iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 12 **Issue:** VIII **Month of publication:** August 2024

DOI: https://doi.org/10.22214/ijraset.2024.63956

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 12 Issue VIII Aug 2024- Available at www.ijraset.com

Nigeria Poverty Prediction Using Statistical Analysis and Machine Learning

Salisu Abubakar Gambo[1], Zhang Diping[2]
School of Science, Zhejiang University of Science and Technology

Abstract: ***A multitude of deficiencies are referred to be poverty, including the inability to supply basic needs like clothing, food, clean water, and shelter. In the modern world, this also includes having access to education and health care. Numerous researchers make a great effort to comprehend, analyse, and forecast poverty using various methods. As a result, the Nigerian government was provided access to these other methods, which are ineffective. The traditional method of prediction in Nigeria involves site surveys, which are costly, labour-intensive, and a waste of time and energy before actual results are obtained. The predictions are also not accurate. The primary issue and barrier to making well-informed policy decisions and efficiently distributing resources in those areas that need the greatest assistance in Nigeria is the lack of trustworthy data on poverty. To determine which prediction model is more accurate for predicting poverty in Nigeria, we will attempt to compare many of them. The use of large data for measurement has been made possible by strategies based on machine learning development. Additionally, in this thesis, we will compare the prediction accuracy using models based on decision trees, binary logistics regression, and random forests. We will also utilise multiple correspondent analysis to do various analyses and a k means cluster to analyse the poverty level. The goal of the World Bank group is to eradicate extreme poverty and advance prosperity for all. To keep track of developments and comprehend the kinds of poverty alleviation tactics that are effective in Nigeria. To determine whether the world is on pace to eradicate extreme poverty, it is critical that we monitor poverty on a regular basis. For many years, a fundamental component of the World Bank's mandate has been the measurement and analysis of poverty, involving the discovery and dissemination of best practices and approaches for increasing the frequency and accuracy of poverty assessments. We can determine whether methods of poverty prediction are effective and ineffective for Nigeria by assessing poverty.***

Keywords: poverty measurement and prediction, machine learning, decision trees, binary logistic regression, random forests, multiple correspondence analysis, k-means.

I. BACKGROUND OF THE STUDY

In Nigeria, poverty is a significant issue that various scholars have defined differently. Measure it only in terms of salaries, even if experts feel that it also takes into account social status, political rights, health, and education. However, what unites all of these studies is the primary issue that led to and continues to produce poverty, as well as the search for a stress-free, straightforward method of identifying it. However, the goal of this research is to identify poverty through various statistical and economic analyses utilising machine learning techniques in order to classify poverty and learn how to identify and resolve issues permanently. Many researchers focus on just one approach, but in this case, we developed three approaches and compared their accuracy to determine whether each model could accurately forecast poverty in Nigeria. In addition, we evaluated each model's performance to determine its merits and limitations. Because this comparison has never been made in Nigeria previously, my work is all the more unusual and distinctive. Nigeria provided the data for this thesis. The entire bank studied poverty in Nigeria. Additionally, while their surveys and research are excellent, they are not precise enough to measure household poverty. Therefore, the data includes both statistical and economic analyses of poverty prediction, which we will incorporate into our four models to visualise and accurately predict. However, we are forced to exclude a significant proportion of missing values from our research and analysis for numerical variables like revenues and expenses. Since incomes are significant but not inclusive, we believe that this could be a reason for the lower models. The best way to predict the phenomenon in the first place is to use the multidimensional concept of poverty to determine the level of poverty in Nigeria. To do this, we must first identify the best cluster respondents for the household condition level of health, education, and many other issues that are related to poverty. We then used the method that classified respondents accordingly. Subsequently, we will employ our three machine learning models in the second technique and evaluate their respective performances.

A. *Statement Of The Problem*

Nigeria's population of over 160 million is split into ethnic and linguistic groups as well as religious divisions. The Hausa Fulani, who live primarily in the northern region and are Muslims, the Yoruba, who are divided into Muslims and Christians, and the Igbo, who live in the southeast, make up the Yoruba ethnic group. There are hundreds of distinct religious sects living in the middle belt. According to a 2020 National Bureau of Statistics study on Nigeria, 50% of the country's population makes less than $1 per day, while some people live below $380 per year. The research is based on data from the 2018–2019 Living Standard Survey, for which the entire bank provided significant support for the analysis. However, they still leave Borno out of the research because of the state's history of terrorism. However, the northern region of Nigeria, which has a high level of insurgency, has the greatest proportion of poverty in the nation: over 52% of Nigerians living in rural regions and only 18% in urban areas are considered to be living in extreme poverty. Nigeria's high and rising unemployment rate has also increased the country's level of poverty. The degree and trajectory of poverty in Nigeria have also been influenced by challenges facing the country's productive sector, growing income inequality, ineffective government that fosters corruption, social unrest, and environmental concerns. However, because 75% of rural residents rely on natural resources for their living, environmental deterioration limits the options that the poor have to earn a stable income. And after some investigation, I see that Nigeria is quickly turning into the global centre of poverty. Nigeria's poverty has surpassed India's to become the nation with the greatest percentage of the poor. Nigeria is larger than India by seven, yet it has surpassed India. Despite Nigeria's growth, over 9 million people still required humanitarian aid, and the country's poverty rate is rising annually as a result of the government's inability to identify the area's most in need of aid or those living in extreme poverty. In Nigeria, there is a significant poverty problem. Due to the present global financial crisis and its negative effects on the domestic economy, exports have decreased, putting millions of people at risk of falling into poverty by 2022. Nonetheless, protecting the impoverished from the crisis and reducing poverty even more must continue to be top priorities.

II. LITERATURE REVIEW

Millions of people in Nigeria are still affected by poverty, which hinders socioeconomic progress. Precise estimation of poverty is essential for efficient policy creation and resource distribution. This review of the literature looks at the methods used to predict poverty in Nigeria. It focusses on statistical analysis, machine learning, especially supervised learning, and poverty measurements.

A. *Poverty Measurements*

There are many different indicators that can be used to quantify the complex problem of poverty. Common measurements in Nigeria consist of:

1) *Income-based Measures:* These include the poverty line, defined as the minimum income required to meet basic needs. The World Bank's international poverty line of $1.90 per day is often used, alongside national poverty lines established by the National Bureau of Statistics (NBS).
2) *Consumption-based Measures:* These assessments evaluate household consumption patterns, often using surveys to gather data on expenditure on essential goods and services.
3) *Multidimensional Poverty Index (MPI):* This approach considers multiple deprivations experienced by individuals, including education, health, and living standards, providing a more holistic view of poverty.
4) *Qualitative Measures:* These involve subjective assessments of well-being and quality of life, often gathered through interviews and focus group discussions.

