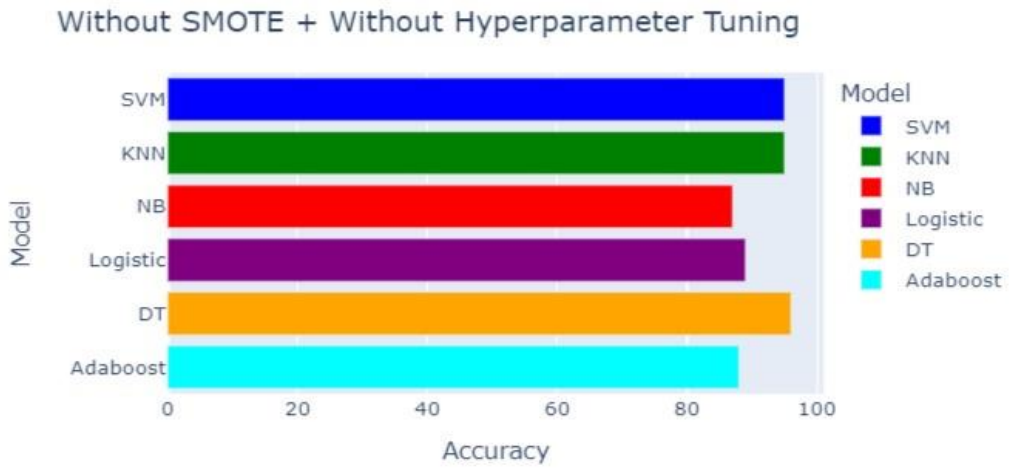

Figure 4Comparison of Accuracies of ML models (without Smote and HT)

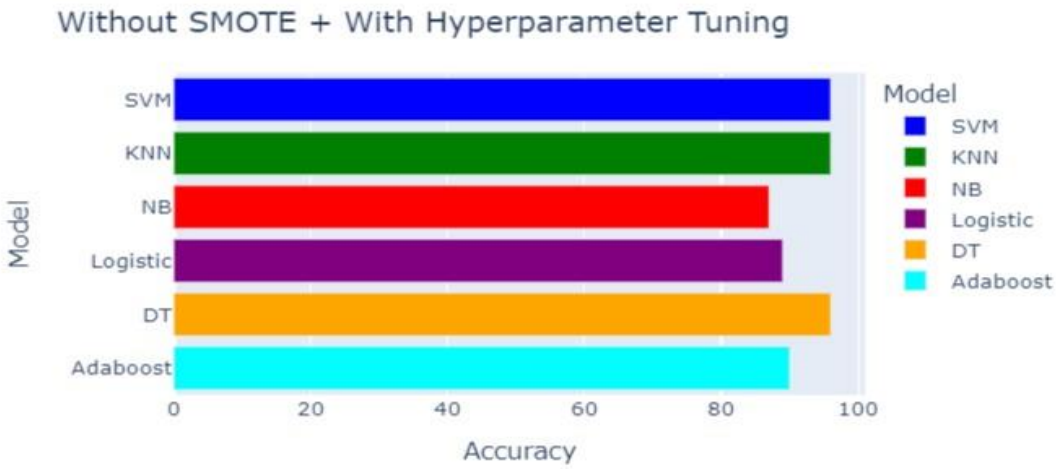

Figure 5Comparison of Accuracies of ML models (without Smote and with HT)

MSW MANAGEMENT -Multidisciplinary, Scientific Work and Management Journal
ISSN: 1053-7899
Vol. 34 Issue 1, Jan-June 2024, Pages: 207-219

