Child Mortality Prediction in Nigeria: A Machine Learning Approach

Muhammad Nuruddeen Idris
Computer Science Department,
Federal University Gusau,
Zamfara-Nigeria
minuruddeen66@gmail.com

Sahabi Ali Yusuf
Computer Science Department,
Ahmadu Bello University,
Zaria-Nigeria
sahabiali@yahoo.com

Mohammed Abdullahi
Computer Science Department,
Ahmadu Bello University,
Zaria-Nigeria
abdullahilwafu@abu.edu.ng

Isuwa Jeremiah
Computer Science Department,
Federal University of Kashere,
Gombe-Nigeria
isuwajeremiah@gmail.com

Ahmed Abdulrazaq Bello
Computer Science Departme
Federal Polytechnic Bauch
Bauchi, Nigeria
aabello.cs@fptb.edu.ng

Muhammad Lawal Tijjani
E-Extension Department
National Agricultural Extension and Research
Liaison Services(NAERLS)
Zaria, Nigeria
mltbnahmad@gmail.com

***Abstract*— Under 5-child mortality, which is the death of children under the age of five, is a crucial sign of the health of a nation which in turn affects its developmental strides. Despite international efforts to lower child mortality rates, Nigeria still has a heavy burden in this area. Nigeria accounted for 8% of worldwide child mortality in 2019, with an estimated 714,000 under-5 mortalities. These concerning numbers demonstrate the critical need for quick-acting efforts to reduce the mortality in the country. There has been a lot of study on child mortality in Nigeria, but majority of the approaches relied on traditional statistical analyses with weak predictive power. Although machine learning (ML) techniques can offer precise predictions of child mortality, no research utilizing ML techniques has been done in Nigeria. Therefore, this study used ML techniques to predict under-5 child mortality in the country. For this study, we ranked 30 factors that could influence under-5 child death and selected the top 10 features using data from the Multiple Indicators Cluster Survey (MICS). To predict the features that contribute to under-5 mortality, we utilized three machine learning models: decision tree, k-nearest neighbors, and logistic regression. The predictive power of each model was assessed using evaluation metrics. The decision tree model had the best performance in terms of predictive accuracy (92.49%), precision score of (94.48%), recall (94.49%) and F1 scores (94.48%), among all the models evaluated. According to the prediction models, there are several factors that affect the under-five mortality rate in Nigeria, including household residential area, household wealth index, child breast-feeding, type of toilet facility, child's sex, birth intervals, prenatal care, mother's education level, and mother's age at birth among others. For evaluating the risk factors for under-5 mortality, the decision tree machine learning model has superior predictive power, which could aid in improved policy making in this area. Understanding the causes of under-5 child mortality in Nigeria is made possible by our findings. Additionally, it shows that machine learning models are more effective than conventional analytical methods.**

Keywords—machine learning, child mortality, Nigeria, predictive models, evaluation metrics.

I. INTRODUCTION

Globally, child mortality continues to be a serious problem for public health, especially in low- and middle-income nations [1]. Sub-Saharan Africa and Southeast Asia are home to the majority of children who passed away before turning five, and the causes of these deaths can be avoided or treated [2]. Nigeria, the most populous nation in Africa, has high rates of child mortality that call for efficient interventions and legislative changes to address the problem. In Nigeria, one in every 13 newborns die before they turn one, and one in every eight does not live to turn five [3]. When compared to the levels in the early 1990s, these estimates show a significant decrease in under-5 mortality, yet "one preventable death of a child is still too many" [4]. The United Nations Development Program (UNDP) has made it clear that one of its goals is to reduce child mortality [5]. To achieve this goal, the UNDP advocates for changes to harmful conditions that increase the risk of child mortality as well as more research to better understand the specific conditions that contribute to child mortality in various ways. The United Nation Sustainable Development Goal and the Global Strategy for Women's, Children's, and Adolescents' Health (2016-2030) both highlight the critical need to eradicate unnecessary child deaths [6], [7]. Therefore, knowing the specific factors (or combinations of factors) that have the greatest impact on predicting under-5 child mortality will help create interventions that are more specialized and efficient.

The causes of under-5 child mortality are proximal and distal. Proximal causes are typically diseases or biological processes that end children's life, such as pneumonia, congenital defects, sepsis, diarrhea, and malaria [8]. It can already be too late to save a child's life by the time the proximate cause of death is discovered. Over the years, proximal causes of child mortality have deservedly received a lot of attention. But known and unknowable distal causes give birth to proximal causes. On account of this, the global health community should be equally concerned with identifying proximal causes of child death that are likely to result from distal causes. Many times, distal causes can be treated or even avoided, thereby eradicating proximal causes [9]. For instance, whereas malaria is a prevalent proximal cause of under-5 child mortality, inadequate sanitation or the use of contaminated water are significant distal causes of the proximal malaria that may lead to the under-5 child mortality.

Chen and Mosley (1984) were the first to outline the framework of distal, social-economic, and proximal variables as related to child mortality. Unfortunately, many developing countries, including Nigeria, are unable to address all of the factors that affect children's health at once due to their low financial resources, health budgets and high level of corruption. Therefore, it is becoming more and more crucial to identify the most crucial factors that should be prioritized as well as the most urgent socioeconomic problems that can serve as a springboard for the government and other stakeholders to concentrate on an intervention approach.

