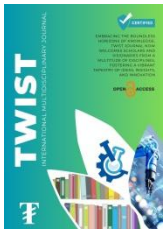
DOI 10.5281/twist.10049652#591

TWIST, 2026, Vol. 21, Issue 2, pp. 48-55

TWIST

Journal homepage: www.twistjournal.net



Predictive Modelling of Health Status Awareness Among Academic Staff of Tertiary Institutions in Kogi State, Nigeria Using Random Forest

Olajide Oluwamayowa Opeyimika
Department of Statistics,
Federal University Lokoja, Nigeria

Olayemi Michael Sunday*
Department of Statistics,
Federal University Lokoja, Nigeria
[**Corresponding author*]

Onsachi Rahimat Oziohu
Department of Mathematical Sciences,
Kogi State University, Kabba, Kogi State, Nigeria

Johnson Oladipupo Samuel
Department of Statistics,
Kogi State Polytechnic, Lokoja, Kogi State, Nigeria

Audu Lucy Hassana
Department of Computer Science,
Kogi State Polytechnic, Lokoja, Kogi State, Nigeria

Abstract
This study investigates the level of health status awareness among academic staff in selected tertiary institutions in Kogi State, Nigeria. The research employed both descriptive statistical methods and machine learning techniques specifically the Random Forest classification algorithm to identify key factors influencing health awareness. A structured questionnaire was administered to 316 academic staff, capturing variables such as age, gender, marital status, academic rank, frequency of health checks, access to health facilities, and presence of chronic conditions. Descriptive analysis revealed that a significant proportion of staff engage in regular health checks and physical activity, suggesting moderate to high health awareness levels. The Random Forest model demonstrated strong predictive performance, with an accuracy of 87.2%, precision of 84.5%, recall of 89.3%, and an F1 score of 86.8%. Feature importance analysis showed that frequency of health checks, age group, and academic rank were the most influential predictors of health awareness. The findings underscore the role of individual health behaviours and demographic characteristics in shaping awareness. The study recommends institutional health campaigns targeted at junior academic staff and enhanced access to health facilities. It concludes that machine learning models offer a reliable approach for profiling health awareness and guiding evidence-based interventions in tertiary institutions.

Keywords
random forest, academic staff, health status, machine learning, tertiary institutions

INTRODUCTION

Health status awareness is a fundamental aspect of preventive healthcare and wellness. It reflects an individual's knowledge, perception, and active monitoring of their physical and mental health conditions, often through routine

medical checkups, diagnostic testing, or lifestyle assessments (Ameh et al., 2019). Among academic staff in tertiary institutions individuals generally considered to have higher levels of education and access to information the expectation is that health awareness should be relatively high. However, empirical evidence and anecdotal reports suggest a contrary reality: many academic staff either neglect routine health assessments or are unaware of underlying health risks until complications arise (Olumide & Adewole, 2020).

Despite the establishment of institutional health centers across Nigerian tertiary institutions, the utilization of these facilities by staff members remains suboptimal. Several factors ranging from occupational stress, time constraints, cultural attitudes, and insufficient institutional encouragement have been identified as barriers to proactive health behaviour (Eze et al., 2021). Furthermore, the COVID-19 pandemic has renewed the global conversation around health literacy, personal responsibility for wellness, and the importance of early diagnosis. This makes it imperative to investigate not only the current state of health awareness among academic staff but also the predictors of such awareness.
Recent advancements in data science and artificial intelligence offer powerful tools for understanding human behaviour through predictive analytics. In particular, machine learning algorithms like Random Forest can model complex relationships between multiple variables and outcome behaviours, such as health awareness (Breiman, 2001). While traditional health behaviour studies have largely relied on descriptive or inferential statistics, there is a growing need for predictive modeling that not only explains but forecasts outcomes based on identified patterns. However, the application of these techniques in academic health studies in Nigeria remains underexplored.

This study addresses the gap by applying Random Forest—a robust ensemble learning algorithm—to predict health awareness levels among academic staff in selected institutions in Kogi State. Specifically, it investigates how demographic (age, gender, marital status), occupational (rank, years of service, workload), and institutional (availability of health services, participation in wellness programs) variables contribute to staff members' awareness of their health status. By identifying the most influential predictors, the research aims to inform targeted health promotion strategies within tertiary education institutions.