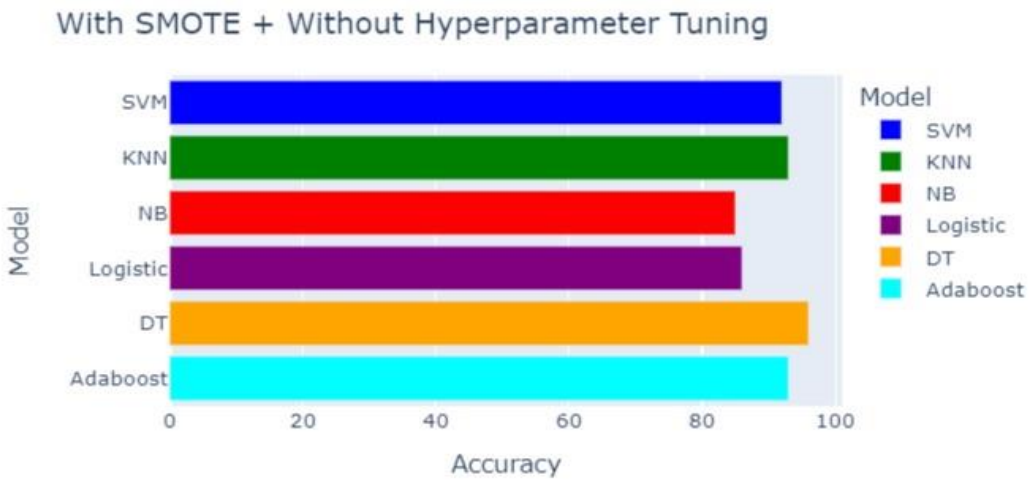

Figure 6Comparison of Accuracies of ML models (with Smote and without HT)

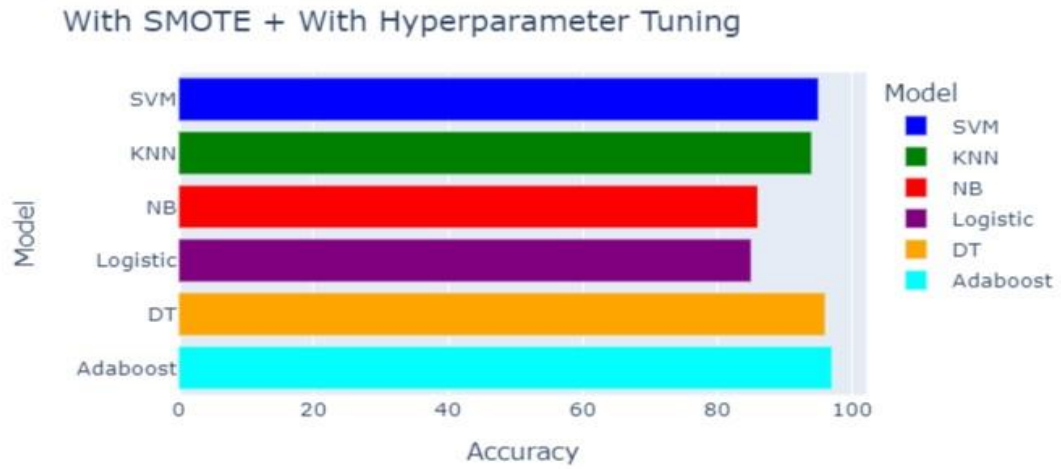

Figure 7Comparison of Accuracies of ML models (with Smote and HT)

5. Conclusion

In our research, the proposed model incorporates "SMOTE" for data balancing and "AdaBoost classifier with hyperparameter tuning" to predict air quality category.
Among the six models examined, the Support Vector Machine (SVM) performed well but had drawbacks when the hyperparameters were not tweaked. The K-Nearest Neighbors (KNN) model produced consistent results, particularly after hyperparameter adjustment, although its efficacy with SMOTE was less obvious. Across different settings, the Naive Bayes (NB) model remained consistent but had lesser accuracy. Decision Trees (DT) demonstrated great accuracy with clearly stated decision rules, but they failed to surpass other models. Logistic Regression delivered consistent results, although it lagged behind more advanced models. AdaBoost, on the other hand, outperformed other alternatives, earning the highest accuracy of 97% across all situations, including those with hyperparameter adjustment and SMOTE.

MSW MANAGEMENT -Multidisciplinary, Scientific Work and Management Journal
ISSN: 1053-7899
Vol. 34 Issue 1, Jan-June 2024, Pages: 207-219

References

Chen, R. C., Dewi, C., Huang, S. W., & Caraka, R. E. (2020). Selecting critical features for data classification based on machine learning methods. Journal of Big Data, 7(1), 52.

Erickson, B. J., & Kitamura, F. (2021). Magician's corner: 9. Performance metrics for machine learning models. Radiology: Artificial Intelligence, 3(3), e200126.https://doi.org/10.1148/ryai.2021200126

Fan, S., Hao, D., Feng, Y., Xia, K., & Yang, W. (2021). A hybrid model for air quality prediction based on data decomposition. *Information*, *12*(5), 210. https://doi.org/10.3390/info12050210

Gladkova, E., & Saychenko, L. (2022). Applying machine learning techniques in air quality prediction. *Transportation Research Procedia*, *63*, 1999-2006.