Moreover, in light of the commitment to the sustainable development goals (SDGs), analyzing the huge datasets that is still being produced by ongoing household surveys like the Demographic Health Survey (DHS) and Multiple Indicator Cluster Survey (MICS) may call for new methodological approaches [12]. To efficiently track the coverage and effects of health interventions, an expansion of the conventional analytical technique may be important and crucial. The development of systems with intelligence-like abilities is the goal of artificial intelligence (AI), which is "rooted in mathematics, philosophy, and computer science" [13]. A branch of artificial intelligence known as machine learning (ML), studies how computers can learn from example data to solve problems for which they were not explicitly programmed [14]. By extending current statistical models and providing new tools, AI and ML techniques enable public health milestones that might not have been previously possible. Machine learning (ML) techniques build on existing statistical approaches and use techniques that are not reliant on a priori assumptions about the distribution of the data in order to uncover previously unknown trends [15]. Numerous empirical research have been conducted on the various ML applications in hospital settings and mortality prediction around the globe [14], [16], [17]. However, the use of ML in population health studies in Nigeria, where it can be a game-changing tool, has yet to be proven. Our work aims to close the gap on ML application in population health studies across the nation. One of the previously mentioned gaps relates to the requirement to pinpoint the key causes of under-5 mortality in Nigeria. Instead of utilizing the conventional hierarchical strategy for multivariate analysis, which tends to be largely user-driven, we used a data-driven approach by performing feature selection to investigate these factors.

Previously, under-5 child mortality in Nigeria has been predicted using traditional statistical approaches. Several studies have identified factors associated with under-5 child mortality in Nigeria, including geopolitical zone, household socioeconomic status, parental care, birth order, birth interval, source of drinking water, children's sex, contraceptive use, and delivery method [18]–[20]. Bayesian hierarchical Poisson regression models have been used to estimate the factors influencing childhood deaths in Nigeria [21]. Additionally, the Auto-Regressive Integrated Moving Average (ARIMA) model has been employed for time series analysis to forecast under-5 mortality in Nigeria, predicting a reduction of up to 37.3% by 2030 [22]. Socioeconomic inequalities in under-5 mortality have also been investigated, highlighting the need for interventions to improve socioeconomic conditions and perinatal care. These findings provide valuable insights for policymakers and stakeholders in planning effective strategies to reduce under-5 child mortality in Nigeria. However, the examined literature demonstrates that there is a lack of research on machine learning modeling techniques for under-five mortality prediction in Nigeria. Huge amount of heterogeneous data generation is already taking place, and machine learning algorithms must interpret and present these health-related data. For successful action, policymakers and healthcare practitioners need precise forecasts on a range of health indices and socioeconomic factors. To produce findings that are extremely accurate, researchers must test and compare a variety of prediction and classification systems.

This research study is organized as follows: Section I introduce the entire research process, while section II present the methodology used to achieve the set out objectives. Section III present the feature selection techniques, whereas section IV the processes involved in the model selection. Section V present the results obtained while the discussion is presented in section VI. Section VII shows the key findings and contributions of the research study. Finally, section VIII present the implications of the findings as well as the recommendation.

II. METHODS

A. Dataset and Study Variables

The Multiple Indicator Cluster Survey (MICS), which is a global household survey created and supported by UNICEF, and conducted in 2021 by the National Bureau of Statistics (NBS) in Nigeria, provided the dataset for this study. MICS aims to gather statistically reliable and globally comparable estimates of the critical variables that are used to evaluate the condition of women and children in the areas of health, education, and child safety. Additionally, it is employed as a tool to produce data for tracking the advancement of national objectives and international commitments aimed at enhancing the welfare of children and women, such as the Sustainable Development Goals (SDGs). With a total sample size of 31,000 children drawn from 1850 clusters across the country, under-5-year-child mortality is the subject of our research. The study variables utilized in this study is based on specific information about children that was gleaned from Fertility/Birth history module, Women (Aged 15-49) module, Under-5 Children module and Household module. Child survival status, which is defined as the probability of under-5 child being alive or died before reaching the fifth birthday was the target variable for this study. The independent variables include residential area, mother's education level, prenatal care, wealth index, mother's age at first birth, gender of the child, age of child, child received vaccination, the spacing between births, children ever born, child ever breastfeed, the source of drinking water, toilet facility, the type of cooking fuel, geographical zone, health insurance, place of cooking, household have refrigerator, energy source for cooking among others. The available literature research on child mortality was used to help choose these variables [23], [24].

B. *Data Processing*

In order to optimize the number of variables available in the dataset, we manually select 30 indicators from the Households (HH), Children under five (CH), Fertility/Birth history (BH) and Women (age 15-49 years) (WM) modules from the 2021 MICS dataset. We specifically chose indicators from HH that were related to the availability of home appliances (such as having a refrigerator and/or electricity) or household demography (such as the number of households living in the house). We also took indicators from WM (such as whether they ever given birth and whether they are married). Then we chose indicators from BH (such as Mother's age at birth). We chose details like the child's gender and vaccination history from CH. We then linked the HH, CH, WM and HH data to form a single entity dataset. The list of the 30 selected indicators is presented in the Table 1 below.

LIST OF INDICATORS SELECTED FROM THE HH, CH, BH AND WM MICS MODULES

Acronyms	**Label**
HH2	Household Number
HC8	Household have electricity
HC9B	Household have: Refrigerator
EU1	Type of cook-stove mainly used for cooking
EU4	Type of energy source for cook-stove
EU5	Place of cooking
WS1	Main source of drinking water
WS15	Toilet facility shared
HHSEX	Sex of household head
Zone	Geopolitical zone
Windex5	Wealth index quintile
UB2	Age of child
UB6	Ever attended early childhood education programme
UB9	Covered by any health insurance
UCD2D	Shouted, yelled or screamed at child
BD2	Child ever been breastfed
IM11	Child ever received any vaccinations
Melevel	Mother's education
Chweight	Children under 5's sample weight
CM1	Ever given birth
CM8	Ever had child who later died
CM11	Children ever born
MN2	Received prenatal care
MA6	Marital status
WB5	Ever attended school
BH5	Still alive
BH3	Sex of child
Magebrt	Mother's age at birth
Birthint	Previous birth interval
Brthord	Birth order

To prepare the data for analysis, data cleaning and labeling were carried out using IBM SPSS statistical software. Data pre-processing (also known as feature selection) was done using Python version 3.10.7 in order to find the key attributes that may be used to predict factors related to under-5 child mortality.

