ABUAD Journal of Engineering Research and Development (AJERD)
ISSN (online): 2645-2685; ISSN (print): 2756-6811

Volume 7, Issue 2, 541-554

Machine Learning for Health Insurance Prediction in Nigeria

Oluwasogo Adekunle OKUNADE[1], Victor Enemona OCHIGBO[2], Emmanuel Gbenga DADA[3], Olayemi Mikail OLANIYI[1], Oluwatoyosi Victoria OYEWANDE[1]

[1]*Department of Computer Science, Faculty of Sciences, National Open University of Nigeria, Abuja, Nigeria*
aokunade@noun.edu.ng/omolaniyi@noun.edu.ng/ofayemi@noun.edu.ng

[2]*Department of Knowledge Management and Communication, Agricultural Research Council of Nigeria, Abuja, Nigeria*
v.ochigbo@arcn.gov.ng

[3]*Department of Computer Science, Faculty of Physical Sciences, University of Maiduguri, Maiduguri, Nigeria*
gbengadada@unimaid.edu.ng

Corresponding Author: v.ochigbo@arcn.gov.ng, +2348063888982

Date Submitted: 07/06/2024
Date Accepted: 26/11/2024
Date Published: 26/12/2024

Abstract: *Health insurance coverage remains critical to healthcare accessibility, particularly in developing nations like Nigeria. This paper focused on predicting the likelihood of medical insurance coverage among individuals in Nigeria by employing four prominent Machine learning techniques: Logistic Regression, Random Forest, Decision Tree, and Support Vector Machine classifiers. The dataset utilized for analysis comprises demographic information, socioeconomic factors, and health-related variables collected from a diverse sample across Nigeria. Four models are trained and evaluated: Logistic Regression widely accepted for its simplicity and interpretability. Random Forest is a robust ensemble learning algorithm capable of capturing complex relationships within the data. The decision Tree model is simple to understand and visualize and the Support Vector Machine model is known for producing a very good classification. Furthermore, the performance metrics utilized to rate the predictive capabilities of the models are Accuracy, Precision, Sensitivity, F Score, and area under the Receiver Operating Characteristic (AUC & ROC Curve). Additionally, a features importance analysis is conducted for the identification of the dominant factors contributing to the prediction of the spread of medical insurance in Nigeria. The outcome of this paper gives insights in the efficiency of each machine learning models used to forecast medical insurance coverage, and identifying key determinants influencing insurance coverage can assist policymakers and healthcare stakeholders in devising targeted strategies to improve healthcare access and affordability for the Nigerian people.*

Keywords: *Ensemble Technique, Odd Ratio, Confusion Matrix, Feature Importance, Medical Insurance.*

1. INTRODUCTION

Health insurance is crucial to the healthcare system, providing individuals as well as families with financial security against high medical costs. This assent involves a person or a group and an insurance company in which the company undertakes part-payment or all of the medical expenses in exchange for suitable premiums. The Nigerian National Health Insurance Scheme (NHIS) is a government program designed to improve access to quality healthcare services for Nigerian citizens and persons of other nationalities who legally live in Nigeria.

The National Health Insurance Scheme (NHIS) is an organization enacted by Degree 35, of 1999 (now Act 35). It started operations officially in 2005 and operates as Public Private Partnership, providing convenient, affordable, and quality healthcare to all Nigerians by providing monetary risk protection and reduction in the difficulty of the out-of-pocket expenses incurred by individuals or households [1]. It was originally called the National Health Insurance Scheme (NHIS) Act but presently referred to as National Health Insurance Authority (NHIA) Act. An Act that repeals the NHIS Act, CAP. N42, Laws of the Federation of Nigeria, 2004, and validate the National Health Insurance Authority (NHIA) Act, 2022, was formally documented by the Federal Government of Nigeria on 19th May, 2022 to promote, manage, and integrate medical insurance plans in Nigeria, to strengthen and utilize the involvement of private sector to provide healthcare and achieving universal health coverage to all Nigerians and other affiliated matters.

The mandate of NHIA is to achieve "Universal Health Coverage (UHC) in Nigeria by 2030" with an insight into building a foremost government organization committed to acquiring the availability of funds for good medical services in Nigeria. NHIA aims to make funds available deliberately for everyone in Nigeria to access good medical services. The World Health Organization says the most notable barrier to healthcare coverage and usage is medical bills. In Nigeria, medical expenditure is mainly via out-of-pocket. Out-of-pocket expenses presently account for about 70% of all medical expenses incurred [2]. However, the major way to mitigate direct out-of-pocket spending is for the government to bring up risk-pooling policies to share the risks involved among the health insurance companies. Although health expenses went up to 98.2 million in 2008 as against 12.5million Naira in 1970, healthcare is still not efficient as 3% of Nigerians aged 15-49 years are secured by medical insurance [3]. The inadequacy recorded by NHIS can also be attributed to the high poverty rate, "brain drain" and inadequate medical resources in Nigeria [4]. Addressing these challenges and expanding coverage to reach more Nigerians remains a priority for the scheme. The application of machine learning techniques will help substantially to curb the disputes in the spread of health insurance in Nigeria.

Machine learning can be referred to as an area of artificial intelligence (AI) which centers on the use of algorithms to design models, aiding computers in revamping their performance on specific jobs by learning from available data [5]. Although the effect of machine learning on health insurance coverage in Nigeria is still emerging, machine learning has the potential to impact health insurance in different ways. One of the ways machine learning can have a strong effect on health insurance coverage in Nigeria is by analyzing existing data for subsequent decision-making. It can help insurance companies analyse the vast amount of data to make more informed decisions [6]. This can be crucial for health insurance expansion to under-served areas and populations. However, the impact of machine learning on health insurance in Nigeria may be affected by features like regulatory policies, accessibility of data, and the adoption of the technique used.

Sequel to the creation of the National Health Insurance Authority (NHIA) in Nigeria, the country continues to grapple with significant challenges in achieving comprehensive health insurance coverage. The core problem lies in the NHIA's inability to fulfill its mandate of providing population coverage with adequate monetary risk protection [1]. This shortfall has resulted to restriction in accessing affordable medical aid for the masses, especially those in villages and under-resourced areas. Consequently, the goal of universal health coverage (UHC) remains elusive, raising concerns about the effectiveness of the NHIA's strategies, policies, and implementation frameworks. The aim of this study is to investigate the underlying factors contributing to the NHIA's challenges by the application of machine learning techniques including Logistic Regression (LR), Random Forest (RF), Decision Tree (DT), and Support Vector Machine (SVM) for prediction using R programming software and the Nigeria National Longitudinal Phone Survey (NLPS) 2021-2023 dataset, assess the impact on healthcare access and financial protection, and propose sustainable solutions to enhance health insurance coverage in Nigeria.

