Predicting Respiratory Diseases Attributed to PM2.5 Air Pollution in Nairobi County Using Random Forest Model

Valine Atieno Okeyo[1]*
[1]University of Nairobi, Department of Mathematics, Kenya

Idah Orowe[2]*
[2]University of Nairobi, Department of Mathematics, Kenya

Nicholas Otienoh Oguge[3]*
[3]University of Nairobi, Center for Advanced Studies in Environmental Law and Policy, Kenya

Correspondence Author:- Valine Atieno Okeyo[1]*; Idah Orowe[2]*; Nicholas Otienoh Oguge[3]*

Abstract:- This study investigates the predictive capability of a Random Forest model in identifying respiratory diseases attributed to PM2.5 exposure in Nairobi County. Leveraging a comprehensive dataset encompassing demographic and air quality variables, the model demonstrated robust performance metrics, achieving an accuracy of 79.97% and an area under the curve (AUC) of 0.872. These results highlight the model's effectiveness in distinguishing between respiratory and cardiovascular conditions. The model's sensitivity and specificity were 81.88% and 73.27%, respectively, indicating a strong ability to correctly identify both true positives and true negatives. Analysis of feature importance revealed that age and PM2.5 concentrations were the most influential factors in predicting health outcomes, emphasizing the significant impact of air pollution and demographic factors on respiratory and cardiovascular health. Furthermore, the consistent train and test error rates across varying training set sizes suggest the model's stability and generalizability. This study underscores the importance of addressing air quality issues to mitigate the health impacts of PM2.5 exposure in urban settings.

Keywords:- *Respiratory Diseases, PM2.5, Random Forest, Accuracy, Feature Importance.*

I. INTRODUCTION

Air pollution, particularly fine particulate matter (PM2.5), is a critical environmental and public health concern worldwide. Nairobi, the capital city of Kenya, is undergoing rapid urbanization and industrialization, contributing to worsening air quality. The city's population has surged in recent decades, leading to increased motor vehicle emissions, construction activities, and industrial operations. Respiratory diseases are already a significant burden, and the additional strain from pollution-related health issues poses a challenge to the healthcare system. To address these challenges, there is an urgent need to develop a robust predictive model that can bridge existing gaps by providing timely insights into potential health risks and enabling proactive measures. Machine learning techniques, particularly the Random Forest algorithm, offer a promising approach to addressing these challenges. Random Forest algorithm is a versatile and powerful tool for predictive modeling, capable of handling complex datasets with numerous variables. It works by constructing multiple decision trees during training and outputting the mode of the classes for classification tasks or the mean prediction for regression tasks. The utilization of machine learning techniques for predictive analytics in the context of respiratory diseases and PM2.5 air pollution represents a novel approach to public health research. This study therefore aimed at developing a predictive model using random forest algorithm to forecast respiratory diseases attributed to PM2.5 exposure in Nairobi, Kenya.

II. MATERIALS AND METHODS

To achieve the objectives of this study, three-year data spanning from 2021 to 2023 of hospital data and PM2.5 data were collected from the East Africa Global Environmental and Occupational Health Research and Training Center. Monthly health records from various files were consolidated into a single folder. Similarly, daily PM2.5 records were gathered into another folder. These datasets were then individually imported into R, explored, and subsequently merged. Additionally, descriptive statistics was performed by examining the summary statistics of each variable and visualizing the data distributions to gain an initial understanding of the data. Feature selection played a significant role in improving the performance of machine learning algorithms by reducing the time to build the learning model and increasing the accuracy of the learning process. Out of the 16 features initially considered, five features were selected to train the model: Real-time PM2.5 Concentrations , Hourly PM2.5 Concentrations, Age, Sex, and Diagnosis. Addressing potential class imbalances in respiratory disease data ensured that the model learns from a representative dataset, minimizing bias towards the majority class and

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24JUL1521

improving overall predictive performance. The Random Over-Sampling technique was applied to randomly synthesize new examples by interpolating from the minority class to balance the class distribution. The dataset was partitioned into a 70% training set and a 30% test set to train the model on a sufficient amount of data and evaluate its performance on unseen data. Various evaluation metrics were employed to assess the model's performance. Confusion matrix provided a detailed breakdown of true positive, true negative, false positive, and false negative predictions, offering insights into the model's strengths and weaknesses across different classes. While Random Forest excels in predictive accuracy, efforts were made to interpret feature importance rankings derived from the model. This analysis elucidates which factors, such as PM2.5 concentrations or demographic variables, exert the most significant influence on respiratory health outcomes in Nairobi County.