III. Feature Selection

The process of automatically selecting important features for a machine learning model based on the kind of problem one is attempting to solve is known as feature selection (FS) [25]. We can accomplish this by adding/removing significant/insignificant features without altering them. FS assists in minimizing the amount of noise in our data and the quantity of our input data [26]. The basic goals of FS are to control the dataset's noise, outliers, and consistency. To assess feature relevance in relation to its predictive power, we employ the filter method approach. Features are chosen in filter methods based on the properties of the data rather than using learning models. This approach is fast and effective. Typically, a filter method consists of two phases. Features are ranked according to a set of criteria in the first step. The features with the highest rankings are selected in the second step. These criteria measure various aspects of the features, such as their dependence on the class label, their correlation with one another, and their dependence on one another [27]. To be more specific, SelectKBest of the scikit-learn library was utilized in our study. In machine learning, the SelectKBest feature selection strategy is used to choose the k-best features from a dataset based on specific scoring functions. It is frequently employed in classification or regression tasks for feature selection [28]. One of the score options for SelectKBest is the chi2 (chi-squared) statistical test. Categorical features in a dataset are tested using the Chi-square test. We determine the Chi-square between each feature and the target variable and then choose the features with the highest Chi-square values. In order to properly apply the chi-squared to test the relationship between various features in the dataset and the target variable, the following prerequisites must be satisfied: the variables must be categorical and sampled independently. We initialize SelectKBest with the chi2 scoring function and set k to 10 to find the top 10 features. Next, we apply the SelectKBest transformation to the dataset using fit transform function, which selects the top 10 features based on their chi2 scores, as shown in the fig. 1 below.

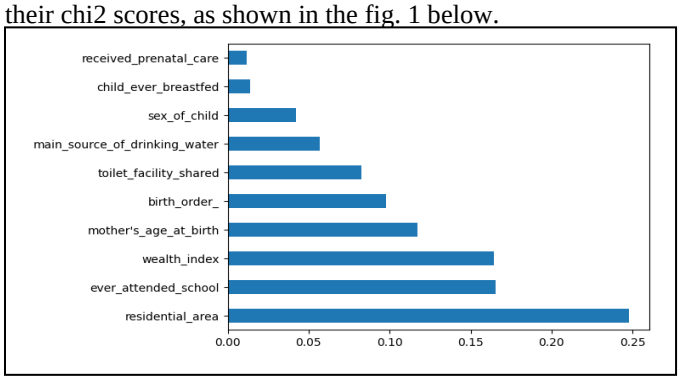

Fig.1: important selected features

IV. Model Selection

Building models that can produce predictions based on input data is the goal of predictive modeling, commonly referred to as supervised learning. These models are trained on labeled dataset where the target variable (the value to be predicted), in this case the probability of under-5 child been dead or alive, and the matching input variables are both known. The objective is to discover the underlying relationships or

patterns in the training data and then apply them to predict new, unseen data. We utilize three machine learning models in predicting the under-5 child mortality namely; Decision Tree (DT), k-Nearest Neighbor (k-NN) and Logistic Regression (LR). We use 75% of the data to train our models, while the remaining 25% is set aside for testing. This decision is consistent with the body of research and comes very near to the 80/20 training-to-testing ratio, also known as the Pareto principle [29]. For a number of reasons, these three models were chosen. In population health research, the logistic regression (LR) is frequently employed as an inferential tool to examine binary data, but it may also be applied as a binary classification model. When the study's dependent variable comprises features like decision-making (yes or no), disease state (ill or healthy), etc., LR is utilized. Multinomial regression analysis can be used to solve some complex LR models when the target variable has more than two categories [30], but certain presumptions must be made in order for a model to fit. However, the LR model does not rely on any form of linear relationship between the dependent variable and the independent variables. In the majority of research analyses, LR can be utilized to determine how the independent factors in a given study are affecting the dependent variable [31]. For instance, in healthcare, we can think about whether a newborn baby would live or die after an intervention, or the likelihood of an underweight or overweight newborn. Therefore, the LR model can be employed for this type of prediction purpose. Based on its capacity to identify both linear and nonlinear group neighbors, the k-NN model was selected. The key determining factor in this classifier is the k value, which stands for the quantity of nearest neighbors [32]. In order to determine its closest neighbors, the primary idea of k-NN relies on measuring the distances between the trained and tested data samples. The simplest machine learning algorithm is KNN. It has numerous uses in areas including pattern recognition, internet marketing, image analysis, healthcare, database clusters, etc. [33] A Decision Tree is a simple decision-making model that predicts the value of a target variable by learning simple decision rules from the data features [34]. A single decision tree can be represented as a flowchart-like structure. At each node, a decision is made based on a specific feature's value [35]. The process continues until a leaf node (terminal node) is reached, and the target variable's prediction is made based on the majority class in that leaf node. Decision tree can be used for both classification and regression tasks. In our research, a decision tree have been utilized to identify the most significant features in predicting under-5 mortality and create interpretable rules for decision-making.

A. *Evaluation Metrics*

The effectiveness of the above machine learning models in predicting under-5 child mortality must be evaluated using evaluation metrics. There are a number of metrics that are frequently used to assess the success of models in classification tasks, such as determining whether or not a child lived until the age of five or not. The specific evaluation metrics employed in this study are:

1) *Confusion Matrix:*

The confusion matrix is a tabular representation that summarizes the performance of a predictive model by comparing its predictions to the true labels. It consists of four entries as indicated in the fig. 2 below.