The training goal of this machine learning application is to build a good model with precise prediction that can address the challenges affecting health insurance coverage in Nigeria. The justification for using machine learning in this context is multi-faceted;

i. **Income Analysis:** Machine learning enables the identification of income-related barriers that prevent a significant portion of the population from accessing health insurance. By training models on income data, the NHIA can better understand which segments of the population are most at risk and develop targeted interventions to make health insurance more affordable.

ii. **Geographical Disparity Identification:** Machine learning models trained on regional data can uncover patterns in health insurance enrollment across different geographical zones. This allows for the identification of areas with low coverage, enabling the NHIA to focus its efforts on regions that require the most attention. The insights from this study, for example, show that the South-West zone requires more targeted initiatives to improve coverage.

iii. **Improved Predictive Accuracy:** The use of machine learning, particularly the SVM algorithm, has shown to provide superior predictive accuracy in forecasting health insurance coverage trends. This accuracy is crucial for planning and policy-making as it allows the NHIA to anticipate challenges and allocate resources more effectively. The study's suggestion to explore additional machine learning models further justifies the need for continuous improvement in predictive capabilities to better serve the population.

2. RELATED WORKS

The evolution of machine learning over the years has led to global acceptance across different fields of endeavour, paving the way for growth and improvement in various facets of our lives. The usage of machine learning is encouraged not just by its predictive competencies, but additionally employing the documented opportunity of bringing new perceptions. In addition, machine learning makes use of data to collect information and make predictions, consequently enabling new insights that have been previously unforeseen. Researchers have come up with numerous machine learning classifiers in different works of life, especially in the health sector [7]. Table 1 below is a thematic succinct of the contributions of scholars.

Table 1: Systematic literature review

Author and Year	Technique Used	Contributions	Research Gap
Han and Suh (2023)	Logistic Regression, C5.0, Random Forest	To predict unsatisfactory healthcare needed in post-disaster	Due to the lack of robustness of the dataset, the seriousness of the disease was not accounted for as it is an important factor in determining the need for

Author and Year	Technique Used	Contributions	Research Gap
			healthcare.
Chen, et al (2023)	Logistic Regression Optimization models: Batch gradient descent (BGD-LR), Differential Privacy with Batch gradient descent (BDP-LR)	Breast cancer prediction	These hybrid models have a longer time to compute data. Also restricted to the combination of divergent solitude techniques with a machine learning model.
Sun and Pan (2023)	Logistic Regression, K-Nearest Neighbours, Support Vector Machine, Decision Tree, Random Forest	To predict heart disease	The dataset utilized is a portion of the original data. There are other variables not included in the dataset that could have enhanced the correctness and minimized the rate of false negative outcomes.
Reghunathan et al (2024)	Logistic Regression, K-Nearest Neighbours, Support Vector Machine	To identify autism spectrum disorder across different age categories	A need for the use of large datasets to increase performance by utilizing better feature selection techniques.
Cai, (2023)	Random Forest	To diagnose breast cancer tumours	A significant amount of missing data in the data-set might affect the outcome of the model's predictive capability
Wei, et al (2023)	Logistic Regression, Random Forest, and Decision Tree	Prediction of breast cancer for early detection and diagnosis	Dependence on the Wisconsin dataset alone is not reliable. Bigger and divergent data is necessary
Dahal and Gautam, (2020)	Logistic Regression, Classification Tree with Bagging (Bagging CART), Random Forest, Support Vector Machine, and k-nearest Neighbors.	Diagnosis of coronary artery disease.	With accuracy of 89%, there is need for further studies to improve on performance. For instance, the utilization of Artificial Neural Networks on larger data-set. or researching other methods to identify important features to be used.
Gill et al (2023)	Gradient Boosted Trees, Logistic Regression, Support Vector Machine.	To detect the early stage of Mesothelioma	A larger and more robust dataset is required for validation to ensure general applicability.
Chae et al (2024)	Logistic Regression, Decision Tree, Random Forest, Support Vector Machine, eXtreme Gradient Boosting, Light Gradient Boosting Machine	To predict hearing recuperation for people suffering from chronic otitis media who undergo Canal-Wall-Down surgery.	Lack of adequate dataset to ensure a generalized validation of the models.
Zhang et al (2023)	Logistic Regression, Random Forest, Support Vector, and Fully Connected Neural Networks	To predict the results of people with Aneurysmal Subarachnoid Hemorrhage	The use of a small dataset of 301 patients may instigate bias which is customary with small-size datasets. Imbalanced distribution of the dataset can lead to overfitting or

Author and Year	Technique Used	Contributions	Research Gap
			underfitting as the case may be.
Zheng (2018)	Linear Regression, Lasso, Support Vector Regression and Random Forest models	Verification of the global warming and identify factors contributing to global warming	The major greenhouse gases used were CO_2, $NO_{2,}$ and CH_4. Fluorinated gases and water vapour were omitted. These would have affected the results of the research.
Dipto, et al (2020)	Logistic Regression, Support Vector Machine, and Artificial Neural Network models	designed a prototype system using the clinical dataset	The data utilized is small which would affect the models' performances.
Santana, et al (2023)	Decision Tree (DT), Multilayer Perceptron (MLP), Gradient Boosting Machine (GBM), Random Forest (RF), eXtreme Gradient Boosting (XGBoost), K-Nearest Neighbors (KNN), Support Vector Machine (SVM) and Logistic Regression (LR) models	Classification of COVID-19 and influenza during concurrent outbreak	Patients having small symptoms can constitute difficulties for the machine learning models.
Oyoo et al. (2024)	First layer: k-Nearest Neighbor, AdaBoost, Naïve Bayes, and Decision Tree using Logistic Regression as meta-classifier	using two-layer ensemble machine learning algorithms to forecast road traffic collision	More time is required to run the ensemble model compared to the single models. Also, the simulated crash dataset was used but the road collision dataset is recommended.
Almayyan (2016)	Random Forest, Decision Trees K-Nearest Neighbor, and Artificial Neural Network models	Enhancement of lymphatic disease diagnosis	The lack of adequate data affects the performance of the model negatively. Also, the problem of imbalanced clinical dataset.
Gai and Zhang (2023)	Optimized logistic regression algorithm	Prediction of water quality for agriculture	Only four features selected were used and this is not easy to work with due to highly sophisticated alterations in quality of water
Colot et al (2021)	Random Forest Model	Investigation of the value of multiple fine-grained data sources for churn detection within telephone companies	The factual outcomes are dependent on a dataset of a single telephone firm. Therefore, it's imperative to acquire dataset for analysis of other phone companies if the need arises.
Getu and Bhat (2024)	Binary Logistic Regression	Analysis of driving factors of urban growth in Bahir Dar city, Ethiopia	A lower image Resolution was used, although images with high resolution give a better, quality outcome
Wang et al (2024)	Random Forest	Characterization of tropospheric ozone pollution, predictions and analysis of influencing factors in South Western Europe.	The inclusion of the year 2020 in the predictive model explains the marginal bias in the model and, if the new crown epidemic is not included in the investigation, the model's accuracy can further