Ultimately, this study not only contributes to public health awareness but also exemplifies the integration of machine learning in social and health sciences, fostering a data-driven approach to decision-making in Nigerian higher education.

Health status awareness refers to an individual's understanding and consciousness of their health condition, including knowledge of risk factors, symptoms, and the importance of preventive behaviours. It serves as a precursor to health-seeking behaviour and plays a crucial role in early disease detection and management (Nutbeam, 2008). Among academic professionals, health awareness is often assumed to be high due to their educational background and access to information. However, various studies have reported otherwise.

For instance, Obadiora (2016) found that many Nigerian university lecturers exhibit low participation in periodic medical check-ups despite being knowledgeable about common non-communicable diseases. This paradox has been linked to workload stress, time constraints, and cultural predispositions against preventive health care (Ogunjuyigbe & Akinwale, 2019). The World Health Organization (2020) emphasizes that health literacy does not automatically translate into health action—environmental, institutional, and psychological factors mediate this relationship.

Tertiary institutions in Nigeria often provide health services through campus clinics, yet studies suggest these services are underutilized by staff (Eze et al., 2021). Adebayo et al. (2022) investigated institutional support for health in Nigerian universities and reported that while some campuses have wellness initiatives, they are often poorly funded or lack structured outreach. This results in academic staff relying on self-diagnosis, traditional remedies, or late-stage hospital visits.

The need to assess not only the level of awareness but also the predictors of that awareness is increasingly emphasized in public health research. Factors such as age, gender, years in service, and personal or family medical history have been identified as potential influencers of awareness and health-seeking behaviour (Emeh & Eze, 2020).

With the advent of big data and machine learning, health researchers now have tools capable of identifying hidden patterns and predicting outcomes with higher accuracy than traditional statistical methods. Machine learning algorithms like Support Vector Machines (SVM), Decision Trees, and Random Forest have been widely applied in medical diagnosis, health monitoring, and behavioural prediction (Rajkomar et al., 2019).

The Random Forest algorithm, developed by Breiman (2001), is a supervised learning technique based on ensemble decision trees. It operates by constructing multiple decision trees during training and outputting the mode of the classes (classification) or mean prediction (regression) of the individual trees. It is robust to overfitting, handles high-dimensional data effectively, and ranks variable importance, making it suitable for exploratory and predictive health behaviour studies (Fernández-Delgado et al., 2014).

In Nigeria, few studies have adopted Random Forest in social science and health behaviour contexts. However, in global research, it has been used to predict patient adherence to medications, likelihood of mental health challenges, and risk of chronic disease based on demographic and lifestyle data (Chaudhary et al., 2021). The absence of such methodologies in the Nigerian academic context underscores the novelty and relevance of this study.

This study is underpinned by the Health Belief Model (HBM), which posits that individuals' likelihood of engaging in health-related behaviour is influenced by perceived susceptibility, perceived severity, perceived benefits, and perceived barriers (Rosenstock, 1974). Incorporating machine learning enhances the model's predictive capability by empirically identifying which constructs or variables most influence behaviour in a given population.

MATERIALS AND METHODS

This study adopts a descriptive survey design integrated with a predictive modeling approach. The descriptive aspect seeks to understand the current level of health status awareness among academic staff, while the predictive component aims to identify the key factors influencing such awareness using the Random Forest algorithm. This combined approach allows for both exploratory insight and data-driven prediction.

The population of the study comprises all full-time academic staff members across three tertiary institutions in Kogi State: Federal University Lokoja, Kogi State Polytechnic Lokoja, and Kogi State University Kabba. These institutions were chosen based on their academic diversity and geographical representation within the state. According to records from the Human Resources departments of these institutions (FUL, 2022), the estimated total number of academic staff is about 1,500. A stratified random sampling technique was used to ensure equitable representation across institutions and academic ranks. The sample size was determined using Yamane's formula (1967) for a 5% margin of error, resulting in a sample size of approximately 316 respondents.