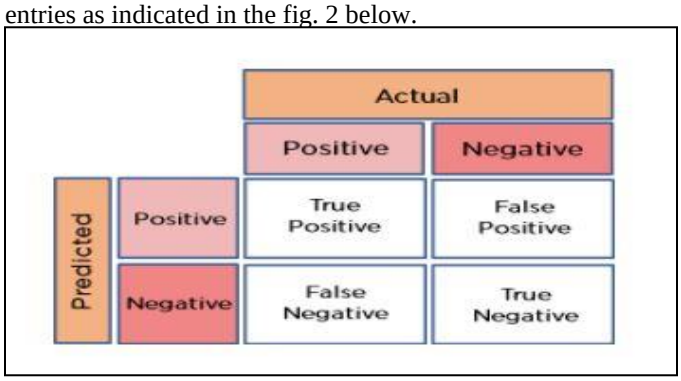

Fig. 2: confusion matrix

True Positive (TP): The number of correctly predicted positive instances (child is alive).

True Negative (TN): The number of correctly predicted negative instances (child is dead).

False Positive (FP): The number of incorrectly predicted positive instances (predicted as alive but did not live).

False Negative (FN): The number of incorrectly predicted negative instances (predicted as died but live).

The confusion matrix allows us to calculate other metrics such as accuracy, precision, recall, and F1-score.

a) *Accuracy*

The most fundamental evaluation metric is accuracy, which represent the proportion of correctly classified instances (both positive and negative) to the total number of instances in the dataset. It provides an overall measure of the model's performance.

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

2) *Precision*

Precision is the ratio of true positive predictions to the total number of positive predictions made by the models. It measures the ability of the model to avoid false positives.

$$Precision = \frac{TP}{(TP + FP)}$$

3) *Recall (Sensitivity):*

Recall is the ratio of true positive predictions to the total number of actual positive instances in the dataset. It measures the model's ability to identify positive instances correctly.

$$Recall = \frac{TP}{(TP + FN)}$$

4) *F1-Score*

The F1-score is the harmonic mean of precision and recall. It is a balanced metric that considers both false positives and false negatives. The F1-score is useful when the dataset is imbalanced.

$$F1 - Score = 2 * \frac{(precision * recall)}{(precision + recall)}$$

V. RESULT

A. Characteristics of the study population

The international community has kept the broad goal of lowering child mortality with the SDG target for child mortality, on ending preventable deaths of children under the age of five. The worldwide goal is to reduce under-5 mortality to at least 25 deaths per 1,000 live births [36]. The sample for the Multiple Indicators Cluster Survey (MICS), 2021 was created in order to provide estimates for a wide range of indicators on the condition of women and children at the national, rural, and urban levels for 36 states, the Federal Capital Territory (FCT), Abuja, and the six geopolitical zones of Nigeria. Our research's mortality rates were computed from the MICS using data gathered from the Women's module, Under-5 child module, Fertility/Birth history module, and household module.

Approximately 8.35% of the 31,000 children in the sample died before turning five as shown in Table 2. The greater part of the under-5 children (79.8%) came from rural areas. Most of the children in the survey are male child (54.23%), as compared with the female child (42.17%). More so, greater part of the children (56.8%), were born to women who had completed primary school, junior secondary and senior secondary schools, 38.63%, had no formal education, and 4.55%, had at least completed higher institution. 64.64% of the family are poverty-stricken. Compared to 36.20% of births by mothers with birth intervals of less than 2 years, the majority (63.80%) of births was given by mothers whose previous births were 4 years or more. The majority of the children (71.93%) were breastfed. The greater percentage (68.57%) came from households with unimproved water sources. While more than 65.61% of children were given birth by women who got prenatal care, and over 68.62% of children belonged to families with unimproved toilet facilities.

Further analysis revealed that there appears to be a significance difference in death prevalence between the genders, with male children experiencing greater mortality rates (8.80%) than female child (6.12%). The under-5 mortality rate was higher among children with 7+ siblings (7.10%), followed by among those with 4-6 siblings (5.29%) and then 2-3 years (3.54%). The least is those with none (2.08%). In the same vain, the under-5 mortality is higher in rural areas (12.41%) than in urban areas (5.83%). Similarly, poorer households record higher child mortality (13.32%), followed by second (9.53%), then middle (8.39%), 5.73% of the mortality are fourth, the least is the richest household (3.84%). Households who used unimproved water source such as unprotected dug well or untreated spring water had a considerably higher rate of under-5 mortality (7.62%) than those who utilized improved water sources such as piped or package water (3.34%). Antenatal care also showed significant differences, with under-5 mortality being much more common among children whose mothers did not receive antenatal care (9.06%) as compared to mothers who received antenatal care (4.90). Majority of the mortality are found in the children born to women aged less than 20 years (11.01%), compared to (6.62%) and (4.40%) of children given birth to mothers under the age of 20-34 and mothers who were 35 years or older, respectively. In general, lower death rates are related to higher educational attainment levels. The highest under-5 mortality rate is found in children born to mothers with no formal education (12.03%). As mother's education level rises, the under-5 mortality rates dramatically decrease. Finally, children who received vaccination consistently have lower rates of mortality (4.88%) than children without vaccination (10.32%).