Author and Year	Technique Used	Contributions	Research Gap
			improve
Chen et al (2021)	Improved Naïve Bayes Classification algorithm	Traffic risk management	Inability to take into consideration the interaction of variables. Example: sample size, category, and others
Liu et al (2017)	Logistic regression	The logistic regression fusion rule enhances the performance of wireless sensor networks with less complex calculations	To find the solution to the logistic regression fusion rule weights, the samples need to be trained
Cho et al (2023)	Light Gradient Boosting Machine, Logistic Regression, Decision tree, Random Forest, Support Vector Machine, Deep and Neural Network,	To predict university student's dropout using machine learning algorithms	Support Vector Machine and Deep Neural Network both have very high execution time compared to the rest models
Nordin, et al (2023)	Hybrid modeling of Support Vector Machine and Logistic Regression	To enhance the classification accuracy of COVID-19	Distribute the training process of the Support Vector Machine on a non-linear kernel is difficult to achieve
Zhang et al (2023)	Support Vector Machine, eXtreme Gradient Boosting, Decision Tree, k-Nearest Neighbors, naïve Bayes, Neural Network, Penalized logistic Regression and Random Forest	Prediction of Parkinson's Disease	Lack of access to the Parkinson's Disease dataset, the research findings are mainly dependent on the PPMI dataset. The performance of Polygenic Risk Score prediction is dependent on Genome Wide Association Study and personal statistics summary mostly of Europe origin were used. Therefore, its application on persons from other parts of the world, the performance of the prediction may differ.
Abbasi, et al (2023)	Logistic Regression, Decision Tree, Gradient Boosting, Random Forest, Extra Tree, AdaBoost, Light Gradient Boosting Machine, Gaussian Naïve Bayes	To predict the optimization of skin cancer survival	The research was restricted to a particular dataset. A better result, perhaps can be achieved different datasets are used
Olaguez-Gonzalez, et al (2023)	Artificial Neural Network, Support Vector Machine, Random Forest,	Using Machine Learning Algorithms to predict autism spectrum disorder pertaining to gut microbiome composition	Limited dataset. Machine learning models are deficient to identify casualty. It's not possible to spot occurrence of mistakes during the process of sampling since the data was collected from public domain
Tu, et al (2023)	Logistic Regression, Random Forest, Light Gradient Boosting Machine, eXtreme Gradient Boosting	Prediction of mortality risks of patients with the trauma of brain injury in the intensive care unit.	This research, being a retroactive study, potential research is required for the validation of these study findings. Also, the picture variables like "midline shift" including presence or absence of brain ventricles were not quantifiably included in the model.
Rojek, et al (2024)	Linear Support Vector Classifier, Logistic	To develop an artificial intelligence-based	The model could not take into consideration the patient's

Author and Year	Technique Used	Contributions	Research Gap
	Regression, K-Nearest Neighbors Classifier, and Random Forest Classifier	prediction of the risk of heart attack as an element of preventive medicine	behavioral changes that could notably affect the real risk of myocardial infarction, nevertheless making it hard to make an accurate prediction
Singh, et al (2024)	Logistic Regression, Random Forest, Support Vector Machine, K-Nearest Neighbor, Decision Tree and Deep Neural Network	Prediction for early detection of congestive heart failure	Inappropriate and inconsistent dataset makes it difficult to train deep learning frameworks

3. MATERIALS AND METHODOLOGY

We shall discuss the specific procedure used in predicting households or individuals covered by National Medical Insurance in Nigeria in line with the research objectives. The four Machine Learning algorithms to be utilized for the development of the models to predict health insurance coverage in Nigeria are Logistic Regression, Random Forest Decision Tree, and Support Vector Machine.

3.1 Data Source

In this research, we shall use the Nigeria National Longitudinal Phone Survey (NLPS) 2021-2023 dataset. This survey was carried out by the National Bureau of Statistics (NBS), working together with the World Bank and 60_decibels sponsored financially by the Bill and Melinda Gates Foundation (BMGF), Federal Republic of Nigeria, United States Agency for International Development (USAID), World Bank (WB), Global Financing Facility for Women, Children and Adolescents (GFF) and Busara Center (BC).

3.2 Data Description

The Nigeria National Longitudinal Phone Survey (NLPS) 2021-2023 dataset will be used for this study. It contains eighteen thousand, hundred and nineteen (18,119) entries and a total of eight (8) variables (seven independent variables and one dependent variable). The independent variables are Healthcare needed, Sex, Age, Employment status, Income, the six geographical Zones, and Education, and the dependent or target variable is Health insurance. The dependent feature, “health insurance” is binary with values 0 and 1. 0 stands for “no insurance” and 1 stands for “insurance”. The education variable contains six categories namely: 1-primary education, 2-secondary education, 3-vocational education, 4-university education, 5-Islamic education, and 6-other post-secondary education. The Zone represents the six geopolitical zones in Nigeria namely: 1-North Central (NC), 2-North East (NE), 3-North West (NW), 4-South East (SE), 5-South South (SS) and 6-South West (SW). The description of the dataset, label and values is given in Table 2 below.

3.3 Data Analysis and Pre-processing

Analytic data research is conducted for underlying data insight into data collection by visualizing the data and the results of the performance. This process will help find null entities in the dataset which are all removed to enable the machine learning algorithms to run successfully on the dataset. Thereafter the numerical variables, categorical variables, etc. are all identified and the categorical values are converted to binary values 0 and 1. Furthermore, data correlation is carried out to ensure the extent of association between the independent variables and the target variable. Finally, a statistical summary is executed to verify the non-existence of all null values.

3.4 Data Partitioning

In this step, the data is partitioned randomly into three datasets namely the train data, validation data, and test data using the “Keras” library package multi-assignment operation. 60% of the dataset is allocated to the train data, utilized to train the machine learning models developed by the logistic regression, decision tree, support vector machine and random forest algorithms. 20% to the validation dataset used to validate the models and 20% to the testing dataset for final evaluation. Training data establishes the link between the predictors and target variables while the testing dataset checks the efficiency of the model [30]. The principal aim of data partitioning is to circumvent overfitting. Overfitting takes place when the machine learning models perform very well on the training data-set and badly on the validation and test datasets.

3.5 Machine Learning Algorithm

Recently, machine learning has got more acceptances. There are various machine learning supervised models to solve classification problems, however, we shall be using four algorithms namely: Logistic Regression, Random Forest, Decision Tree, and Support Vector Machine algorithms for this study. This is due to their ability to perform well on a large dataset [7].

3.5.1 Logistic regression

Logistic Regression, a machine learning algorithm emanated from the field of statistics [8] and used primarily to solve binary distribution problems then predict the binary target feature with the use of the sigmoid function [7]. This machine learning technique performs well on large datasets. The foundation of logistic regression is the logistic function or sigmoid function that accepts real numbers and associates them to values bounded by 1 and 0.