Because different methodologies may give diverse insights into the nature and degree of poverty, the choice of measurement has a considerable impact on the outcomes of poverty projections. Approximately 40% of Nigerians were impoverished just before the COVID-19 pandemic. The percentage of Nigerians living in poverty is known as the "poverty headcount rate," and it may be found by comparing the deflated per capita consumption aggregate with the poverty line—as previously mentioned. According to Figure 6, around 40.1% of Nigerians were living below the national poverty level, which is 137,430 naira per person annually. This indicates that 82.9 million Nigerians were below the poverty line. With 52.1 percent of the population living in rural areas vs 18.0 percent in urban areas, poverty was more concentrated there. 84.1% of Nigeria's impoverished population resided in rural areas. This indicates the spatial inequality in Nigeria all by itself.

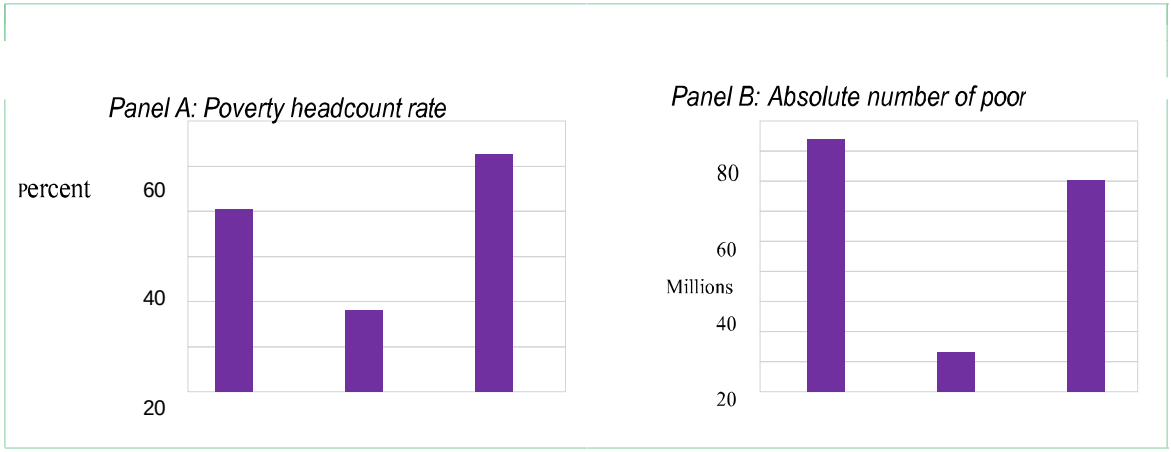


Figure 2.1: Poverty headcount rate and number of poor people in Nigeria in 2018/19, by urban-rural

B. Methods Of Machine Learning

Strong methods for evaluating and predicting complicated datasets are provided by machine learning (ML). The two most often used approaches in machine learning for learning are supervised learning and unsupervised learning. Thus, I'll offer a conversation that you may download that explains how machine learning is categorized. Since all we need to do is train our machines to learn and solve problems, we have chosen to approach our research problems using machine learning methodology. In this case, we would like to use household data from Nigeria to determine and predict the country's poverty level using machine models, a sort of machine learning classification. In the context of poverty prediction in Nigeria, several ML methods have been employed, particularly supervised learning techniques.

1) Supervised Learning

In supervised learning, algorithms are trained using labelled datasets that have known values for the outcome variable (such as poverty status). In order to anticipate poverty, the following techniques have gained traction:

a) *Binary Logistic Regression:* For binary classification issues, such as determining whether a household is above or below the poverty line, this statistical technique is frequently employed. By modelling the probability of the dependent variable depending on independent variables, logistic regression sheds light on the characteristics that affect poverty. We have currently included the logistic regression in the training set, which comprises 70% of the total data. The R has the following syntax and also we used the glm function in r.

We fit the logistic regression method to our model with the following R syntax:

multinom(formula, family = gaussian, data, weights, subset, na.action,
start = NULL, etastart, mustart, offset, control = list(...), model =
TRUE, method = "multinom.fit", x = FALSE, y = TRUE, contrasts = NULL, ...)

b) *Random Forest:* Using this ensemble learning technique, prediction accuracy is increased by building several decision trees and combining them. For complicated variable interactions such as poverty prediction, random forests are a good choice because of their ability to handle huge datasets with multiple variables and their resistance to over fitting.

c) *Random Forest Growing:* To grow the trees for our model, we used the Random Forest R package from R Studio. There is a lot of argument in this package, but we utilised the one we intended to use. The output of the random forest algorithm during training also includes the confusion matrix, thus the syntax used is:

randomForest(x, y=NULL, xtest=NULL, ytest=NULL, ntree=500, mtry=if
(!is.null(y) && !is.factor(y)) max(floor(ncol(x)/3), 1) else
floor(sqrt(ncol(x))), replace=TRUE, classwt=NULL, nodesize = if
(!is.null(y) && !is.factor(y)) 5 else 1, importance=FALSE, ...)

- MTRY: we decided to choose a formula which is $\sqrt{y} \approx 8$ because the predictor number itself must to be randomly choosing
- NTREE: we suggest not to used (0) or small number so to that we keep if in a default
- IMPORTANT: we will indicate or set the value as true if only the significant of the predictor should be assessed
- REPLACE: we used the default true for the indication sample of case should either be done with replacement

d) Decision Trees: Decision trees are simple models that visualise the decision-making process by dividing data into subsets according to feature values. They are especially helpful in comprehending the hierarchical connections between poverty status and predictions. The second step of supervised machine learning is to assess whether the obtained model performs well. The algorithm gives us predicted outcome levels which we compare with the true outcomes. When we reach high levels of accuracy, we can claim that our model predicts the concept well. If we obtain poor accuracy levels, we might have over-trained our model and need to go back and update the algorithm accordingly. The method that is widely used for the minimisation problems is explained next. Supervised learning is further divided into classification and regression. Let's, understand this. Next, we fit our second model using the decision tree method, fitting the decision tree method in R is supported by many packages but we chose to use the rpart package and the syntax used is:

rpart(formula, data, weights, subset, na.action = na.rpart, method,
model = FALSE, x = FALSE, y = TRUE, parms, control, cost, ...)