Gupta, N. S., Mohta, Y., Heda, K., Armaan, R., Valarmathi, B., & Arulkumaran, G. (2023). Prediction of air quality index using machine learning techniques: a comparative analysis. *Journal of Environmental and Public Health*, *2023*(1), 4916267.

Hardini, M., Sunarjo, R. A., Asfi, M., Chakim, M. H. R., & Sanjaya, Y. P. A. (2023). Predicting air quality index using ensemble machine learning. *ADI Journal on Recent Innovation*, *5*(1Sp), 78-86. https://doi.org/10.34306/ajri.v5i1Sp.981

Harishkumar, K. S., Yogesh, K. M., & Gad, I. (2020). Forecasting air pollution particulate matter (PM2. 5) using machine learning regression models. *Procedia Computer Science*, *171*, 2057-2066.

Hastie, T., Tibshirani, R., Friedman, J. H., & Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction* (Vol. 2, pp. 1-758). New York: springer.

Kaggle. (2023). Air quality dataset.https://www.kaggle.com/datasets/nikkiperry/2023-air-quality-data-for-cbsas

Ke, H., Gong, S., He, J., Zhang, L., Cui, B., Wang, Y., Mo, J., Zhou, Y., & Zhang, H. (2022). Development and application of an automated air quality forecasting system based on machine learning. *Science of The Total Environment*, *806*, 151204. https://doi.org/10.1016/j.scitotenv.2021.151204

Liang, Y. C., Maimury, Y., Chen, A. H. L., & Juarez, J. R. C. (2020). Machine learning-based prediction of air quality. *applied sciences*, *10*(24), 9151. https://doi.org/10.3390/app10249151

Leong, W. C., Kelani, R. O., & Ahmad, Z. J. J. O. E. C. E. (2020). Prediction of air pollution index (API) using support vector machine (SVM). *Journal of Environmental Chemical Engineering*, *8*(3), 103208. http://dx.doi.org/10.1016/j.jece.2019.103208.

Mduma, N. (2023). Data balancing techniques for predicting student dropout using machine learning. *Data*, *8*(3), 49. https://doi.org/10.3390/data8030049

Méndez, M., Merayo, M. G., & Núñez, M. (2023). Machine learning algorithms to forecast air quality: a survey. *Artificial Intelligence Review*, *56*(9), 10031-10066.

Phruksahiran, N. (2021). Improvement of air quality index prediction using geographically weighted predictor methodology. *Urban Climate*, *38*, 100890. https://doi.org/10.1016/j.uclim.2021.100890

Pradipta, G. A., Wardoyo, R., Musdholifah, A., Sanjaya, I. N. H., & Ismail, M. (2021, November). SMOTE for handling imbalanced data problem: A review. In *2021 sixth international conference on informatics and computing (ICIC)* (pp. 1-8). IEEE.

Udristioiu, M. T., Mghouchi, Y. E., &Yildizhan, H. (2023). Prediction, modelling, and forecasting of PM and AQI using hybrid machine learning. *Journal of Cleaner Production*, *421*, 138496. https://doi.org/10.1016/j.jclepro.2023.138496

Wang, S., Hu, Y., Burgués, J., Marco, S., & Liu, S. C. (2020, August). Prediction of gas concentration using gated recurrent neural networks. In *2020 2nd IEEE International Conference on Artificial Intelligence Circuits and Systems (AICAS)* (pp. 178-182). IEEE.

World Health Organization. (2024, June 25). What are health consequences of air pollution on populations? [Press release]. World Health Organization. https://www.who.int/news/item/25-06-2024-what-are-health-consequences-of-air-pollution-on-populations

Yang, L., & Shami, A. (2020). On hyperparameter optimization of machine learning algorithms: Theory and practice. *Neurocomputing*, *415*, 295-316.

MSW MANAGEMENT -Multidisciplinary, Scientific Work and Management Journal
ISSN: 1053-7899
Vol. 34 Issue 1, Jan-June 2024, Pages: 207-219

Zahedi, L., Mohammadi, F. G., Rezapour, S., Ohland, M. W., & Amini, M. H. (2021). Search algorithms for automated hyper-parameter tuning. *arXiv preprint arXiv:2104.14677*.

Zhao, X., Song, M., Liu, A., Wang, Y., Wang, T., & Cao, J. (2020). Data-driven temporal-spatial model for the prediction of AQI in Nanjing. *Journal of Artificial Intelligence and Soft Computing Research*, *10*(4), 255-270.