Table 2: Data description

Variable Name	**Variable Label**	**Variable Value**
Healthcare needed	A binary variable indicating persons in need of healthcare	1 = need of healthcare, 0= no need for healthcare
Sex	A binary variable indicating Gender (for example. male or female)	1 = Male, 0=Female
Age	Ages of members of the household	Integer
Employment Status	A binary variable indicating if employed or not employed	1 = employed 0 = not employed
Income	A binary variable indicating any income in the past four weeks or not	1 = income, 0 = no income
Zone	A categorical variable indicating Geopolitical zones in Nigeria	1 = North Central (NC), 2 = North East (NE), 3 = North East (NE), 4=South East (SE), 5=South South (SS), 6=South West (SW)
Education	A categorical variable indicating the highest level of education	1 = Primary education, 2= Secondary education, 3 = Vocational education 4 = University education 5 = Islamic education 6 = other post-secondary education
Health insurance	(Target variable) A binary factor indicating health insurance or no health insurance	1 = health insurance 0 = no health insurance

Each independent variable gets a coefficient from the logistic regression model which calculates its autonomous input to the disparity in the target feature [32]. The target feature records 1 if the reply is “Yes” and records 0 if the reply is “No”. To predict the probabilities, the model is given by the Equation (1) called the sigmoid function.

$$\mathrm{P(Y)} = \frac{e^{\beta 0+\beta 1 X1+\beta 2 X2+\cdots+\beta k\, Xk}}{1+e^{\beta 0+\beta 1 X1+\beta 2 X2+\cdots+\beta k\, Xk}} \tag{1}$$

In which, Y is the outcome, X_1 X_2….X_k are the predictor variables, $\beta 0 + \beta 1\, X1 + ...+ \beta k\, Xk$ represent the model coefficients while $\beta 0$ represents the intercept. The coefficients specify the level of correlation existing in each independent variable and the target variable. Each coefficient indicates an expected change in the target variable if the independent feature changes by one unit. Logistic regression intends to forecast appropriately the class of outcome for every single case by utilizing the finest model. To achieve this, a model is developed including all the independent features that could be used to predict the target variable. The Logistic Regression computes the odd ratio by taking the exponential of each coefficient. It is a means used to compare the odds of an event occurring in two separate groups. Odd ratio aids in investigating the probability of the outcome of a given event, under certain conditions. For example, an odd ratio of 2 shows that the odds of the occurrence of an event are two times higher in one group in comparison to the other.

3.5.2 Random forest (Ensemble technique)

Random forest algorithms fall among the most common machine learning techniques [30]. In the past two decades, random forest classifiers have acquired great recognition due to their very impressive classification outcomes and the fast-processing rate [33]. It is an integrated machine-learning technique that creates multiple learners for a task [34]. The ideal goal of the combination is to find high accuracy with better performance. The ensemble algorithm uses a different method to prioritize items in the dataset compared to the single classifiers. Multiple ensemble learners are created and all the learners are merged in conformity with some voting system. Random forest algorithm which is an extension of the Decision Tree technique can be used to predict future events using multiple classifiers instead of a single one to achieve accuracy and precision.

3.5.3 Decision tree

Decision tree algorithm is a common machine learning algorithm suitable for both regression and classification trees. This works by partitioning the data repeatedly by features, creating a tree-like structure where each node, based on a feature, stands for a decision [7]. This algorithm was used in this study due to its simplicity in understanding and interpretation. It gives clear rules that are not difficult to comprehend and robust to outliers.

3.5.4 Support vector machine (SVM) algorithm

SVM being a supervised algorithm, is suitable for regression and classification purposes. The idea behind this algorithm is to separate the classes with linear hyperplanes. Data points nearest to the decision boundary (hyperplane) are the support vectors. Because of its impressive accuracy rate, SVM has been utilized in different areas of life [36].

Firstly, data importation is initiated on which predictions will be done. Thereafter, covert the dataset into the appropriate format by manipulating the data then observe the hidden patterns in the data by analyzing the data.

Subsequently, the partitioning of the data-set namely: “training”, “validation” and “testing” data-sets as measures to avoid overfitting, then develop and execute the four machine learning models. The training dataset will be used to train the model, after which validation will be done utilizing the validation data-set. Finally, the models will be tested on the testing data-set. A comparison of the four models will be done using the confusion matrix as shown in Table 3 for all models to ascertain their accuracy, precision, sensitivity, F-score, and Area under the Receiver Operating Characteristic (AUC & ROC Curve).

Table 3 Confusion matrix

	Actual Positive	**Actual Negative**
Predicted Positive	True Positive (TP)	False Positive (FP)
Predicted Negative	False Negative (FN)	True Negative (TN)

True Positive (TP): households or individuals with medical insurance, and are correctly predicted “covered”.
False Positive (FP): households or individuals without medical insurance but falsely predicted “covered”.
False Negative (FN): households or individuals with medical insurance but falsely predicted “not covered”.
True Negative (TN): households or individuals without medical insurance, and are correctly predicted “not covered”.

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP+FN+FP+TN}} \tag{2}$$

$$\text{Sensitivity or Recall} = \frac{\text{TP}}{\text{TP+FN}} \tag{3}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \tag{4}$$

$$\text{F Score} = \frac{2(\text{Precision}*\text{Recall})}{\text{Precision+Recall}} \tag{5}$$

Equations 2, 3, 4 & 5 above shows the mathematical computation of Accuracy, Sensitivity, Precision and F Score respectively. ROC curve is basically for comparison of model efficiency by computing the Area under the Curve (AUC). Its graph plots True Positive Rate (TPR), against False Positive Rate (FPR), at all thresholds with values between 0 and 1.

4. RESULTS AND DISCUSSION

In this segment, we shall focus our discussion on the effectiveness of all four created models and executed on the test dataset with three thousand six hundred and twenty-five (3625) entries. The data-set for testing is 20% of the whole dataset. Firstly, data correlation was conducted on the dataset to check the level of correlation of all predictors to the target feature.

Figure 1 below depicts correlation amidst predictors and the dependent variable. These values range from 1 to -1. The values between 1 and 0 show a positive correlation while values between 0 and -1 a negative correlation.

In Figure 1, the Income and North West geopolitical zone features both have a high correlation of 0.31 and 0.21 respectively with health insurance compared to the other features. Suffice it to say that households or individuals with income have higher tendencies of having their health insurance covered and the North West geopolitical zone has the likelihood of health insurance coverage compared to other geopolitical zones. The income feature also has a correlation of 0.29 with university education. This means that people with income tend to obtain a university education. The North West geopolitical zone (Zone_NW) has a correlation of 0.27 with Islamic education. This shows that people in the northern geopolitical zone acquire Islamic education more than in other zones.