C. Classification Methods

To classify families according to their poverty condition, classification techniques are necessary. Alongside the previously described approaches, techniques like Support Vector Machines (SVM) and k-Nearest Neighbours (k-NN) have been investigated. Socio-economic datasets frequently contain high-dimensional data and non-linear relationships, both of which these classifiers can handle well. In supervise learning, classification refers to the process of identifying a model that facilitates the division of data into distinct categorical classifications. Based on the input parameters and the projected labels for the data, the classification assigns different labels to the data. When we refer to an output variable as a category, we simply mean that it can be either true or false, red or black, or yes or no. It forecasts and clarifies value. When a model is used to categorise data, it must determine the likelihood that each input data point (x) will belong to each output (y). This is known as a classification issue. The formula for the logistic function, known as the continuous log sigmoid function, is as follows: it takes an input that is in the negative to positive range and maps it to an output that is in the range of 0.0 to 1.0, such as weight, value, or price. Due to the fact that this model computes the mapping function that illustrates the relationship between the dependent variables (y) and the independent variable (x=(x1..., xp). But the classification isn't limited only to two classes for example classification method can help two assess whether a given image contains a boy or a girl. In this case the output will be 3 different value systems (1) the image contains a boy (2) the image contain a girl (3) the image contains neither boy nor a girl. The easy one of the classification is logistic regression because it estimates the probability of occurrence of an event based on one or more inputs.

D. Regression Methods

Regression techniques, such as linear regression and its variations, have been used to predict continuous measures of poverty, such as consumption expenditure or income levels. These techniques aid in the identification of relationships between poverty indicators and socio-economic factors, enabling a deeper comprehension of the dynamics of poverty in Nigeria. James et al. describe regularisation and dimension reduction as models that can enhance prediction accuracy and interpretability. Dimension reduction functions similarly to a zip file, eliminating irrelevant information by reducing the complexity of the data, transforming the predictors, and fitting the models using transform variables, which combine or reduce the input variable. And when we talk about the regularization which is also known as the (The shrinkage) it has a system of reducing the variance which help in handling the over fitting problems. We also have two regressions under the shrinkage which is the ridges and lasso which am going to explain about it below.

E. Regularization And Overfitting

Over fitting, in which a model learns noise in the training data instead of the underlying patterns, is one of the major problems in machine learning. This problem is especially relevant in the context of poverty prediction since socio-economic datasets are highly

dimensional. Regularization methods, such Lasso and Ridge regression, are used to penalize excessively complex models, which reduces over fitting and improves generalisation to unknown data.

F. Unsupervised Learning

Unsupervised learning algorithms work with unlabelled data and add value by assisting in the identification of patterns in the data without assuming any particular structure. We explained the idea of poverty using these kinds of algorithms. In the initial segment of our examination, we aim to determine if each individual is impoverished or not—a problem that is well-suited for the unsupervised learning approach. Which multiple correspondence analysis (MCA) will be used to analyse. Due to the fact that weights are included in the MCA result to provide a numerical poverty indicator. To ascertain which of the homes were impoverished and which were not, we consequently used a K-means approach. An approach that is frequently used to cluster numerical data is the K-means algorithm. It uses Euclidean distance to calculate how close individuals are to each other and clusters them based on their proximity. We will describe in more details how scree plot helped us in our decision in the section dedicated to the k-mean algorithm, and we will categorize the group with the lowest values of our poverty index as poor.

1) Unsupervised learning is based on the approach that it's useful when it's required to learn clustering or grouping of elements. Elements can be grouped (clustered) according to their similarity.

2) In unsupervised learning, data is unlabeled, not categorized and the system's algorithms act on the data without prior training. Unsupervised learning algorithms can perform more complex tasks than supervised learning algorithms.

3) Unsupervised learning includes clustering which can be done by using K means clustering, hierarchical, Gaussian mixture, hidden Markov model.

G. Weight

The energy and strength of the link between the units are represented by this weight. If there is a larger magnitude of weight between nodes 1 and 2, it indicates that neurone 1 has more impact over neurone 2. A weight reduces the input value's significance. Therefore, in a neural network, the weight of a connection is a numerical value that attempts to rearrange everything in an effort to simply eliminate errors. Every matrix element in the CNN filter is a weight that will be trained, and the extracted convoluted will be impacted by these weights. Every neural network calculates an output value derived from the layers by applying a function to the input values. Because the vector of weights and the biases distinguishing feature of CNNs are that many neurones can share the same filter and are also used across all receptive fields sharing that filter, as opposed to each receptive field having its own bias and vector weighting, neural networks attempt to adjust to these biases and weights. The functions that applied to input values are determined by the weight and bias.

H. Gradient Descent

Gradient descent is a mathematical optimisation technique that modifies the original function's factor and lowers some of the loss function's value. It is also an optimisation algorithm that searches for the optimal weight and determines the parameter (0) of a function (z) that minimises the cost (error) function and also reduces prediction errors. Gradient descent is used in machine learning algorithms to update the parameter by finding our models because it is an interactive method that is widely used in reducing cost risk function. The theoretical aspect are by Shalev.Sch wartz & Ben.David (2014). The algorithm seems to take a step in a negative direction and ensured the gradient convergence at a local minimum.

$$w(t+1) = w(t) - nof(w(t)) \quad \text{..... (1)}$$

We'll attempt to obtain the outcome as soon as T iterations are complete. The theory of Shaley, Schwartz, and Ben David (2014) will be utilised to gain further insight into the rate of convergence.It was used to train data models since it is simple to comprehend and apply, and any combination of algorithms has no effect on it. As a result, this kind of method minimises a specified function. That gradient's primary goal is to swiftly shift the weight downward to the regions with smaller mistakes; it is simpler and more effective to apply to convex functions than non-convex ones. so if the function is convex normally the local minimum is one of the global one. The loss function formula is

$$k(\theta 0,\theta 1) = \sum_{i=1}^{m} (h0(xi) - yi)2 \quad \text{..... (2)}$$

Additionally, the loss hypothesis function should be this formula, which combines the hypothesis function with the linear regression equation, while discussing the loss function, which evaluates the total errors between the samples and function output.

$$k(\theta o, \theta 1, ..., \theta n) = \frac{1}{2m} \sum_{i=0}^{m} (h\theta(xo, x1, ...xn) - yi)2 \quad \text{..... (3) .}$$

If we are to reduce the value of the loss function, we will for sure make microcosmic in other to analyze changes of function and also the formula for the loss function is

III. METHODOLOGY

A. Nigeria Living Standards Survey Data

The World Bank Living level Measurement Study (LSMS) provided the data required for training and assessing the machine learning models. The survey was conducted in 2018 and focused on a number of different facets of Nigeria's level of living.2019 saw the World Bank work with the National Bureau of Statistics (NBS) to perform all surveys. Information about households and communities is included in the survey. Questions regarding the household and its members were answered in a variety of household surveys, which included topics including health, education, assets, housing, employment, and income. 116,321 households participated in the survey, yet many of the responses were missing. Then we decided to clean it, after cleaning we were left with information for 8,229 households. The description of the data is presented in the table 3.1 below.