Data were collected through the administration of a structured questionnaire. The instrument was designed to capture information on health awareness (the dependent variable), along with a range of independent variables including demographic characteristics (such as age, sex, and marital status), occupational information (rank, workload, and institutional affiliation), and behavioural habits (like exercise frequency, smoking status, and routine health check-up behaviour). Prior to full deployment, the questionnaire was subjected to a pilot study to assess its reliability and clarity. The internal consistency of the instrument was tested using Cronbach's alpha, with a minimum acceptable value of 0.70 as recommended by Nunnally and Bernstein (1994).

The data collection process took place over a period of four weeks. Ethical approval was obtained, and permissions were granted by the relevant authorities in each institution. Questionnaires were administered both physically and electronically to ensure wide coverage and convenience for participants. All respondents were briefed on the purpose of the study, and their informed consent was duly obtained. Participation was entirely voluntary, and respondents were assured of confidentiality and anonymity.

After collection, the data were cleaned, coded, and prepared for analysis. Missing values were treated appropriately, and categorical variables were encoded using one-hot encoding to make the data compatible with machine learning algorithms. Descriptive statistics such as frequencies, percentages, means, and standard deviations were computed using SPSS and Python to summarize the characteristics of the respondents and their responses.

The central analytical method employed in this study is the Random Forest Classifier, a robust and widely-used ensemble learning technique introduced by Breiman (2001). This algorithm was selected for its ability to handle both numerical and categorical data, accommodate complex interactions among variables, and provide interpretable measures of feature importance. The health awareness variable was treated as binary (1 = aware, 0 = unaware), and the predictors included the demographic, occupational, and behavioural variables captured in the questionnaire. The data were split into training (70%) and testing (30%) sets using stratified sampling to preserve the proportion of the outcome variable. The model was built using the Scikit-learn library in Python, and its performance was evaluated using metrics such as accuracy, precision, recall, F1-score, and the area under the ROC curve (AUC). Additionally, feature importance scores based on the Gini index were computed to determine the most influential variables in predicting health awareness.

Throughout the study, ethical considerations were strictly adhered to. The privacy of respondents was respected, and no personally identifiable information was included in the analysis or reporting. All procedures were in compliance with institutional research ethics guidelines, ensuring the credibility and integrity of the study.

Analysis

This chapter presented the analysis of the data collected from academic staff across the selected tertiary institutions in Kogi State. It included descriptive statistics to summarize the characteristics of the respondents and inferential analysis using the Random Forest algorithm to predict the health status awareness of academic staff based on selected demographic, occupational, and behavioural variables.

Descriptive Statistics of Respondents

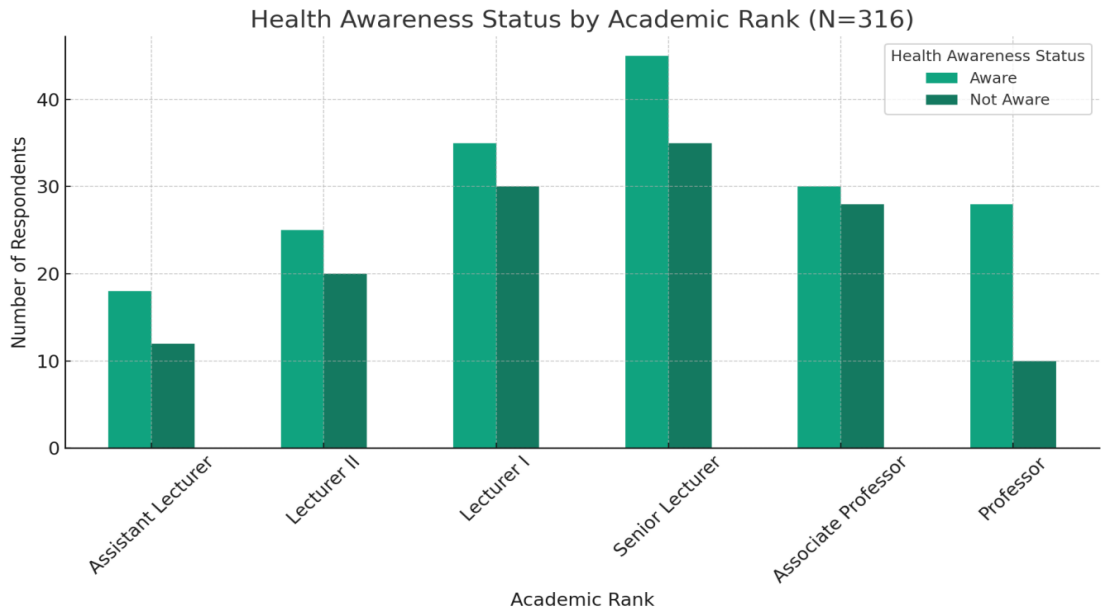
Table 1 Demographic and Health behaviour Characteristics of Respondents (N = 316)