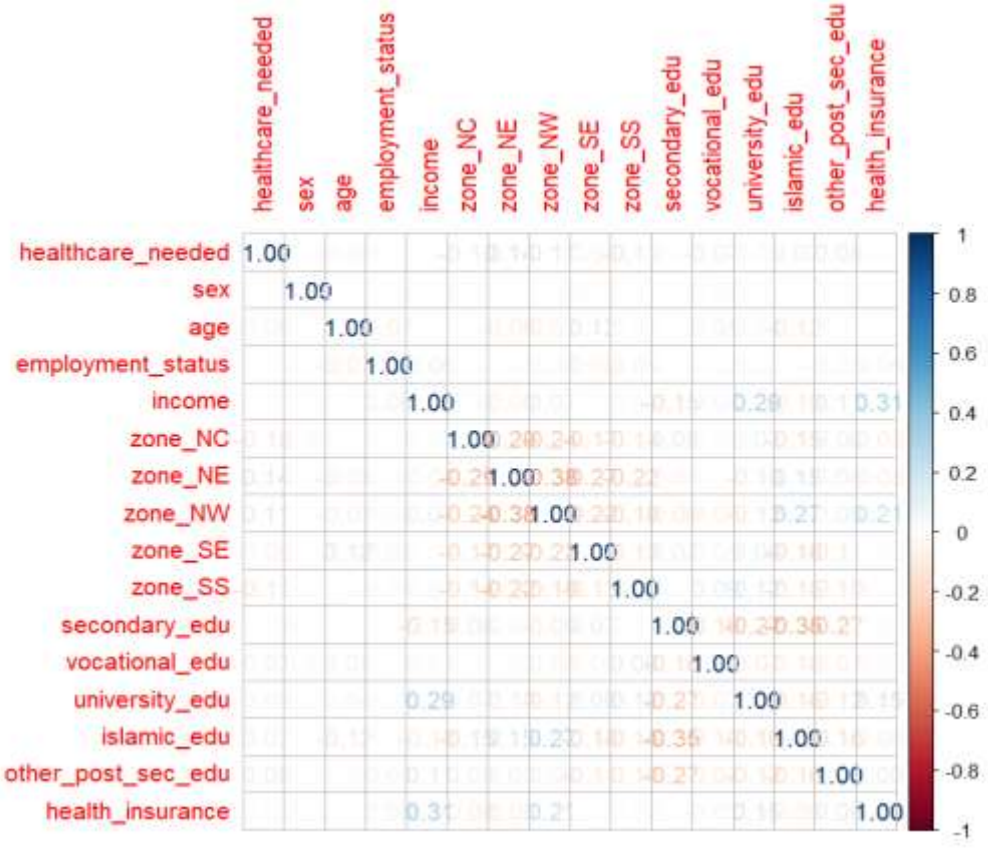

Figure 1: Correlation plot

4.1 Performance Metrics

Performance metrics are utilized to validate the efficacy and quality of machine learning models. It helps to measure how well a model carries out its task. To appraise the prediction of health insurance coverage in Nigeria, each model is assessed with the use of the confusion matrics. The confusion matrics of all various model was used to ascertain the accuracy, sensitivity, precision, F-score, and AUC & ROC curve for the model.

4.1.1 Logistic regression (LR) model performance Metrics

The logistic Regression model's performance could be assessed in several ways. Firstly, we evaluate the summary of the model (correlation of all predictors and the target feature), and the importance of every single independent variable. Secondly, the model's accuracy has to be assessed then validated [32]. Below is Table 4, the summary of the Logistic Regression model.

The asterisks (***) on the extreme right of the model summary in Table 3 above reveal the predictive power of each feature in the model. The likelihood that the true coefficient value is zero, irrespective of the estimated value given in the summary, is provided by the significant level (shown as "Signif. codes:" in the footer of the summary above). Three asterisks represent a significance level of 0, which signifies that the predictor is most likely to be associated with the target variable. Furthermore, a probability value (P-value) signifies a relationship between the predictor and the target. A threshold of 5% or 0.05 is used to establish a relationship. All predictors with a P-value less than 0.05 have a relationship with the target variable. The Logistic Regression computes the odd ratio by taking the exponential of each coefficient which is utilized to report the relatedness existing between the predictors and the target variable.

The odds of having medical insurance are about 14% lower for men compared to women, provided all other variables are constant, given the odds ratio of 0.858. The odds of having health insurance are approximately 2 times higher for the employed compared to the unemployed, provided other features are unchanged, given the odds ratio of 1.867. The odds of having health insurance in Nigeria are 6 times higher for those reporting to have an income than Nigerians without an income, provided all other variables are constant, given an odds ratio of 5.959.

The probability of medical insurance varies with geopolitical zones. Nigerians in North West zone had 13 times the odds of acquiring medical insurance in comparison to those living in North Central zone, ceteris paribus. Similarly, North East, South East, and South-South zones have odds of 2.8, 2.7, and 2.3 times higher respectively than North Central if all other variables are constant. Nigerians living in South West, however, were 69% lower odds of having health insurance compared with those living in North Central if all other variables are constant. Education appears to have an impact on who gets health insurance in Nigeria. The odds of people with a university education are about 6 times higher compared to those with primary or no education (odds ratio 5.878). Similarly, the odds of people with secondary and other post-secondary education are 3 and 4 times higher respectively compared to those with primary or no education (secondary education odds, 3.258 and post-secondary education odds, 4.239), provided every other feature is constant.

Below are Tables 5,6,7 and 8 illustrating the confusion matrix of Logistic Regression, Random Forest, Decision Tree, and Support Vector Machine models respectively.

Table 4 Logistic regression summary

	Estimate	**Std. Error**	**z value**	**Pr(>|z|)**	
(Intercept)	-5.8198700	0.2465731	-23.603	< 2e-16	***
healthcare_needed	-0.0947279	0.0861478	-1.100	0.2715	
Sex	-0.1764317	0.0784180	-2.250	0.0245	*
Age	0.0005245	0.0022069	0.238	0.8122	
employment_status	0.6245813	0.1483833	4.209	2.56e-05	***
Income	1.7850686	0.0865884	20.616	< 2e-16	***
zone_NE	1.0177628	0.1651416	6.163	7.14e-10	***
zone_NW	2.5801751	0.1570082	16.433	< 2e-16	***
zone_SE	0.9789545	0.1849285	5.294	1.20e-07	***
zone_SS	0.8377476	0.1973685	4.245	2.19e-05	***
zone_SW	-1.1851116	0.5262811	-2.252	0.0243	*
secondary_edu	1.1812020	0.1442859	8.187	2.69e-16	***
vocational_edu	-0.5866870	0.3698351	-1.586	0.1127	
university_edu	1.7712440	0.1589337	11.145	< 2e-16	***
islamic_edu	-0.0467619	0.1736214	-0.269	0.7877	
other_post_sec_edu	1.4443383	0.1578611	9.149	< 2e-16	***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Table 5 Confusion matrix of logistic regression model

	Actual Positive	**Actual Negative**
Predicted Positive	TP = 2187	FP = 69
Predicted Negative	FN = 1137	TN = 232

Table 6 Confusion matrix of the random forest model

	Actual Positive	**Actual Negative**
Predicted Positive	TP = 3316	FP = 207
Predicted Negative	FN = 8	TN = 94

Table 7 Confusion matrix of decision tree model

	Actual Positive	**Actual Negative**
Predicted Positive	TP = 3311	FP = 212
Predicted Negative	FN = 13	TN = 89

Table 8 Confusion Matrix for Support Vector Machine Model

	Actual Positive	**Actual Negative**
Predicted Positive	TP = 3290	FP = 167
Predicted Negative	FN = 34	TN = 134

4.1.2 Feature 1mportance

Below are Figures 2 and 3 result of Random Forest, this shows the feature importance of each variable in the data classification.