Table 3.1: Data Description

Dimension	Variable name	Value	Frequency	Percentage	Meaning
Physical capital	RANDWOThr	No Yes	5017 3219	0.6291301 0.4906699	Can the person read or write with any of Nigeria languages
Human capital	SCHOOL	No yes	1231 6971	0.2504436 0.9995564	Is the person educated
Human capital	RANDWENG	No Yes	3759 4670	0.454667 0.625353	Can the person read or write in English
Human capital	HEALTHY	No yes	7513 802	0.93205371 0.09787629	Was the person sick or injured recently
Human capital	EMPLOYED	No Yes	7765 574	0.9348619 0.0687381	Does the person has wage or salary job
Financial assets	ASSETS	No Yes	5948 2581	0.7168574 0.2896726	Does the household own any assets
Physical capital	TAP WATER	No Yes	6463 1976	0.9654619 0.0674381	Does the house hold have water
Physical capital	OWN TOILET	No Yes	5540 3489	0.6249202 0.3786798	Does the house hold have any kind of the toilet
Physical capital	ELECTRICITY	No Yes	2911 5318	0.3549489 0.6462961	Does the house hold has electricity from any kind of source
Financial assets	SAVING	No Yes	7875 594	0.9694619 0.0689381	Does the person has any saving

While this is the data we have gotten after cleaning it from our data survey, we used this data only for machine learning models in other to compare the accuracy of the models.

B. OPHI Multidimensional Poverty index data set

This data set that was used was gotten from Oxford poverty and human initiative (OPHI), as a result of the project carried out to measure a cute poverty all over the world. The data set was created in 2017 and it contains country the following:

1) Country: Nigeria
2) MPI Urban: Multi-dimensional poverty index for urban areas within the country
3) Headcount Ratio Urban: Poverty headcount ratio (% of population listed as poor) within urban areas within the country
4) Intensity of Deprivation Urban: Average distance below the poverty line of those listed as poor in urban areas
5) MPI Rural: Multi-dimensional poverty index for rural areas within the country
6) Headcount Ratio Rural: Poverty headcount ratio (% of population listed as poor) within rural areas within the country
7) Intensity of Deprivation Rural: Average distance below the poverty line of those listed as poor in rural areas

C. Multiple Correspondence Analyses

Our research began with a multi-dimensional poverty analysis using different correspondence analyses on our living standard data set. The variables in the data set, which were all coded as factor variables with two levels, underwent MCA. Additionally, we factored our data so that it was prepared for several correspondence analysis. The data set containing 8,229 respondents (households) was subjected to the MCA.

According to the findings, dimension 1 accounts for 26.6% of total inertia. We display the result visually. The figure depicts a scatter plot with Dimensions 1 and 2 displayed on the x and y axes, respectively. The primary axes, or distances from the columns, are represented by the values for these dimensions.

Upon examining and visualising the plot, we can observe that the categories linked to non-poverty are grouped on the left side of the plot, indicating a pattern in the data. Multidimensional poverty analysis (MCA) was employed to ascertain the household's poverty status by examining many household indicators. We employed function MCA and the Facto MineR package.

The function has the following syntax:
MCA(X, ncp = 5, ind.sup = NULL, quanti.sup = NULL, quali.sup = NULL,
graph = TRUE, level.ventil = 0, axes = c(1,2), row.w = NULL, method = "
Indicator", na.method = "NA", tab.disj = NULL)

We ultimately decide that since using the Burt was our option, we must utilise it.The default option for the number of dimensions is the only one we've retained. The data show a pattern, and you can see that all of the categories related to non-poverty are grouped together on the left side of the plot. The loading factor for each category, which highlights the total inertia collected by our algorithm, is the most important component of the result. Therefore, we utilise the loading factor in the following equation to get the poverty index: $PI = \sum_i (XijWj)$, since our thesis does not contain the entire output from the Multi Corresponding Analysis, we utilise the factors in order to do so. Additionally, the Wj in this factor indicates the weight of the jth variable, which is derived from the MCA, and PI stands for Poverty Indicator. However, the weight itself is a component that loads from Dimension 1. Which means that the weights that we assign to each category are a reflection of the MCA's result. And according to what we visualize on our map is that the negatives scores of our variables are associated with the poor while the positive scores are associated with the none poor .So our next step was that we will used the K-means algorithm in other to record the poverty index into a binary variable.

D. K-means algorithm

Finding the poverty levels is the second stage of our investigation. For this issue, K-means clustering is an appropriate option. The requirement to ascertain the number of clusters is one issue with this technique. It is not always the case that we are immediately aware of this.

Making a scree plot with various selections for K and their corresponding within-cluster variation is one method to solve this problem. Figure 3's scree graphic for our data illustrates how adding more clusters reduces the within-cluster variance. The point at

which the variance starts to decrease makes sense and is the best option for K. We can spot this point by visual inspection of the scree plot as it resembles an elbow (Ng. 2015).

Figure 3.1: Number of cluster

We utilised the K-mean clustering algorithm in R, and the function syntax is as follows: as we can see on the screen plot of our data set, the variance faced a big decrease. We chose three clusters, but the function only offered four, so we ended up choosing the fourth group.

k-means(x, centers, iter.max = 10, nstart = 1, algorithm = c("Hartigan.Wong", "Lloyd", "Forgy", "MacQueen"), trace=FALSE)
We try to use this following arguments for our function because to that it can only give us the output that we want throw the argument.

1) x . this is the matrix of data, and its numeric; The index in our case is poverty
2) centers . in this argument the number of clusters that we used is 4 instead of 3 And we also include the scatter plot for our poverty index
3) The x.axis it represents an index for all observations and while the y.axis shows the main poverty index. The color of the coding is based on the clustering that we obtained from K-Means. We also categorize the poor as those who belong to the black group. In this way we try to classified our data set.

We are working to establish an additional variable called "poverty" for the following set of algorithms. It has two values: 0 for non-poor and 1 for impoverished. Thus, the discussion of unsupervised learning methods comes to a close in this part.