DESCRIPTIVE ANALYSIS RESULTS OF THE STUDY POPULATION

Feature	*Frequency*	*percentage*
Child survival status		
a. Alive	28,410	91.65
b. Death	2,590	8.35
Residential area		
a. Rural	22,835	73.68
b. Urban	8,165	26.32
Ever attended school		
a. None	11,973	38.63
b. Primary	7,934	25.56
c. Junior secondary	6,694	21.59
d. Senior secondary	2,991	9.65
e. Higher/tertiary	1,409	4.55
Wealth index		
a. Poorest	10,019	32.32
b. Second	10,019	32.32
c. Middle	6,104	19.69
d. Fourth	3,916	12.63
e. Richest	939	3.03
Mothers' age at birth		
a. <20	13,898	44.83
b. 20-34	7,412	23.91
c. 35+	9,681	31.23
Birth order		
a. 1	5,862.1	18.91
b. 2-3	5,360	17.29
c. 4-6	7,877	25.41
d. 7+	11,895	38.37
Toilet facility shared		
a. Yes	21,273	68.62%
b. No	9,727	31.38%
Main source of drinking water		
a. Piped water	5,831	18.80
b. Tube well/ borehole	6,175	19.92
c. Dug well	6,705	21.62
d. Spring	4,845	15.63
e. Rainwater	2,608	8.41
f. Tanker-truck	1,401	4.52
g. Water kiosk	928	2.99
h. Packaged water	3,908	12.60
Sex of child		
a. Male	16,812	54.23%
b. Female	14,188	45.77%
Child ever been breastfed		
a. Yes	22,300	71.93%
b. No	8,700	28.06%
Received Prenatal Care		
a. Yes	20,342	65.61

b. No	10,658	34.38

B. Result of Predicting Child Mortality Rate

A predictive model of under-5 child mortality was created using the ML techniques approach models, specifically Decision Tree, Logistic Regression, and K-Nearest Neighbor classifiers. All the prediction models undergo training on a 75% training dataset and testing on a 25% test dataset. Utilizing evaluation criteria such as the Confusion Matrix, Accuracy, Precision, Recall, and F1 Score, the performance of the prediction models was assessed and compared. Our findings revealed that 2,228 children died before the age of five, whereas 24,743 children were still alive, as predicted by the logistic regression model. It incorrectly predicted that 2,192 children were still alive and that 1,837 children did not survive beyond five years. According to our research, the logistic regression model has an overall accuracy of 87.00% in predicting under-5 child mortality, with recall scores of 93.09%, precision scores of 91.86%, and F1 scores of 92.47%. The decision tree model accurately predicted that 26,018 children were alive and 2,112 children have died. It incorrectly predicted 1,518 children as been alive while 1,352 children died before the age of five. The decision tree model was found to have an overall accuracy of (92.49%) in predicting the under-5 child mortalities, recall scores of (93.09%), precision (91.86%), and F1 score (94.48%). 1,988 children died and 25,376 children were alive, as predicted by the k-nearest neighbor model. It incorrectly predicted 1,904 children as alive and that 1,732 children have died. Under-5 mortality was generally predicted with 88.27% accuracy, recall 93.61%, precision 93.02%, and F1 score (93.31%). Based on the results of the predictive models and their performance discussed above, Decision Tree was the most accurate model found in this study for predicting under-5 mortality as shown in Table 3 below.

PERFORMANCE COMPARISONS OF THE PREDICTIVE MODELS AS USED ON THE TEST DATASET

			Predictive models					
Performance metrics			***KNN***		***DT***		***LR***	
			Actual		***Actual***		***Actual***	
Confusion metric			***Alive***	***Dead***	***Alive***	***Dead***	***Alive***	***Dead***
	Predicted	Alive	25,376	1,904	26,018	1,518	24,743	2,192
		Dead	1,732	1,988	1,352	2,112	1,837	2,228
		%			%		%	
Accuracy		88.27			92.49		87.00	
Precision		93.02			94.48		91.86	
Recall		93.61			94.49		93.09	
F1-score		93.31			94.48		92.47	

VI. DISCUSSION

This work described how ML techniques were used to predict under-5 mortality in Nigeria. The study demonstrates that ML approaches are capable of outperforming conventional methods at predicting the variables linked to under-5 mortality. This is understandable given that ML techniques have been shown in numerous studies to perform better than conventional statistical techniques [15], [37]. These results were consistent with those of earlier research [38], [39]. In our research, we discovered that variety of bio-demographic and socioeconomic factors influence under-5 mortality in Nigeria. According to our research, residential area, household wealth index, mother's education level, and birth order were the four most significant risk factors associated with the under-5 child mortality. This finding was in congruent with similar studies [16], [40], who claimed that mother's age at birth, wealth index, birth order and type of toilet facility are the primary and comparatively more significant predictors of under-5 child mortality. They stated, for instance, that it has long been known that the pregnancy timing related with the healthiest outcomes for children occurs in mothers who gave birth at their 20s and that children born to extremely young (teenage) mothers are at an elevated risk of poor health outcomes and even death. Our research also discovered that breastfeeding duration was also one of the major predictor of under-5 mortality in the country. Previous research supports our findings [41], stating that children who are breastfed for longer period of time have lower rates of infectious disease morbidity and mortality as well as a higher probability of surviving than those who are breastfed for shorter period of time or not at all. Additionally, numerous studies have demonstrated that early breastfeeding initiation and exclusive breastfeeding lower child mortality [23], [42]. In addition to breastfeeding, we discovered that a number of other elements of the continuum of fundamental obstetric care, such as prenatal care and children vaccination, were significant predictors of death among children under the age of five. The provision of critical obstetric care is essential for survival during the child's first year of growing, according to a number of earlier studies. According to the research of [43] and [44], there is a substantial correlation between prenatal care and a lower risk of under-5 mortality. The consequence of children whose mothers do not received prenatal care and/or postnatal care services may encounter more close-by under-5 mortality risk factors than their counterparts, including congenital disorders and infectious diseases. Additionally, compared to children belonging to mothers with birth order of only one, our findings revealed that children with birth order of more than 7 are much higher at risk of dying. There is strong evidence to support the claim that larger birth order increase the likelihood that successive