III. RESULT AND DISCUSSION

➢ *Confusion Matrix and Statistics:*

Table 1 Model Performance Summary

Metric	Accuracy	Sensitivity/Recall	Specificity	Precision	Balanced Accuracy	F1 Score	AUC Score
Value	0.7997	0.8188	0.7327	0.9148961	0.7757	0.86418	0.87196

Table 2 Feature Ranking

Feature	Mean Decrease Gini
ConcRT.ug.m3	13217.7847
ConcHR.ug.m3	13285.0345
AGE	45895.2029
SEX	838.4769

➢ *Learning Curve*

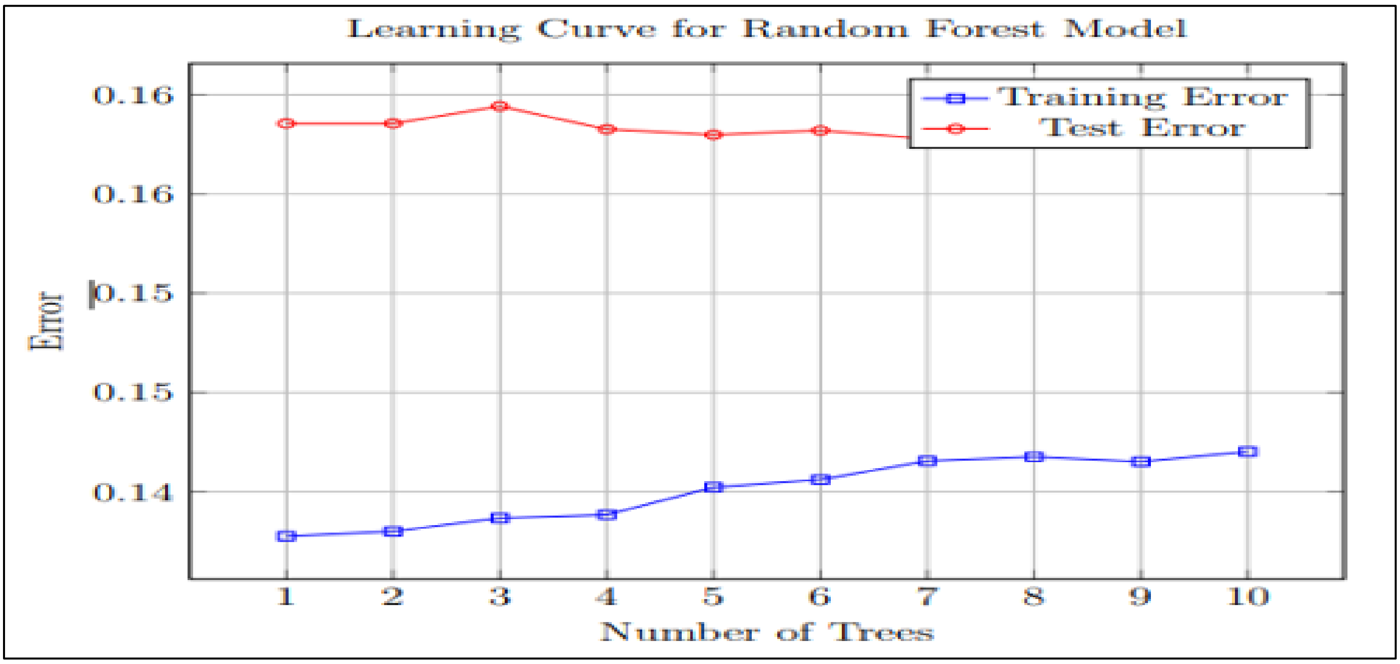

Fig 1 Learning Curve for Random Forest Model

➢ *Confusion Matrix Analysis*

The confusion matrix provides a detailed breakdown of the model's predictions compared to the actual outcomes. From the confusion matrix:

- **True Positives (Respiratory):** 40,002 cases
- **False Positives (Respiratory):** 3,721 cases
- **True Negatives (Cardiovascular):** 10,200 cases
- **False Negatives (Cardiovascular):** 8,853 cases

➢ *Model Accuracy*

The model's accuracy is a measure of its overall ability to correctly classify cases.

- **Accuracy**: 79.97%
- **95% Confidence Interval (CI)**: (0.7965, 0.8028)
- **No Information Rate (NIR):** 77.82%
- **P-Value [Acc > NIR]**: < 2.2e-16

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24JUL1521

This indicates that the model performs significantly better than random guessing.

➢ *Sensitivity and Specificity*

Sensitivity and specificity assess the model's performance in detecting positive and negative cases, respectively.

- **Sensitivity (Recall) for Respiratory Diseases: 81.88%**
- **Specificity for Cardiovascular Diseases: 73.27%**

The high sensitivity indicates the model's strong ability to identify true positive cases of respiratory diseases, while the specificity shows a moderate ability to correctly classify cardiovascular cases.