IV. RESULT AND DISCUSSION

We then looked at the OPHI data to analyse it and comprehend poverty in Nigeria. First, we attempted to see the regions of the country with the highest levels of poverty by viewing Nigeria as a whole. Next, we dissected the data to identify the states with the highest and lowest rates of poverty. The outcome of our analysis was visualised as follows:
First of all we visualize the global world multidimensional poverty inde

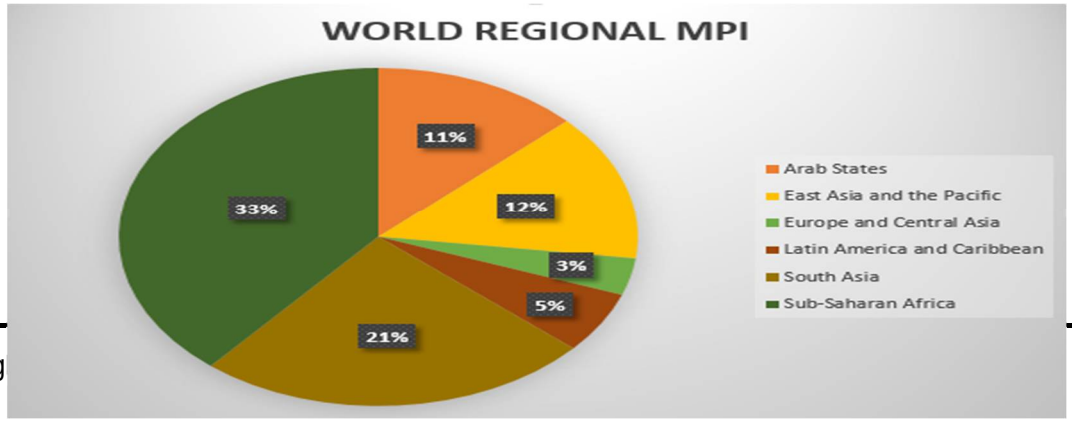


Figure 4.1: world multidimensional poverty index

From the pie chart above, it is clear that sub-Saharan Africa has the greatest multidimensional poverty index, with 33 percent of the world's population living below the poverty line. Since Nigeria is situated in this region, we will now turn our attention to it.

As we can see from the plot above, Nigeria and Chad have the highest percentages of people living in poverty. Let's zoom in and examine Nigeria. Nigeria, which has 36 states including its capital, is one of the most economically significant countries in Africa. Its two areas, the North and the South, are further divided into six geopolitical zones: North East, North West, North Central, South East, South West, and South South. We'll examine the country's data and contrast its many areas and sub-regions to determine which has the highest percentage of the population living in poverty. Based on the plots, we can see that Yobe has the greatest proportion of the poor in the nation, while Lagos has the fewest; Zamfara has the highest headcount ratio, while Lagos has the lowest; Yobe has the highest intensity of deprivation, and Ekiti has the lowest intensity. Since the north of the country is more rural and has a higher population, it makes sense that there are more poor people living there who lack access to adequate education, clean water, and a decent standard of living.