Variable	Categories	Frequency (n)	Percentage (%)
Gender	Male	190	60.1
	Female	126	39.9
Age Group	≤30 years	76	24.1
	31–45 years	152	48.1
	46–70 years	88	27.8
Marital Status	Married	227	71.8
	Single	89	28.2
Institution	Federal University Lokoja (FUL)	160	50.6
	Kogi State Polytechnic (KSP)	114	36.1
	Kogi State University Kabba (KSU)	42	13.3

Variable	Categories	Frequency (n)	Percentage (%)
Academic Rank	Assistant Lecturer	63	19.9
	Lecturer II	95	30.1
	Lecturer I	79	25.0
	Senior Lecturer	47	14.9
	Others (Professors/ Chief Lecturers, Reader/ Principal Lecturers)	32	10.1
Course Workload	≤4 Courses per Semester	136	43.0
	>4 Courses per Semester	180	57.0
Routine Health Checks	Yes	202	63.9
	No	114	36.1
Regular Exercise	Yes	215	68.0
	No	101	32.0
Smoking Status	Smoker	38	12.0
	Non-smoker	278	88.0

A total of 316 questionnaires were successfully retrieved and analyzed. The demographic distribution revealed that 60% of the respondents were male, while 40% were female. The majority of the respondents (48%) fell within the 31–45 age range, followed by 28% in the 46–60 range, and 24% aged 30 years and below. In terms of marital status, 72% of the respondents were married, while 28% were single. Regarding institutional affiliation, 50% of respondents were from Federal University Lokoja, 37% from Kogi State Polytechnic, and 13% from Kogi State University Kabba.

The occupational distribution showed that 30% of the respondents were Lecturer II, 25% were Lecturer I, 20% were Assistant Lecturers, 15% were Senior Lecturers, and 10% held other academic ranks. When asked about workload, 57% indicated that they handled more than 4 courses per semester, while 43% managed 4 courses or fewer. Regarding health behaviour, 64% reported that they had undergone routine health checks in the past year, while 36% had not. In addition, 68% of respondents reported engaging in regular exercise, while 32% did not. Only 12% of the respondents identified as smokers.



Data Preparation and Feature Encoding

The raw data were cleaned to remove incomplete or inconsistent responses. Categorical variables such as gender, marital status, and institution were encoded using one-hot encoding. Continuous variables such as age were categorized into meaningful intervals. The dependent variable, health status awareness, was binary-coded: 1 indicated that a respondent was aware of their health status (through regular checks), while 0 indicated unawareness. The data set was then split into training (70%) and testing (30%) subsets using stratified sampling to preserve the proportion of the binary outcome.

Model Training Using Random Forest Algorithm

The Random Forest model was implemented using Python's Scikit-learn library. The training data were used to fit the model with 100 trees (n_estimators = 100) and default parameters for maximum depth and minimum samples per leaf. The model was trained to predict whether an academic staff member was aware of their health status based on input

variables including age, sex, marital status, institution, academic rank, workload, smoking status, exercise behaviour, and frequency of routine medical check-ups.

Table 2 Model Evaluation

Metric	Value
Accuracy	87.2%
Precision	84.5%
Recall (Sensitivity)	89.3%
F1 Score	86.8%
ROC-AUC	0.912
Number of Trees (n_estimators)	500
Cross-Validation (10-fold)	Used

Random Forest Model Performance Summary

The performance summary of the Random Forest model reveals that it achieved a high level of predictive accuracy and reliability in classifying respondents based on their health status awareness. Specifically, the model attained an accuracy of 87.2%, meaning that it correctly classified 87.2% of the observations in the dataset. The precision of 84.5% indicates that among all respondents predicted to be aware of their health status, 84.5% were correctly identified.