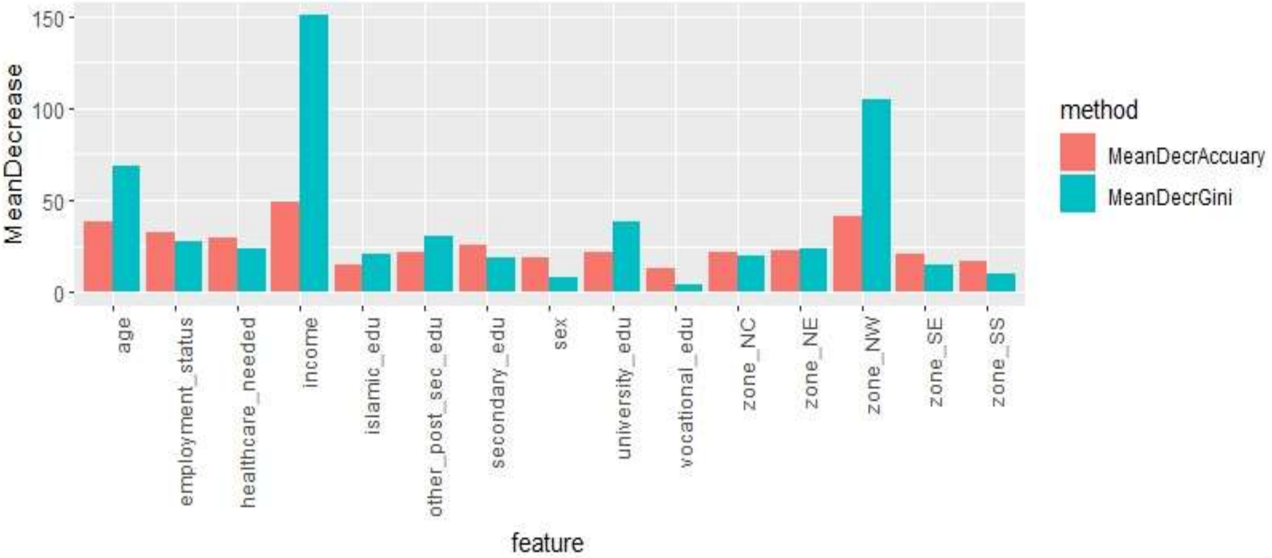

Figure 2: Feature importance

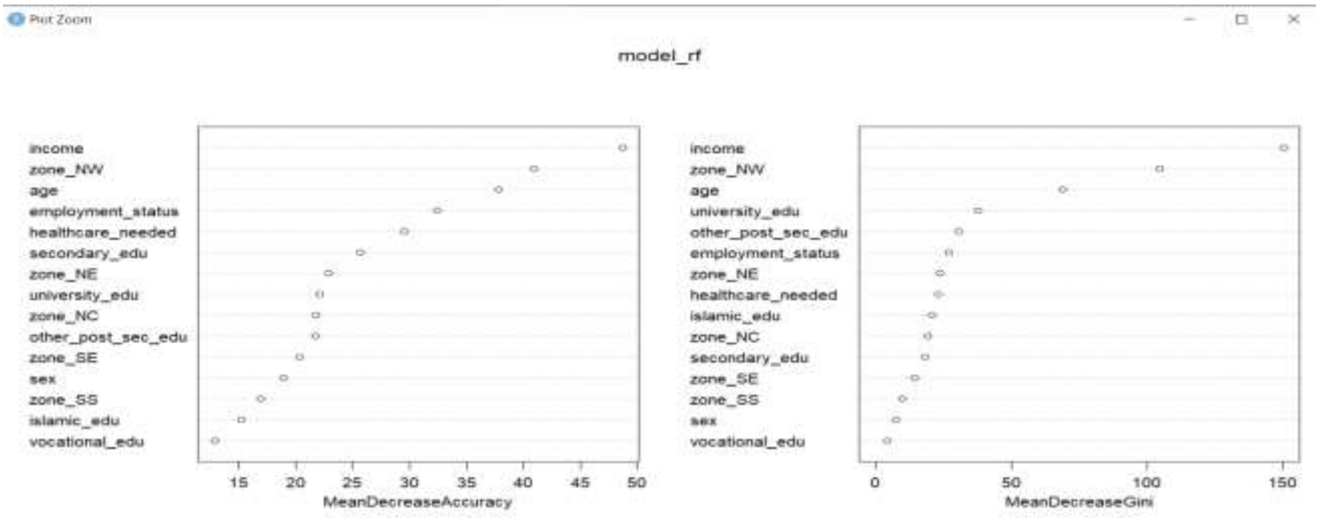

Figure 3: Variable importance plot

The Mean Decrease Accuracy plot as seen in Figure 2 above shows measure of accuracy loss in the model provided the variable is excluded. “The more the accuracy reduces as a result of the exclusion of the variable, the more important the variable is for successful classification” [35]. The Variables are determined according to their importance level. “The Mean Decrease in Gini coefficient measures how each variable contributes to the homogeneity of nodes and leaves in the resulting random forest. The higher the value of mean decrease accuracy or mean decrease Gini score, the higher the significance of the variable in the model” [35]. In Figures 2 and 3 above the variable “Income” is more important compared to the rest of the features. It suffices to say that financial income plays a major role to establish health insurance scheme in the community. Similarly, North West has a higher health insurance coverage compared to other geopolitical zones in the country.

4.1.4 Performance comparison of models

The performance metrics will enable us to conduct analytic comparison on all four models which will aid in identifying the model with better performance.

Table 9: Performance comparison of models

Metrics	Logistic Regression Model	Random Forest Model	Decision Tree Model	Support Vector Machine Model
Accuracy	66.73%	94.07%	93.73%	94.46%
Sensitivity	65.79%	99.79%	99.61%	98.98%
Precision	96.94%	94.13%	93.98%	95.17%
F Score	78.38%	96.86%	96.71%	97.03%
AUC	71.44%	65.49%	64.59%	71.75%

Above is Table 9 which illustrates the performance of all models using five different performance metrics. Logistic Regression has the least predictive accuracy 66.73% with an error rate of 33.27%, while the SVM with 94.46% accuracy rate indicates that SVM(Support Vector Machine) model is better in terms of model accuracy in prediction. A model's sensitivity ("true positive rate"), calculates the proportion of positive outcomes which were accurately predicted (Lantz, 2013). The logistic Regression model remains lowest in sensitivity with 65.79%, whereas random forest model scored highest with 99.79%. This indicates that random forest model is more sensitive to predict positive outcomes compared to the other models. In terms of precision ("positive predictive value"), Logistic Regression model is slightly better with 96.94%. However, the F Score metric also referred to as F-measure merges precision and sensitivity metrics using the harmonic mean. The harmonic mean is preferred to the arithmetic mean because both precision and sensitivity are expressed as proportions between 0 and 1 (Lantz, 2013). The Support Vector Machine model, once again, has a better percentage in F-Score and Area Under ROC Curve values scoring 97.03% and 71.75% respectively.