children will survive [45], [46]. Under-5 mortality can be said to be influenced by a large birth order through three primary mechanisms: first, closely spaced births may exhaust the mother. The second effect involves the children competing for limited home resources, and the third involves the spread of contagious diseases between children who are close to one another [16]. Furthermore, we discovers that using unimproved drinking water is linked to a higher risk of mortality in children under the age of five. Insufficient access to clean water has been identified as the cause of 1,800 children deaths worldwide daily [47]. The link between household sanitation and access to clean water in promoting children's health and survival has also been demonstrated by numerous studies conducted in developing nations [10], [40]. Only 19% of the population in Nigeria has access to improved drinking water sources [48]. The variances in the country's under-5 mortality rate could be seriously impacted by this.

While the current work on utilizing machine learning approaches to predict under-5 mortality in Nigeria has produced insightful results, it is important to recognize some limitations that may affect the interpretation and generalizability of the findings. These limitations are inherent to the data, methods, and scope of the research and warrant careful consideration: Firstly, the research totally relies on the Nigeria 2021 MICS datasets as the primary source of information. Although the MICS surveys are comprehensive and widely used, they may have certain limitations such as reporting bias and data collection errors. These factors could introduce uncertainties in the analysis and potentially affect the accuracy of the predictive model. Second, the study concentrated on predicting under-5 mortality within the Nigerian setting using the Nigeria 2021 MICS datasets. As a result, care must be taken when extrapolating the results to other datasets or countries that have distinct demographic and socioeconomic characteristics. The model's performance might vary in other contexts, necessitating validation on different datasets for broader generalizability. Furthermore, the prediction of under-5 mortality inherently faces the challenge of class imbalance, with the number of instances representing child mortality being significantly lower than instances of no child mortality. While synthetic minority oversampling technique (SMOTE) were employed to address this issue, there remains the possibility of model biases, especially in favor of the majority class. Additionally, the selection of features for the predictive model was based on available data and domain knowledge. However, it is possible that other relevant features were not included, which may have contributed to a more comprehensive analysis. The omission of critical predictors could impact the model's ability to fully capture the complexity of factors influencing under-5 mortality. Yet again, three machine learning algorithms namely; decision tree, k-nearest neighbor and logistic regression were used for the predictive modeling. While these algorithms are commonly employed and have shown promising results, the selection of these models is not exhaustive. Different algorithms might yield different performance characteristics, and alternative models might be worth exploring to gain a more comprehensive understanding of the data. The study's dataset includes information up to the year 2021. Therefore, the predictive model is limited to this time period and might not take into account future changes in under-5 mortality patterns over time. Future research should consider updating the dataset to capture temporal variations. Machine learning models have the potential to impact real-world decisions, and the prediction of under-5 mortality carries significant ethical implications. Although every effort was made to ensure responsible and ethical use of the data and models, it is essential to acknowledge the potential for unintended consequences or biases in predictions. Finally, the evaluation metrics used, including accuracy, precision, recall, and F1-score, provide valuable insights into models performances. However, these metrics may not fully capture all aspects of the predictive model's effectiveness, and additional metrics or domain-specific evaluation may be warranted for a more comprehensive assessment.

While this study contributes to understanding the factors associated with under-5 mortality in Nigeria using machine learning, there are ample opportunities for future research. Areas of future work may include exploring additional features, incorporating more sophisticated machine learning algorithms, validating the model's predictions with external datasets, and considering dynamic models that account for temporal changes. By openly acknowledging these limitations, this research strives to maintain transparency and ensures a nuanced interpretation of the results. The insights gained from this study can serve as a foundation for further investigations into predicting under-5 mortality, ultimately contributing to evidence-based interventions and policy-making in addressing this critical public health issue in Nigeria.

VII. KEY FINDINGS AND CONTRIBUTION

A. Identifying Predictive Factors

Our study identified several important factors, including household residential area, household wealth index, sex of children, mother’s education level, source of drinking water, mother’s age at birth, birth order, child breastfeeding among others, that are related to under-5 mortality. These results are mostly consistent with the body of literature and emphasize how critical it is to address these issues in order to increase child survival rates.

B. Machine Learning Model Performance

The predictive models, including DT, KNN, and LR, demonstrated promising performance in predicting under-5 mortality. By employing resampling techniques to address class imbalance, we attempted to mitigate potential biases and improve the model's ability to predict child mortality accurately.

C. Limitations and Ethical Considerations

We recognize certain limitations in the research, such as data quality and generalizability. The class imbalance issue and the ethical implications of using machine learning models for predicting critical outcomes were thoroughly acknowledged. These limitations provide opportunities for future research and model refinement.

VIII. Implications and Recommendations

The findings from this study have several implications for public health policy and interventions aimed at reducing under-5 mortality in Nigeria. By understanding the key factors associated with the under-5 mortality, policymakers and healthcare professionals can develop targeted strategies and interventions to improve children health. Investing in maternal & child healthcare most especially in rural areas, providing poverty alleviation programs for poverty-stricken households, discouraging early girl child marriage, encouraging girl-child education and increasing accessibility to healthcare facilities in remote areas can contribute significantly to reducing under-5 mortality rates.

IX. Conclusion

In conclusion, this research contributes to the body of knowledge on under-5 mortality in Nigeria by harnessing the power of machine learning techniques. By identifying significant factors associated with the under-5 mortality and developing predictive models, we have taken a step towards evidence-based decision-making and targeted interventions to improve child survival rates. As we move forward, it is imperative to continually refine and validate predictive models, considering the dynamic nature of public health challenges. Ultimately, with collaborative efforts from policymakers, healthcare professionals, and researchers, we can strive towards reducing under-5 mortality and ensuring a healthier future for the children of Nigeria.