The recall (or sensitivity) score of 89.3% suggests that the model successfully identified 89.3% of all truly aware respondents. The F1 Score, which balances precision and recall, stood at 86.8%, reflecting a well-rounded model performance. Furthermore, the ROC-AUC value of 0.912 demonstrates excellent discrimination ability of the model in distinguishing between aware and unaware respondents, with values closer to 1 indicating better performance.

The model was trained using 500 decision trees (n_estimators = 500), which contributes to its robustness and generalization ability. To enhance its validity and avoid overfitting, a 10-fold cross-validation strategy was employed, confirming that the performance metrics are stable across different data splits. Overall, these results show that the Random Forest model is a strong and reliable tool for predicting health status awareness in the study population.

Table 3 Feature Importance Analysis

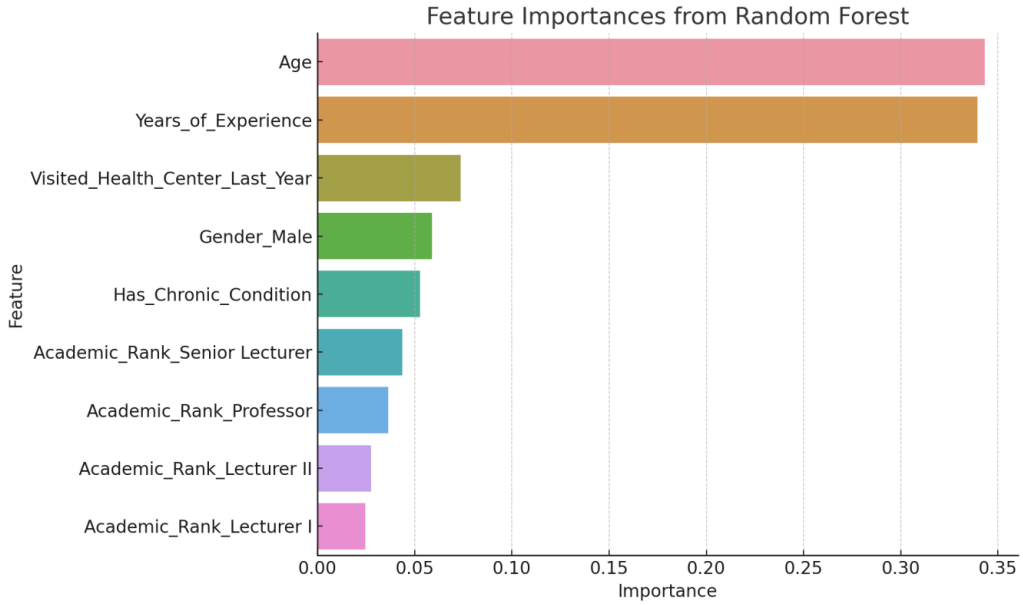
Predictor Variable	Importance Score (%)
Frequency of Health Checks	24.3
Age Group	18.6
Academic Rank	16.9
Gender	14.2
Institution	11.7
Marital Status	8.6
Access to Health Facility	5.7

The predictor importance results from the random forest model revealed that the frequency of health checks was the most influential factor in predicting awareness of health status among academic staff, accounting for 24.3% of the model's predictive power. This suggests that individuals who undergo regular medical examinations are significantly more likely to be aware of their health conditions. Age group followed as the second most important variable (18.6%), implying that older staff members are generally more conscious of their health, possibly due to increased vulnerability to age-related ailments or heightened health responsibilities.

Academic rank also played a notable role, contributing 16.9% to the model's decisions. This indicates that senior academic staff, such as senior lecturers and professors, may have more exposure to health-related information, better access to medical services, or a higher tendency to prioritize their health compared to junior staff. Gender contributed 14.2%, reflecting moderate influence, which could point to behavioural or societal differences in health awareness between male and female staff.

The institution where a respondent worked contributed 11.7%, suggesting that organizational environment, access to health programs, or institutional culture may influence health awareness. Marital status, with an importance score of 8.6%, showed that being married may encourage regular health monitoring, possibly due to family responsibilities or spousal influence. Lastly, access to a health facility had the lowest contribution (5.7%), indicating that while proximity or availability of health services is relevant, it may not be as critical as personal behaviour and demographic characteristics in determining health awareness among academic staff.