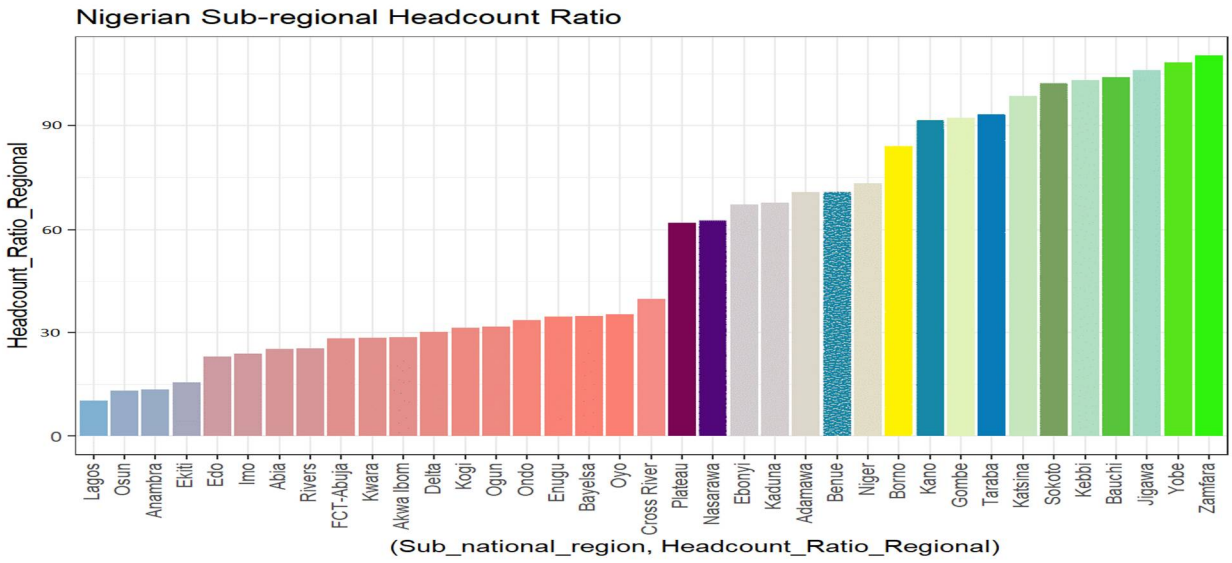


Figure 4.2: Nigeria sub-regional multidimensional poverty index

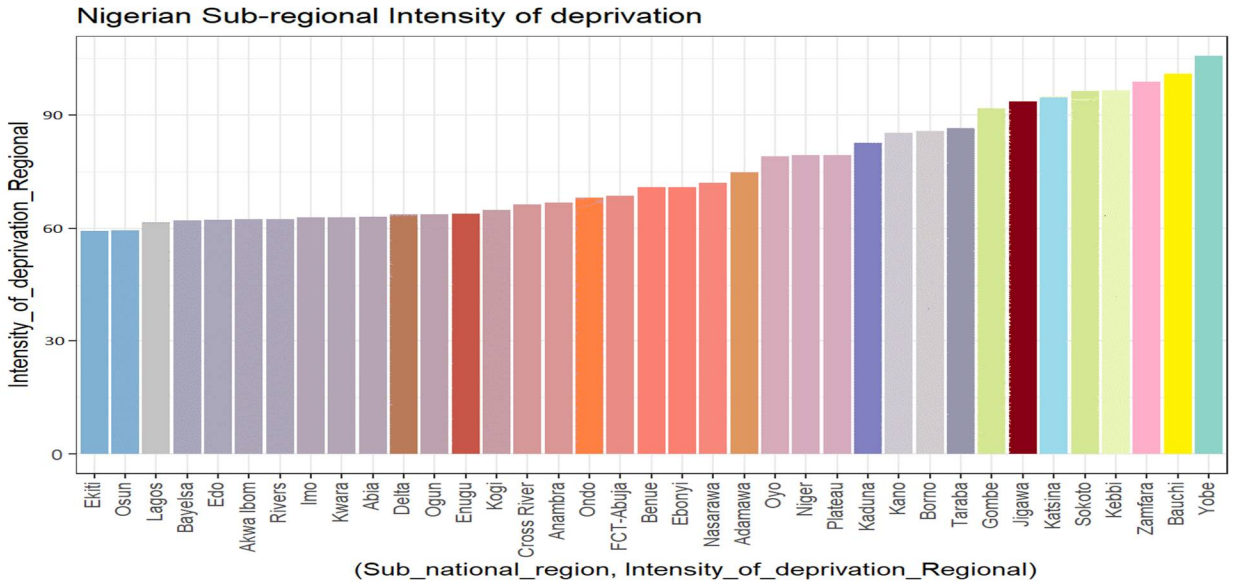


Figure 4.3: Nigeria sub-regional headcount ratio

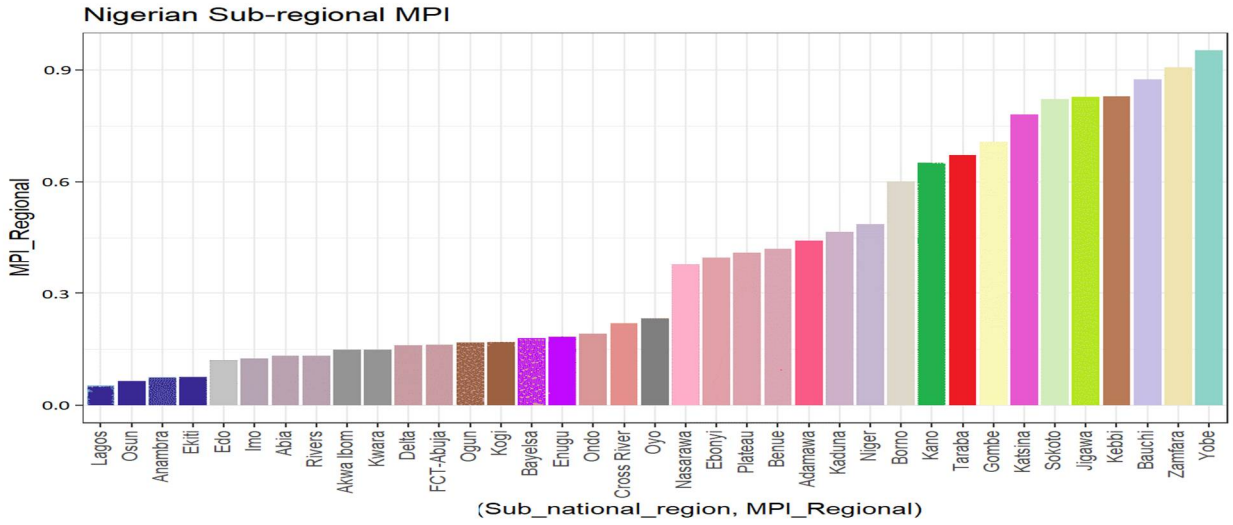


Figure 4.4: Nigerian sub- regional intensity of deprivation

A. Output Of The Models

Convolutional neural networks, logistic regression, random forests, and decision trees are the four models we will fit and compare to detect poverty. We will also drop the household ID since it is not relevant, and we will convert our target column to a factor so that it can fit into our models and provide the result after our factors 5761 samples with 10 predictors and 4 classes—"1,"","2,""3," and"4"—with no preprocessing are used to fit the models. Resampling was cross-validated ten times. The sample size is summarised as follows: 5185, 5185, 5185, 5186, 5185, 51884, and 5184. The results of the resampling across tuning parameters are as follows, based on each model.

B. Random Forest

The tuning parameter "min.node.size" was kept constant at 1, and the best model was chosen using the greatest value by applying kappa. The model's final settings were min.node.size=1 and mtry=6, splitrule=extratrees.

Table 4.1 Random forest result

MTRY	SPLITRULE	ACCURACY	KAPPA
2	gini	0.9855924	0.9770885
2	extratrees	0.9854191	0.9768125
6	gini	0.9854188	0.9768113
6	extratrees	0.9857660	0.9773656
10	gini	0.984377	0.9751518
10	extratrees	0.9843771	0.9751518

C. *Binary Logistic Regression*

We used the package $multinum and also Kappa was used to select the optimal model using the largest value. The final value used for the model was decay = 1e.04.

Table 4:2 binary logistic regression result

Decay	Accuracy	kappa
0se+00	0.9855933	0.9770894
1e.04	0.9857669	0.9773632
1e.01	0.9854197	0.9768092

D. Decision Tree

We used the package $rpart and also Kappa was used to select the optimal model using the largest value. The final value used for the model was cp = 0.08540576.

Table 4.3 Decision tree result

cp	Accuracy	kappa
0.08540576	0.9055744	0.8440188
0.11649215	0.8476229	0.7382927
0.65837696	0.6772289	0.4071587

E. Accuracy Comparison

The decision tree model, which has the lowest accuracy, performed poorly until some of its hyper parameters were adjusted to improve performance, but even then, it is still the least performed of the three. The figures below summarise the three models. The random forest-based model performed similarly to the logistic regression model. We matched our findings to their performance. Without generating a distinct prediction vector, we extract accuracy from CNN in a single step by class. The results are as follows.

Table 4.4 models accuracy comparisons

NAME	ACCURACY	SENSITIVITY	SPECIFICITY	SPEED
Binary logistic regression	0.9882496	0.9936292	0.9952964	15.11elapsed
Decision tree	0.8666937	0.6999812	0.9429608	1.09 elapsed
Random forest	0.9882496	0.9936292	0.9952964	33.34 elapse

Figure 4.5: Plotting the results for speed vs accuracy comparison

The weighted harmonic mean of the test's precision and recall is known as the F1 score, or F score, and it serves as a gauge of the test's accuracy by balancing precision and memory. It is further explained by the formula below. F is equal to 2 * (recall * precision) / (recall + precision).

Table 4.5 viewing the data frame of models

NAME		F1.SCORE	MACRO F1	WEIGHTED F1
Binary logistic regression	MULTINOM	1.0000000	0.9928992	0.98827651
Decision tree	RPART	0.9185431	0.9342116	0.82207943
Random forest	RANGER	1.0000000	0.9928943	0.9882656

V. DISCUSSION

Our goal was to predict poverty using the data set we had acquired, and then to compare the predictive power of our models. Our data was divided into four groups—1, 2, 3, and 4—after the k means operations. We then factored the four categories, resulting in two—poor and non-poor. We factored out our data to establish two groups because there is a close association between 1, 2, and 3, 4 and the k means helped us create those categories based on their clusters. The next step involved training our models after the new target column was added to the dataset based on the k-means clustering result. However, before we could do that, we split the data into two groups: 70% for training and 30% for testing. Once the data was split, we chose the machine learning algorithms we wanted to use and got them ready for training. The training data was used to fit the three trained models—Random forest, decision tree, and logistic regression—and subsequently the test data was used to gauge the models' performance and generate predictions. The model predicted poverty based on the poverty levels in the target column which was either 1, 2, 3, or 4 that is 1 or 2 for non.poor and 3 or 4 for poor the result of the predictions is illustrated in the confusion matrix below