➢ *Precision and F1 Score*

Precision and F1 score provide insights into the balance between the model's accuracy in identifying positive cases and its overall performance:

- **Precision: 91.49%**
- **Recall (Sensitivity): 81.88%**
- **F1 Score: 86.42%**

The high precision and F1 score reflect the model's effectiveness in correctly identifying positive cases and balancing precision and recall.

➢ *Model Error Rates*

The model's error rates are consistent, reflecting its robustness and reliability:

- **Training Error: Ranged from 13.77% to 14.20%**
- **Test Error: Ranged from 15.75% to 15.91%**

The small difference between training and test error rates indicates good generalization capability, with minimal overfitting.

IV. CONCLUSIONS

By demonstrating the effectiveness of machine learning algorithms, notably Random Forest, in predicting respiratory disease outbreaks in relation to PM2.5 air pollution levels, this study contributes to evidence-based health and environmental policy-making. Age and PM2.5 concentrations were identified as the most significant predictors of respiratory disease outcomes. The prominence of age as a critical feature suggests that younger populations are more vulnerable to respiratory diseases in the context of PM2.5 pollution. This study therefore, recommends targeted health interventions for younger populations, who are identified as more susceptible to respiratory diseases. We also encourage policies that promote cleaner technologies and reduce pollution from major sources such as traffic and industrial activities and utilization of predictive models to forecast high-risk periods and inform the public and healthcare providers in advance. While this study provides valuable insights, several areas warrant further investigation. Future studies should consider incorporating more comprehensive datasets, including socio-economic status, lifestyle factors, and genetic predispositions, to improve model accuracy. Additionally, exploring and comparing different machine learning models can help identify the most efficient and accurate approaches for health outcome predictions.

ACKNOWLEDGEMENT

This research was funded in part by the Advancing Public Health Research in Eastern Africa through Data Science Training (APHREA-DST) project (Grant No. 1U2RTW012123), with support from the U.S. National Institutes of Health (NIH)

REFERENCES

[1]. Dockery DW, Pope CA. Acute respiratory effects of particulate air pollution. Rev Public Health. 1993;15(1):107-32.
[2]. Pope CA, Burnett RT, Thun MJ, Calle EE, Krewski D, Ito K, et al. Lung cancer, cardiopulmonary mortality, and long-term exposure to fine particulate air pollution. JAMA. 2002;287(9):1132-41.
[3]. Health Effects Institute (HEI). Understanding the health effects of ambient ultrafine particles. Research Report 155. Boston, MA: Health Effects Institute; 2010.
[4]. Amegah AK, Agyei-Mensah S. Urban air pollution in sub-Saharan Africa: Time for action. Environ Pollut. 2017;220:738-43.
[5]. European Environment Agency (EEA). Air quality in Europe - 2020 report. Copenhagen, Denmark: European Environment Agency; 2020.
[6]. Kanyiva KW, Mwalukumbi JM, Chege W, Juma PA, Mutemi J. Air quality in Nairobi, Kenya: A review of monitoring and policy gaps. Atmosphere. 2021;12(4):508.
[7]. Githinji G, Wanyua J, Karu J, Muchiri EM. Assessment of ambient air quality and its health impact in Nairobi City, Kenya. Int J Environ Res Public Health. 2019;16(11):1987.
[8]. Breiman L. Random forests. Mach Learn. 2001;45(1):5-32.
[9]. Lall R, Kendall M, Zhao Y, Wesson B, Harlan S, Jones M. Machine learning approaches for estimating spatial PM2.5 concentrations across the continental United States. Environ Sci Technol. 2017;51(21):12449-58.
[10]. Hu X, Waller LA, Al-Hamdan MZ, Crosson WL, Estes MG Jr, Estes SM, et al. A systematic review of machine learning applications in air quality research. Environ Res Lett. 2020;15(6):063001.
[11]. World Health Organization. Air pollution. Available from: https://www.who.int/airpollution. 2018.
[12]. Liu Y, Chen X, Yan B. The impact of PM2.5 on respiratory diseases: Evidence from hospital admissions in China. J Environ Manag. 2020;274:111214.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24JUL1521

[13]. Anderson JO, Thundiyil JG, Stolbach A. Clearing the air: A review of the effects of particulate matter air pollution on human health. J Med Toxicol. 2012;8(2):166-75.

[14]. Gatari MJ, Kinyari BN, Gaita SM, Wafula G, Blake DR, Harrison RM. The state of air quality in Nairobi, Kenya. Atmos Environ. 2015;123:177-84.