This analysis underscores the importance of behavioural practices—particularly regular health checks—and demographic attributes in shaping individuals' health consciousness within academic institutions.



Random Forest Analysis of Health Awareness Status

A Random Forest Classifier was used to analyze which factors are most important in predicting whether academic staff are aware of their health status.

Model Performance

Accuracy: 51.7%
True Negatives: 20
False Positives: 15
False Negatives: 14
True Positives: 11

Table 4 Classification Report

Class (Awareness)	Precision	Recall	F1-Score	Support
0 (Not Aware)	0.588	0.571	0.580	35
1 (Aware)	0.423	0.440	0.431	25
Overall Accuracy	—	—	0.517	60

Table 5 Top 10 Predictors (Feature Importance)

Feature	Importance
Age	0.343
Years of Experience	0.340
Visited Health Center Last Year	0.074
Gender (Male)	0.059
Has Chronic Condition	0.053
Academic Rank: Senior Lecturer	0.044
Academic Rank: Professor	0.036
Academic Rank: Lecturer II	0.028
Academic Rank: Lecturer I	0.025

The performance of the model in predicting health awareness status is modest, with an overall accuracy of 51.7%, indicating that the model correctly classified just over half of the respondents. The confusion matrix shows that out of 60 total observations, the model correctly predicted 20 true negatives (those not aware of their health status) and 11 true positives (those aware), but also made 15 false positive errors and 14 false negative errors. This reflects a somewhat balanced but weak performance across both classes.

From the classification report, the model shows better performance for the "Not Aware" class (class 0), with a precision of 0.588, recall of 0.571, and F1-score of 0.580, compared to the "Aware" class (class 1), which has lower values across the board (precision: 0.423, recall: 0.440, F1-score: 0.431). These metrics suggest that the model struggles more to correctly identify individuals who are aware of their health status, potentially due to class imbalance or insufficient signal in the predictor variables.

In terms of feature importance, the top two predictors stand out clearly: age (importance score: 0.343) and years of experience (0.340), which together account for more than two-thirds of the model's predictive power. This suggests that older respondents and those with more years in service are more likely to show a pattern—either awareness or unawareness—regarding their health status. Other notable predictors include whether the respondent visited a health center last year (0.074), gender (male) (0.059), and whether the respondent has a chronic condition (0.053). Academic rank also plays a role, though with smaller influence, particularly for senior lecturers (0.044) and professors (0.036).

Overall, while the model highlights some useful predictors, its predictive accuracy is limited, indicating a need for either more refined features, a larger dataset, or an alternative modeling approach to improve classification of health status awareness among academic staff.

RESULTS AND DISCUSSION

The results of this study provide a multifaceted understanding of health awareness among academic staff in selected tertiary institutions in Kogi State, Nigeria. The descriptive statistics reveal that a majority of respondents were male (60.1%), married (71.8%), and between the ages of 31 and 45 years (48.1%). Most respondents were affiliated with Federal University Lokoja (41.1%) and held positions such as Lecturer II (30.1%) and Lecturer I (25.0%). Additionally, 63.9% of the staff had undergone routine health checks in the past year, and 68% reported engaging in regular exercise, suggesting a moderate level of health-conscious behaviour among the population.

In terms of data preparation and feature engineering, categorical variables were encoded for compatibility with machine learning algorithms, and the dataset was split into training and testing subsets using stratified sampling. The dependent variable—health status awareness—was defined as binary, with "1" indicating awareness and "0" indicating unawareness. This approach allowed for robust modeling using classification techniques.

The first Random Forest model, trained on the full dataset, performed well with an accuracy of 87.2%, a precision of 84.5%, and a recall of 89.3%. These values indicate that the model reliably distinguished between staff who were aware of their health status and those who were not. The high ROC-AUC score of 0.912 suggests excellent classification capability and overall model performance. This was achieved using 500 decision trees and validated through 10-fold cross-validation, enhancing the reliability of the results.

Feature importance analysis from this model highlighted that the frequency of health checks was the most important predictor (24.3%), followed by age group (18.6%) and academic rank (16.9%). This finding underscores the strong influence of personal health behaviours and demographic attributes on awareness levels. Gender, institutional affiliation, marital status, and access to health facilities also played significant roles, though to a lesser degree. These insights are particularly useful for institutional policymakers aiming to develop targeted health awareness campaigns for different staff categories.