References

[1] T. A. J. Houweling and A. E. Kunst, "Socio-economic inequalities in childhood mortality in low- and middle-income countries : a review of the international evidence," Br. Med. Bull. 2010;, pp. 7–26, 2010, doi: 10.1093/bmb/ldp048.

[2] D. Y. Lucia Hug, David Sharrow, Kai Zhong, "Levels & Trends in Child Mortality 2018," United Nations Inter-agency Gr. Child Mortal. Estim. (UN IGME), 2018.

[3] UNICEF, "Infant, under-five and maternal mortality rates," N I G ER I A, 2015.

[4] T. Ayotunde, O. Mary, A. O. Melvin, and F. F. Faniyi, "Maternal Age at Birth and Under-5 Motality in Nigeria," East African J. Pablic Heal. Voulume 6 Number, no. 1, pp. 4–7, 2009.

[5] P. Conceição, United Nations Development Programme: Human Development Report 2021/2022. 2022.

[6] Pedro Conceição, Human Development Report 2019: Beyond income , beyond averages , beyond today: Inequalities in human development in the 21st century. 2019.

[7] "The Global Strategy for Women's, Children's and Adolescents' Health (2016-2030)," Sustain. Dev. Goals, 2016.

[8] B. O. Ahinkorah et al., "Proximate, intermediate, and distal predictors of under-five mortality in Chad: analysis of the 2014–15 Chad demographic and health survey data," BMC Public Health, pp. 1–12, 2020.

[9] N. S. Mugo, K. E. Agho, A. B. Zwi, E. Y. Damundu, and M. J. Dibley, "Determinants of neonatal , infant and under-five mortality in a war-affected country : analysis of the 2010 Household Health Survey in South Sudan," BMJ Glob Heal., pp. 1–15, 2018, doi: 10.1136/bmjgh-2017-000510.

[10] J. P. S. Cabral, "Water Microbiology . Bacterial Pathogens and Water," Int. J. Environ. Res. Public Health, pp. 3657–3703, 2010, doi: 10.3390/ijerph7103657.

[11] W. H. M. L. C. C. Chen, Lincoln C, "An Analytical Framework for the Study of Child Survival in Developing Countries," vol. 10, no. May 2023, pp. 25–45, 1984.

[12] D. J. Corsi, M. Neuman, J. E. Finlay, and S. V Subramanian, "Demographic and health surveys : a profile," Int. J. Epidemiol., no. November, pp. 1602–1613, 2012, doi: 10.1093/ije/dys184.

[13] A. E. Adegbosin, B. Stantic, and J. Sun, "Efficacy of deep learning methods for predicting under-five mortality in 34 income countries," BMJ Open, pp. 1–8, 2020, doi: 10.1136/bmjopen-2019-034524.

[14] J. Jaskari et al., "Machine Learning Methods for Neonatal Mortality and Morbidity Classification," IEEE Access, vol. 8, pp. 123347–123358, 2020, doi: 10.1109/ACCESS.2020.3006710.

[15] R. A. Trishan Panch, Peter Szolovits, "Artificial intelligence, machine learning and health systems," Viewpoints, vol. 8, no. 2, pp. 1–8, 2018, doi: 10.7189/jogh.08.020303.

[16] A. Bizzego et al., "Predictors of contemporary under-5 child mortality in low-and middle-income countries: A machine learning approach," Int. J. Environ. Res. Public Health, vol. 18, no. 3, pp. 1–12, 2021, doi: 10.3390/ijerph18031315.

[17] M. Podda, D. Bacciu, A. Micheli, R. Bellù, G. Placidi, and L. Gagliardi, "A machine learning approach to estimating preterm infants survival : development of the Preterm Infants Survival Assessment (PISA) predictor," no. May, pp. 1–9, 2018, doi: 10.1038/s41598-018-31920-6.

[18] O. K. Ezeh et al., "Under-5 Mortality and Its Associated Factors in Northern Nigeria : Evidence from 22 , 455 Singleton Live Births (2013 – 2018)," 2021.

[19] A. F. Fagbamigbe, "Approximation of the Cox survival regression model by MCMC Bayesian Hierarchical Poisson modelling of factors associated with childhood mortality in Nigeria," Sci. Rep., no. 0123456789, pp. 1–18, 2021, doi: 10.1038/s41598-021-92606-0.

[20] D. A. Adeyinka, N. Muhajarine, P. Petrucka, and E. W. Isaac, "Inequities in child survival in Nigerian communities during the Sustainable Development Goal era : insights from analysis of 2016 / 2017 Multiple Indicator Cluster Survey," BMC Public Heal. 201613, pp. 1–18, 2020.

[21] D. O. Eke and F. Ewere, "Modeling and Forecasting Under-Five Mortality Rate in Nigeria using Auto-Regressive Integrated Moving Average Approach," Earthline J. Math. Sci., vol. 4, no. 2, pp. 347–360, 2020.

[22] M. A. Ayoade, "Spatial and socioeconomic inequalities in under 5 mortality in Nigeria," African Geogr. Rev., vol. 00, no. 00, pp. 1–16, 2019, doi: 10.1080/19376812.2019.1605913.

[23] S. Yaya, M. Ekholuenetale, G. Tudeme, S. Vaibhav, G. Bishwajit, and B. Kadio, "Prevalence and determinants of childhood mortality in Nigeria," pp. 1–7, 2017, doi: 10.1186/s12889-017-4420-7.

[24] A. I. Bello, R. A, Joseph, "Determinants of Child Mortality in Oyo State , Nigeria," An Int. Multidiscip. Journal, Ethiop., vol. 8, no. 32, pp. 252–272, 2014.

[25] Y. Bouchlaghem, Y. Akhiat, and S. Amjad, "Feature Selection: A Review and Comparative Study," E3S Web Conf., vol. 351, pp. 1–6, 2022, doi: 10.1051/e3sconf/202235101046.

[26] S. Wang, J. Tang, and H. Liu, "Encyclopedia of Machine Learning and Data Science," Encycl. Mach. Learn. Data Sci., no. October 2017, 2020, doi: 10.1007/978-1-4899-7502-7.

[27] and H. L. Jiliang Tang, Salem Alelyani, "Feature Selection for Classification: A Review," Data Classif. Algorithms Appl., pp. 657–665, 2014, doi: 10.1201/b17320.

[28] A. Jović, K. Brkić, and N. Bogunović, "A review of feature selection methods with applications," 2015 38th Int. Conv. Inf. Commun. Technol. Electron. Microelectron. MIPRO 2015 - Proc., pp. 1200–1205, 2015, doi: 10.1109/MIPRO.2015.7160458.

[29] M. Mani, "Application of Pareto's Principle on Deep Learning Research Output: A Scientometric Analysis," Int. Conf. Mach. Learn. Smart Technol., no. October, 2021.

[30] N. R. Panda, J. K. Pati, J. N. Mohanty, and R. Bhuyan, "REVIEW ARTICLE A Review on Logistic Regression in Medical Research," Natl J Community Med, vol. 13, no. 04, pp. 265–270, 2022, doi: 10.55489/njcm.134202222.

[31] N. Srivastava, "Annales Geophysicae A logistic regression model for predicting the occurrence of intense geomagnetic storms," Ann. Geophys. 23, pp. 2969–2974, 2005.

[32] N. Ali, D. Neagu, and P. Trundle, "Evaluation of k - nearest neighbour classifier performance for heterogeneous data sets," SN Appl. Sci., vol. 1, no. 12, pp. 1–15, 2019, doi: 10.1007/s42452-019-1356-9.

[33] D. Dahiwade, G. Patle, and E. Meshram, “Designing Disease Prediction Model Using Machine Learning Approach,” 2019 3rd Int. Conf. Comput. Methodol. Commun., no. Iccmc, pp. 1211–1215, 2019.
[34] J. Kim, J. Lee, and Y. Lee, “Data-Mining-Based Coronary Heart Disease Risk Prediction Model Using Fuzzy Logic and Decision Tree,” Healthc. Informatics Res. HIP, vol. 21, no. 3, pp. 167–174, 2015.
[35] V. Podgorelec, P. Kokol, B. Stiglic, and I. Rozman, “Decision Trees : An Overview and Their Use in Medicine Decision trees : an overview and their use in medicine,” J. Med. Syst. · Novemb. 2002, no. Vol. 26, Num. 5, pp. 445–463, 2002, doi: 10.1023/A.
[36] UNICEF, “Levels & Trends in Levels & Trends in Mortality Child,” Estim. Dev. by UN Inter-agency Gr. Child Mortal. Estim., 2019.
[37] G. Nguyen, “Evaluating statistical and machine learning methods to predict risk of in-hospital child mortality in Uganda,” 2016.
[38] E. Mfateneza, P. C. Rutayisire, E. Biracyaza, S. Musafiri, and W. G. Mpabuka, “Application of machine learning methods for predicting infant mortality in Rwanda : analysis of Rwanda demographic health survey 2014 – 15 dataset,” BMC Pregnancy Childbirth, pp. 1–13, 2022, doi: 10.1186/s12884-022-04699-8.
[39] L. J. B. Caluza, “Machine Learning Algorithm Application in Predicting Children Mortality : A Model Development,” vol. 10, no. 1, pp. 1–6, 2018.
[40] J. M. K. Aheto, “Predictive model and determinants of under-five child mortality : evidence from the 2014 Ghana demographic and health survey,” BMC Public Health, pp. 1–10, 2019.
[41] G. A. Kayode, V. T. Adekanmbi, and O. A. Uthman, “Risk factors and a predictive model for under-five mortality in Nigeria : evidence from Nigeria demographic and health survey.sk factors and a predictive model for under-five mortality in Nigeria,” 2015.
[42] S. A. Adebowale, O. M. Morakinyo, and G. R. Ana, “Housing materials as predictors of under- five mortality in Nigeria : evidence from 2013 demographic and health survey,” BMC Pediatr., pp. 1–13, 2017, doi: 10.1186/s12887-016-0742-3.
[43] F. H. Bitew, S. H. Nyarko, L. Potter, and C. S. Sparks, “Machine learning approach for predicting under-five mortality determinants in Ethiopia : evidence from the 2016 Ethiopian Demographic and Health Survey,” Genus, 2020.
[44] G. A. Tesema, A. B. Teshale, and Z. T. Tessema, “Incidence and predictors of under-five mortality in East Africa using multilevel Weibull regression modeling,” Arch. Public Heal., pp. 1–13, 2021.
[45] A. E. O. ADETORO, G. W, “A Statistical Analysis of Child Mortality: Evidence from Nigeria,” J. Demogr. Soc. Stat. Vol. 1, Mach, pp. 1–11, 2014.
[46] N. Kozuki and N. Walker, “Exploring the association between short / long preceding birth intervals and child mortality : using reference birth interval children of the same mother as comparison,” BMC Public Health, vol. 13, no. Suppl 3, 2013.
[47] UNICEF, “Triple Threat: How disease, climate risks, and unsafe water, sanitation and hygiene create a deadly combination for children,” no. March, 2023.
[48] FMWR, “Water Sanitation and Hygiene National Outcome Routine Mapping Summary of Survey Findings,” 2021.