[15]. Egondi T, Kyobutungi C, Ng N, Muindi K, Oti S, Vijver S, et al. Exposure to airborne particles and respiratory health in Nairobi informal settlements. Environ Health. 2018;17(1):62.

[16]. Onyango C, Wamukoya DK, Macharia E, Ayah R. Air quality monitoring in Kenya: Current status and future perspectives. Environ Sci Policy. 2021;122:36-46.

[17]. Breiman L. Random forests. Mach Learn. 2001;45(1):5-32.

[18]. Ravindra K, Bahadur SS, Katoch V, Bhardwaj S, Kaur-Sidhu M, Gupta M, et al. Machine learning models for predicting respiratory diseases due to air pollution in urban India. Environ Res Lett. 2023;18(1):014003.

[19]. Li L, Sun J, Jiang X, Liu X. Predicting high-cost patients using medical insurance data: A case study in western China. Health Serv Res. 2019;54(1):120-30.

[20]. Patel SJ, Teach SJ, Haynes ML, Mathew M, Mittal MK. Predictive modeling of asthma exacerbations in pediatric patients using machine learning. Pediatr Pulmonol. 2018;53(6):873-82.

[21]. Ravindra K, Bahadur SS, Katoch V, Bhardwaj S, Kaur-Sidhu M, Gupta M, et al. Application of machine learning approaches to predict the impact of ambient air pollution on outpatient visits for acute respiratory infections. Department of Community Medicine & School of Public Health, PGIMER, Chandigarh 160012, India. 2023.

[22]. Harrou F, Dairi A, Sun Y, Kadri F. Detecting abnormal ozone measurements with a deep learning-based strategy. IEEE Sens J. 2018;18:7222-32. doi: 10.1109/jsen.2018.2852001.

[23]. Xi Y, Tian CL, Qian L. A study of deep learning methods for de-identification of clinical notes in cross-institute settings. BMC Med Inform Decis Mak. 2019;19:232. doi: 10.1186/s12911-019-0935-4.

[24]. Gans D, Kralewski J, Hammons T, Dowd B. Medical groups' adoption of electronic health records and information systems. Health Aff. 2005;24:1323-33. doi: 10.1377/hlthaff.24.5.1323.

[25]. Raghupathi W, Raghupathi V. Big data analytics in healthcare: Promise and potential. Health Inf Sci Syst. 2014;2:3. doi: 10.1186/2047-2501-2-3.

[26]. Yu G, Yang Z, Shi Y. Identification of pediatric respiratory diseases using a fine-grained diagnosis system. J Biomed Inform. 2021;117:103754. doi: 10.1016/j.jbi.2021.103754.

[27]. Deo RC. Machine learning in medicine. Circulation. 2015;132:1920-30. doi: 10.1161/CIRCULATIONAHA.115.001593.

[28]. Patrício M, Pereira J, Crisóstomo J, Matafome P, Gomes M, Seiça R, et al. Using resistin, glucose, age, and BMI to predict the presence of breast cancer. BMC Cancer. 2018;18:181-88. doi: 10.1186/s12885-017-3877-1.

[29]. Abera A, Friberg J, Isaxon C, Jerrett M, Malmqvist E, Sjöström C, et al. Air quality in Africa: Public health implications. Annu Rev Public Health. 2021;42:193-210. doi: 10.1146/annurev-publhealth-100119-113802.

[30]. Agbo KE, Walgraeve C, Eze JI, Ugwoke PE, Ukoha PO, Van Langenhove H. A review on ambient and indoor air pollution status in Africa. Atmos Pollut Res. 2021;12:243-60. doi: 10.1016/j.apr.2020.11.006.

[31]. Kurmi OP, Lam KBH, Ayres JG. Indoor air pollution and the lung in low- and medium-income countries. Eur Respir J. 2012;40(1):239-54. doi: 10.1183/09031936.00193311.

[32]. Abegaz SB, Zereyesus YA, Dalie FS, Belay KA. Air pollution and respiratory health: A review. Int J Environ Res Public Health. 2021;18(4):1947. doi: 10.3390/ijerph18041947.

[33]. Amegah AK, Agyei-Mensah S. Urban air pollution and noncommunicable diseases in low- and middle-income countries: A narrative review. J Environ Public Health. 2021;2021:9747538. doi: 10.1155/2021/9747538.

[34]. Chowdhury S, Dey A, Smith KR. Ambient PM2.5 exposure and premature mortality burden in the 10 most populous urban localities in India: An assessment of exposure-response relationships. Environ Health Perspect. 2021;129(5):057004. doi: 10.1289/EHP7071.

[35]. Limaye VS, Schraufnagel DE. Impact of air pollution on lung health—Strategies for global action. Glob Heart. 2021;16(1):28. doi: 10.5334/gh.897.