5. CONCLUSION

The utilization of machine learning in this research is concentrated on enhancing distribution and access to medical insurance across Nigeria. The core idea is that machine learning models can analyze large datasets to uncover critical challenges within the healthcare system, particularly those that impede the effectiveness of the National Health Insurance Authority (NHIA). A notable challenge ascertained is the insufficient financial income among large portions of the Nigerian population, which restricts their ability to participate in health insurance programs. By using machine learning to analyze financial data, the study highlights how income disparities contribute to the overall deficit in health insurance coverage. Understanding these financial barriers is essential for NHIA as it strives to accomplish Universal Health Coverage (UHC) by 2030. The vision to build a foremost government organization committed to acquiring the availability of funds for good medical services in Nigeria can be actualized when the populace have a constant income.

In addition, the study employs machine learning to identify geographical areas with varying levels of health insurance enrollment. For instance, the North-West geopolitical zone is observed to have the highest health insurance coverage, while other regions, particularly the South-West zone, have significantly lower enrollment rates. These insights are crucial for NHIA to direct resources and initiatives towards regions with lower coverage, thereby enhancing overall health insurance accessibility in Nigeria.

Furthermore, Support Vector Machine (SVM) algorithm is proven to be the best classifier among the models tested. The SVM algorithm was able to accurately forecast health insurance coverage, providing valuable predictions that can inform policy decisions. However, the study also recognizes the limitations of relying on just four machine learning models for prediction. It suggests that future research could benefit from exploring additional algorithms like K-Nearest Neighbours (KNN), Gradient Boosting, and Naïve Bayes, or even hybrid models, to potentially achieve more accurate and comprehensive predictions.

In summary, the utilization of machine learning is justified in this research, by its capability of providing detailed, data-driven insights into the financial as well as geographical factors that impact health insurance coverage in Nigeria. By training models on relevant data. The NHIA can develop more effective strategies to attain its goal of Universal Health Coverage by 2030, ensuring that all Nigerians get affordable, quality medical care.

REFERENCE

[1] Onwujekwe, O., Ezumah, N., Mbachu, C., Obi, F., Ichoku, H., Uzochukwu, B., & Wang, H. (2019). Exploring Effectiveness of Different Health Financing Mechanism in Nigeria; What Need to Change and How Can It Happen? BMC Health Service Research 19:661 https://doi.org/10.1186/s12913-019-4512-4

[2] Obikeze, E., Onyeje, D., Anyanti, J., Idogho, O., Ezenwaka, U., & Uguru, N. (2022). Assessment of Health Purchasing Functions for Universal Health Coverage in Nigeria: Evidence from Grey Literature and Key Informant Interviews. Health, 14, 330-341 https://doi.org/10.4236/health.2022.143026

[3] Awosusi, A. (2022). Nigeria's Mandatory Health Insurance and The March Towards Universal Health Coverage. The Lancet Global Health, 10, e1556.

[4] Baba, M., & Omotara, B., (2013). Nigeria's Public Health's Gains and Challenges

[5] Badawy, M., Ramadan, N. & Hefny, H.A. (2023). Healthcare Predictive Analytics Using Machine Learning and Deep Learning Techniques: A Survey. Journal of Electical Systems and Inf Technol, 10:40. https://doi.org/10.1186/s43067-023-00108-y

[6] Shaukat, Z., Zafar, W., Ahmad, W., Haq, I.U., Husnain, G., Al-Adhaileh, M.H., Ghadi, Y.Y. & Algarni, A. (2023). Revolutionizing Diabetes Diagnosis: Machine Learning Techniques Unleashed. Healthcare. 11, 2864. https://doi.org/10.3390/healthcare11212864

[7] Rahman, M.M., Rahman, A., Akter, S. & Pinky, S.A. (2023). Hyperparameter Tuning Based Machine Learning Classifier for Breast Cancer Prediction. Journal of Computer and Communications. 11**,** 149-165 https://doi.org/10.4236/jcc.2023.114007

[8] Ibrahim Baba Suleiman, Oluwasogo Adekunle Okunade, Emmanuel Gbenga Dada and Uchenna Christiana Ezeanya. (2024). Key Factors Influencing Students' Academic Performance. *Springer Open Journal of Electrical Systems and Information Technology*. 11(41). 1-18. https://doi.org/10.1186/s43067-024-00166-w

[9] Han, H.J. & Suh, H.S. (2023). Predicting Unmet Healthcare Needs in Post-Disaster: A Machine Learning Approach. Int. J. Environ. Res. Public Health, 20, 6817. https://doi.org/10.3390/ijerph20196817

[10] Chen, H., Wang, N., Zhou, Y., Mei, K., Tang, M. & Cai, G. (2023). Breast Cancer Prediction Based on Differential Privacy and Logistic Regression Optimization Model. Appl. Sci., 13, 10775. https://doi.org/10.3390/app131910755

[11] Sun, H.T. & Pan, J.N. (2023). Heart Disease Prediction Using Machine Learning Algorithms with Self-Measureable Physical Condition Indicators. Journal of Data Analysis and Information Processing, 11(1)**,** 1-10. https://doi.org/10.4236/jdaip.2023.111001

[12] Wei, Y.Z., Zhang, D., Gao, M.Y., Tian, Y.H., He. Y., Huang, B.I. & Zheng, C.Y. (2023). Breast Cancer Prediction based on Machine Learning. Journal of Software Engineering and Applications, 6**,** 348-360. https://doi.org/10.4236/jsea/jsea/jsea-2023.168018

[13] Reghunathan, R.K., Venkidusamy, P.N.P., Kurup, R.G., George, B. & Thomas, N. (2024). Machine Learning-Based Classification of Autism Spectrum Disorder across Age Groups. Eng. Proc., 62, 12. https://doi.org/10.3390/engproc2024062012

[14] Cai, M.Y. (2023). A Novel Method for Disgnosis of Breast Cancer Tumors Based on Random Forest. Journal of Biosciences and Medicines, 11**,** 252-259. https://doi.org/10.4236/jbm.2023.114018

[15] Gill, T.S., Shirazi, M.A. & Zaidi, S.S.H. (2023) Early Detection of Mesothelioma Using Machine Learning Algorithms Eng. Proc., 46, 6. https://doi.org/10.3390/engproc2023046006

[16] Chae, M., Yoon, H., Lee, H. & Choi, J. (2024). Hearing Recovery Prediction for Patients with Chronic Otitis Media Who Underwent Canal-Wall-Down Mastoidectomy. J. Clin. Med, 13, 1557. https://doi.org/10.3390/jcm13061557

[17] Zhang, Y., Zeng, H., Zhou, H., Li, J., Wang, T., Guo, Y., Cai, L., Hu, J., Zhang, X. & Chen, G. (2023). Predicting the Outcome of Patients with Aneurysmal Subarachnoid Hemorrhage: A Machine-Learning-Guided Scorecard. J. Clin. Med, 12, 7040. https://doi.org/10.3390/jcm12227040

[18] Santana, I., Sobrinho, A., Silva, L. D. D., & Perkusich, A. (2023). A Machine Learning for COVID-19 and Influenza Classification during Coexisting Outbreak. Appl. Sci., 13, 11518. https://doi.org/10.3390/app132011518

[19] Dipto, I.C., Islam, T., Rahman, H.M.M., & Rahman, A.A. (2020). Comparison of Different Machine Learning Algorithms for the Prediction of Coronary Artery Disease. Journal of Data Analysis and Information Processing 8(1), 41-68. https://doi.org/10.4236/jdaip.2020.82003.

[20] Zheng, H. (2018). Analysis of Global Warming Using Machine Learning. Computational Water Energy, and Environmental Engineering, **7,** 127-141. https://doi.org/10.4236/cweee.2018.73009

[21] Oyoo, J.O., Wekesa, J.S. & Ogada, K.O. (2024). Predicting Road Traffic Collisions Using a Two-Layer Ensemble Machine Learning Algorithm. Appl. Syst. Innow, 7, 25. https://doi.org/10.3390/asi7020025

[22] Almayyan, W. (2016). Lymph Disease Prediction Using Random Forest and Particle Swarm Optimization. Journal of Intelligent Learning Systems and Applications. 8**,** 51-62 http://dx.doi.org/10.4236/jilsa.2016.83005

[23] Colot, C., Baecke, P, & Linden, I. (2021). Leveraging Fine-Grained Mobile Data for Churn Through Essence Random Forest. Journal Big Data, 8**:**63. https://doi.org/10.1186/s40537-021-00451-9

[24] Getu, K. & Bhat, G. H. (2024). Application of Geospatial Techniques in Binary Logistic Regression Model for Analyzing Driving Factor of Urban Growth in Bhar Dar City Ethiopia. Heylion, 10**.** e25137. https://doi.org/10.1016/j.heliyon.2024.e25137

[25] Wang, J., Ju, T., Li, B., Huang, C., Xia, X. & Li, Li. C. (2024). Characterization of Tropospheric Zone Pollution, Random Forest Trend Prediction and Analysis of Influencing Factors in South-Western Europe. Environmental Sciences Europe, 36**:**61**,** https://doi.org/10.1186/s12302-024-00863-3

[26] Chen, H., Hu, S., Hua, R. & Zhao, X. (2021). Improved Naïve Bayes Classification Algorithm for Traffic Risk Management. EURASIP Journal on Advances in Signal Processing, 2021:30. https://doi.org/10.1186/s13634-021-00742-6

[27] Gai, R. & Zhang, H. (2023). Prediction Model of Agricultural Water Quality Based on Optimized Logistic Regression Algorithm. EURASIP Journal on Advances in Signal Processing, 2023:21. https://doi.org/10.1186/s13634-023-00973-9

[28] Liu, L., Luo, G. & Zhang, X. (2017). An algorithm based on logistic regression with data fusion in wireless sensor networks. EURASIP Journal on Wireless Communications and Networking, 2017:10. https://doi.org/10.1186/s13638-016-0793-z

[29] Hancock, J.T., Bauder, R.A., Wang, H. & Khoshgoftaar, T.M. (2023). Explanable machine learning models for Medicare fraud detection. Journal of Big Data. 10:154. https://doi.org/10.1186/s40537-023-00821-5

[30] Dahal, K.R. & Gautam, Y. (2020). Argumentative Comparative Analysis of Machine Learning on Coronary Artery Disease. Open Journal of Statistics, 10**,** 694-705. https://doi.org/10.4236/ojs.2020.104043

[31] Rahman, M.M., Rahman, A., Akter, S. & Pinky, S.A. (2023). Hyperparameter Tuning Based MAchine Learning Classifier for Breast Cancer Prediction. Journal of Computer and Communications. 11**,** 149-165 https://doi.org/10.4236/jcc.2023.114007

[32] Boateng, E.Y. & Abaye, D.A. (2019). A Review of the Logistic Regression model with Emphasis on Medical Research. Journal of Data Analysis and Information Processing. **7**, 190-207. https://doi.org/10.4236/jdaip.2019.74012

[33] Belgiu, M. & Dragout, L. (2016) Random Forest in Remote Sensing: A Review of Applications and Future Directions. ISPRS Journal of Photogrammetry and Remote Sensing. 114**,** 24-31. www.elsevier.com/locate/isprsjprs

[34] Shaik, A.B., & Srinivasan, S. (2018). A Brief Survey on Random Forest Ensembles in Classification Model. International Conference on Innovative Computing and Communications. 253-256

[35] Martinez-Taboada, F. & Redondo, J. I. (2020). The SIESTA (SEAAV Integrated Evaluation Sedation Tool for Anaesthesia) Project: Initial Development of a Malfactoral Sedation Assessment Tool for Dogs. PLoS ONE. 15(4): e0230799 https://doi.org/10.1371/Journal.pone.0230779

[36] Lantz, B. (2013). Machine Learning with R. Packt Publishing Ltd. P308.

[37] Cho, C.H., Yu, Y.W., & Kim, H.G. (2023). A Study on Dropout Prediction for University Students Using Machine Learning. Appl. Sci., 13, 12004. https://doi.org/10.3390/app132112004

[38] Nordin, N.I., Mustafa, W.A., Lola, M.S., Madi, E.N., Kamil, A.A., Nasution, M.D.,Abdulhamid, K.A.A., Zainuddin, N.H., Aruchunan, E, & Abdullah, M.T. (2023). Enhancing COVID-19 Classification Accuracy with a Hybrid SVM-LR Model. Bioengineering, 10, 1318. https://doi.org/10.3390/bioengineering10111318

[39] Zhang, J., Zhou, W., Yu, H., Wang, T., Wang, X., Liu, L. & Wen, Y. (2023). Prediction of Parkinson's Disease Using Machine Learning Methods. Bioengineering, 13, 1761. https://doi.org/10.3390/biom13121761

[40] Abbasi, E.Y., Zeng, Z., Magsi, A.H.., Ali, Q., Kumar, K. & Zubedi, A. (2023). Optimizing Skin Cancer Survival Prediction with Ensemble Techniques. Bioengineering, 11, 43. https://doi.org/10.3390/bioengineering11010043

[41] Olaguez-Gonzalez, J.M., Chairez, I., Breton-Deval, L. & Alfaro-Ponce, M. (2023). Machine Learning Algorithm Applied to Predict Autism Spectrum Disorder Based on Gut Microbiome Composition. Biomedicines, 11, 2633. https://doi.org/10.3390/biomedicines11102633

[42] Tu, K.-C., Tau, E.N.T., Chen, N. C. L., Chang, M.-C., Yu, T. C, Wang, C.-C., Liu, C.-F. & Kuo, C.-L. (2023). Machine Learning Algorithm Predicts Mortality Risk in Intensive Care Unit for Patients with traumatic Brain Injury. Diagonostic, 13, 3016. https://doi.org/10.3390/diagonostics13183016

[43] Rojek, I., Kotlarz, P., Kozielski, M., Jagodzinski, M. & Krolikowski, Z. (2024). Development of AI-Based Prediction of Heart Attack Risk as an Element of Preventive Medicine. Electronics, 13, 272. https://doi.org/10.3390/electronics13020272

[44] Singh, M.S., Thongam, K., Choudhary, P. & Bhagat, P.K. (2024). An Integrated Machine Learning Approach for Congestive Heart Failure Prediction. Diagnostics, 14, 736. https://doi.org/10.3390/diagnostics14070736