Table 4.6 Confusion matrix of Logistic regression and random forest

	1	2	3	4
1	250	0	0	0
2	0	904	27	0
3	0	2	1133	0
4	0	0	0	152

Table 4.7 Confusion matrix of decision tree

	1	2	3	4
1	0	0	0	0
2	111	904	54	23
3	139	2	1106	0
4	0	0	0	129

The confusion matrix above illustrates how close the predictions generated by the random forest and logistic regression models were. According to the matrix's interpretation, (1) was properly predicted 250 times, whereas (2), (3), and (4) were incorrectly forecasted zero times, and so on. Furthermore, according to the four previously mentioned poverty levels, the models indicated that 1154 households would not be poor and 1285 households would be poor, for a total of 2439 correct forecasts and 29 incorrect predictions. Let's now discuss the models' accuracy, which was determined from the confusion matrix above using the following formulas and the formation:

Accuracy = (TP + TN) / (TP + TN + FP + FN)

Sensitivity (recall) = TP / (TP + FN)
Specificity = TN / (TN + FP)
Precision = TP / (TP + FP)

In our model above, where the observation was (1) and the prediction was (1) 250 times, a true positive (TP) is an observation that is positive and is projected to be positive. A false positive (FP) occurs when an observation is negative but is anticipated to be positive. For example, in our model, an observation of (2) was predicted twice, but the result was (3). A true negative (TN) indicates that both the observation and the prediction are negative. An observation that is positive but a negative prediction is called a false negative (FN). The weighted harmonic mean of the test's precision and recall is known as the F1 score, or F score, and it serves as a gauge of the test's accuracy by balancing precision and memory. The formula below explains it further F = 2 * (precision * recall) / (precision + recall)

After reviewing the predictions made by the models how accurately they have done it is clear that the logistic regression and random forest based models were more accurate and the decision tree was less accurate even though being faster. To further explore our findings, let's examine how our models generated their predictions and identify the features that drove them. Using our MCA, we were able to observe that households with similar characteristics were clustered together, i.e., most households without assets, no job, or no education are assumed to be poor, while households with these characteristics are assumed not to be poor. Additionally, the k-means algorithm was able to classify these households using the same data, indicating that our models also made use of those features. What sets our work apart from others' is that, although most studies utilise survey data or satellite imagery to forecast poverty, we used both. We then performed multidimensional poverty analysis on the data to analyse it, and in the end, we were able to determine how the two approaches could be combined to better understand poverty.

Because survey data is akin to multidimensional poverty analysis, which suggests that poverty has several dimensions and is about much more than just income, infrastructure, or healthcare, we may say that survey data is better suited for in-depth examination of poverty. The purpose of the study was to assess poverty in Nigeria using data from the Nigeria Living Standard Survey, forecast poverty using the same data using a straightforward machine learning technique, and then compare the two approaches to determine which is more helpful in comprehending poverty in Nigeria.

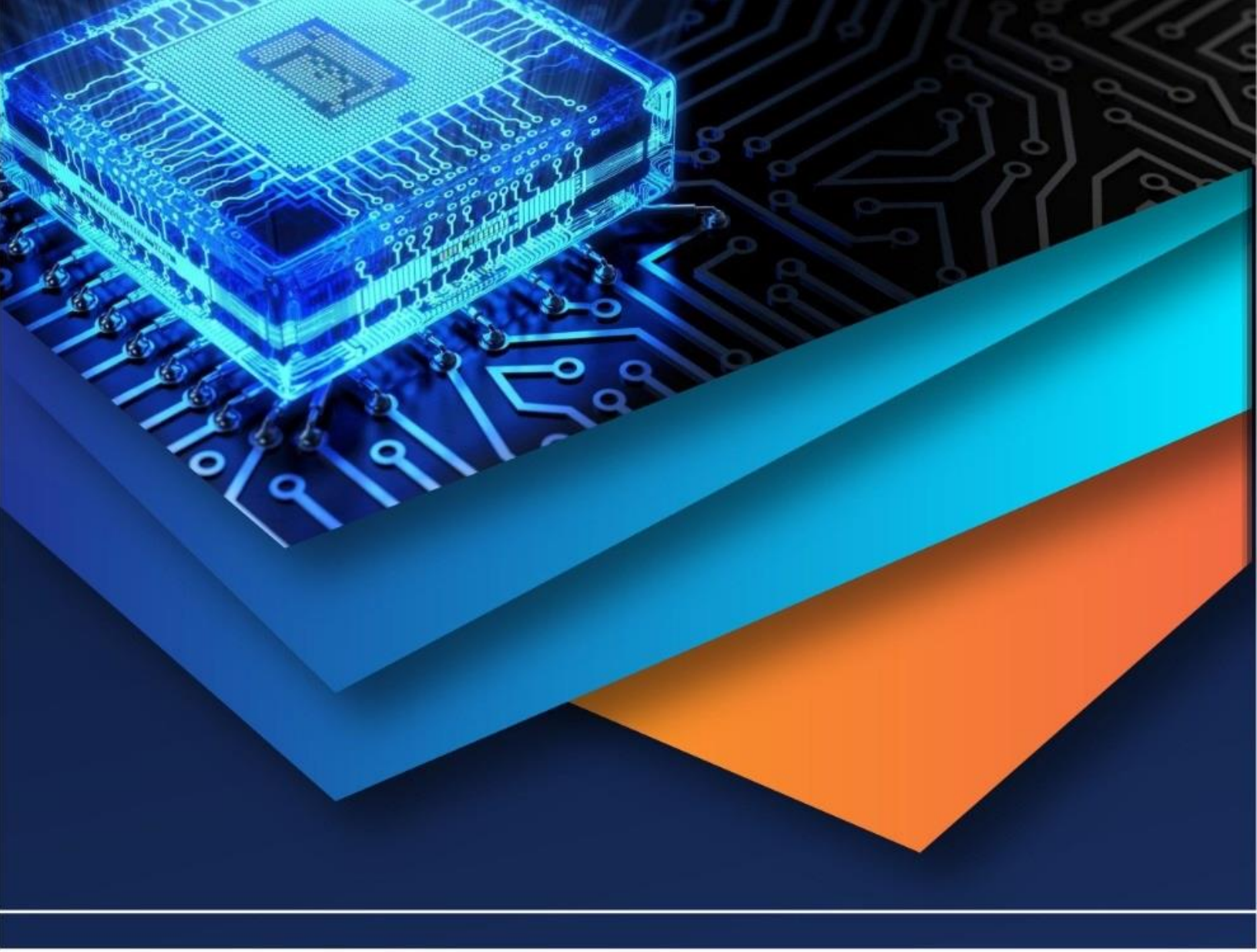
They can be used in tandem to both comprehend and forecast poverty; in fact, we can build a model that predicts poverty using both forms of data. Upon closer inspection, the machine learning techniques are straightforward classification techniques that performed admirably on our dataset. Our research will assist the government and other parties in understanding household poverty as well as predicting poverty by looking at satellite imagery. Our research demonstrates how unsupervised learning algorithms can be used on the survey data to make poverty analysis and then we can use supervised learning algorithms to predict poverty. Ultimately, our findings also imply that combining the two approaches may be a useful tool for forecasting and comprehending poverty.