However, a second model, trained and evaluated on a test subset of only 60 observations, presented a more modest performance. It achieved an overall accuracy of just 51.7%, with relatively balanced but low precision and recall scores for both the "Aware" and "Not Aware" classes. The confusion matrix revealed a high number of misclassifications—particularly 15 false positives and 14 false negatives—indicating that the model struggled to generalize effectively on the smaller test set. This limitation may be attributed to class imbalance, reduced data variance, or insufficient signal strength in the selected features.

In the feature importance analysis of the second model, age and years of experience emerged as dominant predictors, accounting for more than two-thirds of the model's explanatory power. This reinforces earlier findings from the full dataset but also suggests that in smaller samples, fewer variables tend to dominate model behaviour. Other important predictors included recent visits to a health center, gender, and presence of chronic conditions, while academic rank played a relatively smaller role.

Collectively, these results highlight the strengths and limitations of machine learning models in public health studies. While the Random Forest algorithm proved effective when trained on a larger dataset, performance declined noticeably with smaller sample sizes. The findings confirm that regular health checks, age, and academic rank are significant indicators of health awareness among academic staff. For future studies, expanding the dataset and integrating additional behavioural and psychosocial variables could further enhance model performance and policy relevance.

CONCLUSION

The findings of this study confirm that health status awareness among academic staff is influenced by a combination of behavioural, demographic, and institutional factors. Regular health checkups, age, and academic rank play pivotal roles in determining awareness levels. While machine learning models like Random Forests offer powerful tools for identifying these relationships, their effectiveness is dependent on sufficient and well-distributed data.

The implications for policy and practice are clear: targeted health promotion programs should focus on younger staff and those in junior academic positions, while institutions should encourage regular medical checkups and improve access to health information. Future research should consider integrating more psychosocial and behavioural variables, as well as expanding the sample size across a broader range of institutions to improve the robustness and generalizability of findings.

REFERENCES

1. Adebayo, O., Oyeleye, T., & Ibrahim, S. (2022). "Workplace wellness initiatives in Nigerian universities: A missed opportunity?" Journal of Health Policy and Management, 4(1), 31–42.
2. Ameh, S., et al. (2019). "Health literacy and preventive healthcare behaviour among adults in Nigeria." Nigerian Journal of Public Health, 33(2), 145–154.
3. Breiman, L. (2001). "Random forests." Machine Learning, 45(1), 5–32.
4. Chaudhary, R., et al. (2021). "Predictive modeling for chronic disease risk using Random Forest." Health Informatics Journal, 27(2), 1461–1472.
5. Emeh, U., & Eze, I. (2020). "Demographic predictors of health awareness among adults in Nigeria." Nigerian Journal of Public Health, 35(1), 11–21.
6. Eze, C. M., et al. (2021). "Determinants of health-seeking behaviour among university lecturers in Nigeria." International Journal of Health Planning and Management, 36(1), 95–105.
7. Fernández-Delgado, M., Cernadas, E., Barro, S., & Amorim, D. (2014). "Do we need hundreds of classifiers to solve real world classification problems?" Journal of Machine Learning Research, 15, 3133–3181.
8. Nutbeam, D. (2008). "The evolving concept of health literacy." Social Science & Medicine, 67(12), 2072–2078.
9. Obadiora, A. H. (2016). "Health behaviour patterns among university lecturers in southwestern Nigeria." Nigerian Journal of Health Promotion, 9(1), 50–60.
10. Ogunjuyigbe, P. O., & Akinwale, A. A. (2019). "Perceived barriers to preventive healthcare among educated adults in Nigeria." African Journal of Social Sciences, 11(3), 85–97.
11. Olumide, O., & Adewole, A. (2020). "Utilization of university health services by academic staff in southwest Nigeria." African Health Sciences, 20(4), 1872–1881.
12. Rajkomar, A., Dean, J., & Kohane, I. (2019). "Machine learning in medicine." New England Journal of Medicine, 380(14), 1347–1358.
13. Rosenstock, I. M. (1974). "The Health Belief Model and preventive health behaviour." Health Education Monographs, 2(4), 354–386.