VI. CONCLUSION

The integration of statistical analysis and machine learning techniques presents a promising avenue for poverty prediction in Nigeria. The choice of poverty measurement, along with the application of various supervised learning methods, plays a crucial role in the accuracy and reliability of predictions. The significant of this project is to look deep down into the predictive nature of the models and their accuracy. In other to get the result of what we are looking for we prepare our data that are suitable when it comes to prediction with our models. However we analyse models. We learn from the pre-training that the concept of anticipating poverty has produced better and more favourable outcomes for us. Many procedures were taken in order to accomplish the goal of the research, including data sourcing and preparation. Despite some problems with the data we got, we managed to make it appropriate for the study and employed various functions to ensure that it would work with all of the methodologies and algorithms that were employed. In conclusion, this research is a very good way to help critically understand poverty and how to predict areas with tendency of poverty. It can also help experts come up with ways to combat poverty. We collected data, analysed the data, and used k-means clustering to categorise the household by their level of poverty. We also tried to use machine learning models to predict the poverty levels.Future research should focus on refining these models and exploring hybrid approaches that combine multiple methodologies for improved insights into poverty dynamics. Moreover, addressing the challenges of data availability, quality, and representativeness remains essential for effective poverty prediction and policy formulation in Nigeria.

REFRENCES

[1] Jean, N., Burke, M., Xie, M., Davis, W.M., Lobell, D.B., Ermon, S.: Combining satellite imagery and machine learning to predict poverty. Science 353(6301), 790–794 (2016)
[2] Ngestrini, 2019 Ngestrini, R. (2019). Predicting poverty of a region from satellite imagery using cnns. Master's thesis, Departement of Information and Computing Science, Utrecht University. Unpublished, http://dspace.library.uu.nl/handle/1874/376648.
[3] Nigeria's Economy". Macro Poverty Outlook for Sub.Saharan Africa. World Bank
[4] Y. Zhang, J. Gao, and H. Zhou, "Breeds Classification with Deep Convolutional Neural Network," in ACM International Conference Proceeding Series, 2020, pp. 145–151
[5] C. Zhao, B. Ni, J. Zhang, Q. Zhao, W. Zhang, and Q. Tian, "Variational convolutional neural network pruning," in Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2019, vol. 2019.June, pp. 2775–2784.
[6] Jean et al., 2016 Jean, N., Burke, M., Xie, M., Davis, W. M., Lobell, D. B., and Ermon, (2016). Combining satellite imagery and machine learning to predict poverty.
[7] Poverty index in Nigeria https://www.statista.com/statistics /1121438/ poverty. head count.rate. in.nigeria.by.state/in the poverty headcount of the northern part of Nigeria
[8] S. Albawi, T. A. Mohammed, and S. Al.Zawi, "Understanding of a convolutional neural network," in Proceedings of 2017 International Conference on Engineering and Technology,
[9] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large.scale image recognition," in 3rd International Conference on Learning Representations, ICLR 2015 . Conference Track Proceedings, 2015. and poverty detection using satellite image
[10] Har.Peled, S., Roth, D., Zimak, D. (2003) "Constraint Classification for Multiclass Classification and Ranking." In: Becker, B., Thrun, S., Obermayer, K. (Eds) Advances in Neural Information Processing Systems 15: Proceedings of the 2002 Conference, MIT
[11] A. Shafiee et al., "ISAAC: A Convolutional Neural Network Accelerator with In.Situ Analog Arithmetic in Crossbars," Proc. . 2016 43rd Int. Symp. Comput. Archit. ISCA 2016,
[12] Abdi, H. and Valentin, D. (2007): Multiple correspondence analysis. Retrieved 27.03.2016 from: https://www.researchgate.net/profile/Dominique_Valentin/publication/239542271_Multiple_Correspondence_Analysis/ links/54a979900cf25 6bf8bb9
[13] Hersh, J., and Newhouse, D. Poverty from space: using high.resolution satellite imagery for estimating economic well.being, 2017. Breiman, L. (2001). Random forests. Machine learning, 45(1), 5.32.
[14] David, W. . H., & Stanley, L. (2000). Applied logistic regression. Willey.
[15] Agresti, A. (2003). Categorical Data Analysis (2th ed.). John Wiley & Sons Asselin, L..M. And Anh, V.T. (2008) : Multidimensional Poverty Measurement with Multiple Correspondence Analysis. In: Kakwani, N. And Silber, J. (eds.) Quantitative Approaches to Multidimensional Poverty Measurement. Palgrave Macmillan James, G; Witten, D., Hastie, T., Tibshirani, R. (2013): An Introduction to Statistical Learning with Applications in R. New York: Springer Greenacre, M.(2010) : Correspondence analysis and related methods [course]. Retrieved 26.03.2016 from http://statmath.wu.ac.at/courses/CAandRelMeth/
[16] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," Commun. ACM, vol. 60, no. 6, pp. 84–90, 2017
[17] K. H. Jin, M. T. McCann, E. Froustey, and M. Unser, "Deep Convolutional Neural Network for Inverse Problems in Imaging," ., vol. 26, no. 9, pp. 4509–4522, 2017.
[18] aches to Multidimensional Poverty Measurement. Palgrave MacmillanK.Means Clustering in Rdocumentation. Retrieved 28.04.2016 from https://stat.ethz.ch /R.manual /R.devel/library/stats/html/kmeans.html
[19] F. Harrell. (2017, January 15). Classification vs. Prediction. Retrieved April 01, 2018, from Statistical Thinking: http://www.fharrell.com/post/classification/
[20] Jean, Neal, Marshall Burke, Michael Xie, W. Matthew Davis, David B. Lobell, and Stefano Ermon. 2016. "Combining satellite imagery and machine learning to predict poverty." Science 353 (6301):
[21] Joshi, R. (2016, September 9). Accuracy, Precision, Recall & F1 Score: Interpretation of Performance Measures. Retrieved from http://blog.exsilio. com/all/accuracy. Precisionre call. f1.score.interpretation.of.performance.measures/
[22] Asselin, L..M. And Anh, V.T. (2008) : Multidimensional Poverty Measurement with Multiple Correspondence Analysis. In: Kakwani, N. And Silber, J. (eds.) Quantitative Appro



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY