University of Nairobi

School of Mathematics

ISSN: 2410-1397

Master Project in Statistics

Using Random Forest (RF) to Identify Key Determinants of Poverty in Kenya

Research Report in Mathematics, Number 40, 2020

Joel Omae Okiabera

November 2020

Submitted to the School of Mathematics in partial fulfilment for a degree in Master of Science in Social Statistics

Using Random Forest (RF) to Identify Key Determinants of Poverty in Kenya

Research Report in Mathematics, Number 40, 2020

Joel Omae Okiabera

School of Mathematics
College of Biological and Physical sciences
Chiromo, off Riverside Drive
30197-00100 Nairobi, Kenya

Master Thesis

Submitted to the School of Mathematics in partial fulfilment for a degree in Master of Science in Social Statistics

Submitted to: The Graduate School, University of Nairobi, Kenya

Abstract

Background

The World Bank (WB) has defined poor people as those who live on USD 1.9 or less per day. Most of the poor populations live in sub-Saharan Africa, Kenya being one of them. Although Kenya has made tremendous improvements in poverty reduction, it is unlikely to eradicate it by the year 2030. Its on this backdrop that it is important to understand the determinants of poverty in Kenya. This study intends to apply a random forest technique to KDHS 2014 dataset to explore poverty determinants in Kenya.

Methods

Random Forest is an algorithm used for classification and regression usually constructed from a set of classification and regression trees. The advantage of random forest is that they perform well in large feature setups and is effective in handling complex data. Analysis is carried out using the random forest package in R software, while imputation is carried out in missForest package. Inference is made on the model's classification accuracy, model diagnostics and interpretation.

Results

The results showed that variables had a very small percentage of missingness. There was reduction in error of the classifier as the trees in the classification was increased. The variable with the highest importance was the highest education level attained while marital status and sex of the household head, were variables with least importance. The classification was improved on the extreme ends of the wealth index as the Random Forests highly classified poorest and richest classes, while the middle, poorer and richer indices were not as accurately classified.

Conclusion

This study brings out the application of random forests to classify and predict the wealth index class of an individual in Kenya. The random forests are a significant improvement from classical regression techniques. The multiclass classification problem was effectively captured in this study. Regional residence and level of education details should be considered when interventions are being made for improvements of livelihoods in the country.

Master Thesis in Mathematics at the University of Nairobi, Kenya.
ISSN 2410-1397: Research Report in Mathematics

DISTRIBUTOR: School of Mathematics, University of Nairobi, Kenya

Declaration and Approval

I the undersigned declare that this dissertation is my original work and to the best of my knowledge, it has not been submitted in support of an award of a degree in any other university or institution of learning.

Signature Date

Joel Omae Okiabera
Reg No. I56/8288/2017

In my capacity as a supervisor of the candidate's dissertation, I certify that this dissertation has my approval for submission.

Signature Date

Dr Idah Orowe
School of Mathematics,
University of Nairobi,
Box 30197, 00100 Nairobi, Kenya.
E-mail: .....@uonbi.ac.ke

Dedication

I dedicate this work to God Almighty the creator and the source of all knowledge and wisdom. He has been the source of my strength and inspiration during the course of the research. I also dedicate this work to my wife, Nancy for the encouragement and moral support. To my children Griffins, Purity and Theophilus who have in one way or the other been affected by the time of this research. I also wish to dedicate this work to my mother, Esther, for her compelling concern for me to attain the highest level of education. God bless you all most abundantly.

Contents

Acknowledgments

First, my thanks to God for the enablement and provision of good health in the course of my research. I would also like to thank my employer through my boss, Mr. Rotich for allowing me time to do my studies as I performed my duties. I'm immensely grateful to my supervisor Dr. Idah Orowe, of the School of Mathematics, University of Nairobi, for her unwavering support and guidance in this work. I thank all those with whom I have had the pleasure to work with in this project, including some of my classmates who were very invaluable.

Joel Omae

Nairobi, 2020.

List of Acronyms

- RF- Random Forests model
- DHS - Demographic and health surveys
- PCA - Principal component analysis
- SES - Socio-economic status
- KDHS- Kenya Demographic Health Survey
- ROC- Receiver Operating Curve
- AUC- Area Under a Curve
- USAID- United States Agency for International Development

1 INTRODUCTION

1.1 Background to the study

The World Bank (WB) defines people who are poor as those who live on USD 1.9 or less per day. The WB also estimates that as of 2015, 10% of the world's population lived below the poverty line. In 2016, the United Nations formulated the Sustainable Development Goals (SDGs), and goal one focuses on the elimination of all forms of poverty by 2030 [Griggs et al., 2013]. This suggests that poverty is a significant problem, especially in the low-income countries and in particular Sub Saharan Africa. As of 2015, most of the global poor lived in sub-Saharan Africa. In the Sub-Saharan parts of Africa, 54% were poor in 1990. By 2015, the figure had reduced to 41%, which is a considerable improvement [Beegle and Christiaensen, 2019]. In Kenya, the poverty rate is even much lower. The percentage of those living below the international poverty line was 36.1% in 2015/2016.

However, the world bank in 2018 reported that although Kenya has made significant improvements in poverty reduction, it will not be able to hit the SDG target of eradicating poverty by 2030. There is, therefore, a considerable need to understand the significant poverty determinants in Kenya. The world bank, independent researchers, and other multinational organizations have conducted several studies to identify the determinants of poverty in Kenya. Such reviews help policymakers to set up guidelines and policy recommendations for poverty elimination in Kenya.

The measurement of poverty in developed countries relies on income data. However, this data is hardly available in low and middle-income earning nations (LMICs) and poverty measurement in LMICs depends on household expenditure [Sahn and Stifel, 2003]. Such household expenditure data can be obtained from household surveys like the Kenya Demographic Health Survey. Income and expenditure data are used because they are either seen as "ends" in themselves or because they are considered to be sufficiently well correlated with other welfare indicators such as literacy and nutritional status to suffice by themselves.

Studying poverty determinants in Kenya has mostly been done using classical regression methods such as logistic regression. In 2018, the world bank recommended using random forests as a more robust method of studying determinants of poverty. This study intends to apply a random forest technique to KDHS data to explore the determinants of poverty in Kenya.

1.2 Problem Statement

Poverty incidence in Kenya has been on a decline in the last ten years, but the world bank reported that it is unlikely to be eliminated by 2030, which is the SDG target. It is, therefore, essential to study the determinants of poverty to assist policymakers. Studying poverty determinants in Kenya has mostly been done using classical regression methods such as logistic regression. In 2018, the world bank recommended using random forests technique as a more robust method of studying determinants of poverty. There is, therefore, a need to apply a random forest technique to KDHS data to explore the determinants of poverty.

1.3 Study Objectives

1.3.1 Main Objective

The broad aim of this study is to identify key determinants of poverty in Kenya using the Random Forest technique and KDHS 2014 dataset.

1.3.2 Specific Objectives

1. To use random forests to impute missing values in the KDHS 2014 dataset.
2. To identify critical determinants of poverty using Random Forest from the KDHS 2014 dataset.
3. To fit a predictive model to quantify the contribution of each determinant to poverty in Kenya.

1.4 Significance of the study

Understanding the poverty determinants will help the government and other stakeholders to make more informed planning. This will accelerate the pace at which Kenya fights to eradicate poverty in parallel to the SDGs. Besides, the results of this study will help inform researchers who want to compare the performance of classical regression and RF technique in studying poverty determinants in Kenya.

2 LITERATURE REVIEW

2.1 Measuring poverty and its determinants

In most countries, poverty is a crucial policy concern and many, especially those with high levels of poverty, have strived to measure poverty with consumption expenditures. It has, however, been found that expenditures in consumption are unpredictable to measure. Due to this, poverty status is often determined with poverty proxies which are pointers that are highly associated with poverty and consumption. Ways of tracking of poverty over time through poverty proxies are positive and show that the methodology is capable of tracking poverty in time [Sumarto et al., 2007] [Christiaensen et al., 2012] [Mathiassen, 2013] [Sohnesen et al., 2014]. However, these evaluations also indicate that the method can be sensitive to the specific model applied.

According to [Vyas and Kumaranayake, 2006] measuring household expenditure also requires a considerable amount of resources, and the use of asset indices is a cheaper alternative when expenditure and income data is not available. For this reason he measured differences in the socioeconomic status of household in South Africa using the DHS data set. This analysis of the relationship between population and its welfare, whose data is present in the DHS surveys, will be impossible without such an asset index.

[Sahn and Stifel, 2003] used Demographic and Health Surveys (DHS) dataset to contrast poverty at various points in time among African countries. Their results showed a general decline in poverty during the previous ten years, owing largely to improved living standards in rural areas.

[Achia et al., 2010]) used an asset-index to measure poverty where it involved collecting information on ownership of various permanent assets that included; bicycle, car/track, radio, refrigerator, solar power, telephone and television. Also of interest in the measurement was housing characteristics that consisted of the material of floor, roof and toilet facilities of the dwelling house and access to necessary services that included electric power supply and source of water for drinking.

2.2 Handling missing data

Most researchers prefer to handle missing data by removing observations from the variables with missing entries then proceed to fitting statistical models and then use imputation methods. Most researchers use two standard techniques in handling missing

observations. These are removing observations where there are missing values in any variable to fit statistical models and imputation methods. Although these methods are widely used, there are various shortcomings. For instance, removing observations has been associated with producing biased parameters and estimates. In some cases, the researcher may opt to delete the entire variable with missing values. This should only be done where data is missing for more than 60 observations, but only if that variable is insignificant. Imputation methods can deal with some of these disadvantages associated with dropping observations and variables. These include use of mean, median and mode, linear regression and multiple imputations. Although the use of measures of tendency is fast, mean imputation, for instance, has been linked to reduced variance in the dataset. While linear regression(citation3) has proved theoretically to provide good estimates of missing values, it relies on the assumption that the variables used in the regression equation are linearly related which may not hold in most cases. Moreover, the replaced values are as a result of prediction by other variables; hence they tend to fit 'too well'. This deflates the standard error and thus biased estimates. Multiple imputations have, however, been a breakthrough in the field of missing data in the recent past. It's the use of the Markov Chain Monte Carlo (MCMC) simulation, and pooling of analysis results has been shown to improve efficiency and accuracy. It has been shown to provide unbiased estimates which are more valid than other ad hoc methods in estimation of missing values. It's easy to use due to the availability of various algorithms already developed in various standard statistical software, and it preserves the sample size through the use of all the available data. Moreover, the results are readily interpreted and preserve statistical power(citation2). However, this method is limited to analytical models without interactions and where the proportion of missing data is not too large. These shortcomings could be checked mainly through the use of machine learning techniques developed recently. These include the use of K Nearest Neighbours, XGBoost and Random Forests algorithms.

2.3 Random Forest

Random forests are a non-parametric supervised learning ensemble method used for regression and classification. It is also known as the random decision forest. The implementation is done through building machine learning algorithms than assist in prediction. To achieve the best outcome, the RF develops a forest of random unassociated decision trees to attain the best possible solution [Wambua, 2019]

Random Forests algorithm as machine learning techniques has been proven to conduct classification accurately. Compared to other methods. RF is more attractive due to the fact that it hosts properties for handling missing value for mixed data, and adapts easily to interactions and non – linear settings. Due to its extensive use in most research projects, different algorithms have been developing from the original package by

[Tang and Ishwaran, 2017]. This method would inform the best classification of the determinant of poverty in Kenya, that can also be used in predictive modelling.

The RF technique is part of Machine Learning techniques and is used for to predict a wide area of research fields. In [Verikas et al., 2011] for example,RF was used to predict long disordered regions in protein sequences, classifying farming practices using satellite imagery, environment conditions measurement in space, handwritten digits recognition, and video objects segmentation.

[Varian, 2014], a prediction economist, identifies Machine Learning as having an advantage of being better to predict, but is deficient in estimations and testing of hypothesis. The main reason why Machine Learning is better at prediction is the ability to handle non-linearities, but at the same time working under-performing in predictions involving linear variables.

Though Random Forests application is scant and new in predicting poverty, in Indonesia, [Otok and Seftiana, 2014]found out that Random Forests are precise in identification of households that are poor and are eligible for government assistance packages, while it was successfully applied in Mauritius to predict poverty. [Thoplan, 2014].

In his study, , [Sekhampu, 2017], sought to find key determinants of poverty. In the study, data at the household-level was used to analyze factors associated with household poverty in Bophelong, a town in South Africa. He used Logistic regression in the estimation based on data with the response as socioeconomic status, whether they were poor or non-poor, as the response variable and a set of predictor variables. The results indicated that age, household size, and the household head's employment status of were significant in telling poverty. The chances of being poor was reduced by age and employment status of the head of the household, and the size of the household was associated with an increased chance of poverty. The most important factor associated with poverty was found to be the employment status of the household head.

[Achia et al., 2010] used data from the DHS to examine the factors determining poverty in Kenya. They used Principal Component Analysis(PCA) to develop asset indices which yielded the Social Economic Status (SES) of each household. They used binary Logistic regression to do the estimationwith the SES (that is poor and non-poor) as the response and demographic variables as the predictors. From the results, it was clear that DHS data was useful in determining factors associated with poverty.

[Dartanto and Nurkholis, 2013] used the 'spell' method to identify poverty determinants and applied an ordered logit model to evaluate the determinants of Poverty in Indonesia. They categorized households as transient poor (–), poor, transient poor (+) or non-poor. They found out that 28% of households are classified as very poor, that is, remaining poor

in two periods while 7% of non-poor households are vulnerable to being transient poor (-). Their results found out that attaining education, household members size, employment status, physical assets, health shocks, credit program, electricity access, fluctuations in employment dynamics, status of employment and the number of household members were significantly associated with poverty.

In their paper, [Akerele et al., 2012] explored the poverty situations among urban households in the Nigerian state of Ekiti, with a focus on socioeconomic characteristics of their households and their associated effect on poverty.

[Lekobane and Seleka, 2017] used the 2002/2003 Household Income and Expenditure Survey and the 2009/2010 Botswana Core Welfare Indicator Survey data, to study household determinants of poverty and welfare using regression in Botswana. They established that level of education and status of employment of the head of household are the key determinants of poverty. Further, they found out that living in rural areas increases the chances of one being poor and had negative relation to welfare.

[Habyarimana et al., 2017], in their paper, used PCA to first create asset indices for each household. They later used quantile regression model to model Poverty determinants in Rwandan households. They considered the features of homes, and the household heads. The data for this study was from the Rwanda Demographic and Health Survey (2010). From the findings, individual's gender and age of the head of the household, household size, the level of education, place which one resides were significant predictors of poverty of household in Rwanda. The model was invaluable to these researchers for them to evaluate the effect of covariates on quantiles of the asset index. This helped in getting a clear picture of the association between the index and covariates.

3 METHODOLOGY

3.1 Data

The data to analyze the critical determinants of poverty is sourced from the 2014 Demographic and Health Surveys (DHS) in Kenya. The data collection is done under sponsorship by the USAID with partnerships with other Kenyan research agencies. The data has key indicators such as wealth index, and population characteristics of the samples of the population chosen for the survey. The DHS program is carried out in many nations providing key updates on the population metrics, health status both at individual and regional levels to provide reliable estimates of the whole national characteristics of citizens. Authorisation to use this data was obtained by sending a request to DHS online portal, and granted under the project name using random forest (rf) to identify poverty determinants in the country. Kenyan agencies in the partnership provide the personell to be involved in conducting the survey. The samples are taken from the country's population and housing census estimates usually created before the onset of a national census. Sampling is done in a two-stage sampling approach where the Enumeration Areas (EAs) are the primary sampling units, while the second sampling unit selects the household. The household represents an individual or a group of individuals, who stay together and share the same source of food. Both rural and urban populations will be covered in the study. The information collected in the survey relates to access to public services, asset ownership and housing characteristics.

3.1.1 Study Variables

Key determinants of povery levels are presented in table 1 below. The outcome variable is wealth index, categorized in five levels ranging from poorest, poorer, middle, richer, richest.

Table 1. Description of Key Variables

Determinants	Definition
Outcome Variable	
Wealth Index	The computed wealth index of an individual ranked from poorest, poorer, middle, richer, richest.
Independent Variables	
Region	Various administrative regions in the country, from Coast to North Eastern, Rift Valley, Eastern to Western, and Central to Nairobi.
Type of residence	Categorized as rural or urban
Education Level	Individuals Level of education, ranging from none, to primary, secondary to tertiary
Sex of household head	Categorized as male or female
Marital Status	Individual's status of marriage, categorized as married, divorced, never married, widowed
Number of household members	Continous, listing the number of members in the household.
Age of household head	Age in years of the head of the household .

3.1.2 Data Structure

The KDHS data 2014, has the wealth index of an individual, coded for the categories as outlined above, while the determinant variables are both categorical and continuous.

3.2 Classification and Regression Trees

Classification and Regression Trees (CART) are a more recently adopted name of Decision Trees. A decision tree consist of a tree-like structure with the top node termed as the root of the tree, while the other parts of the tree are recursively partitioned at a set of decision nodes until an optimal decision node is reached. The decision tree works by making a decision to classify an outcome of interest at a certain level of the outcome for a certain node, if that decision is not reached, then it is considered as information gained, and then the next node is considered for a classification decision, and if such a decision is optimal, then the classification is terminated. Figure 1 illustrates how decisions are reached in

a decision tree,if one has no education the tree classifies them as poor, while when one has secondary education, another node is created to determine the classification after putting in more information.

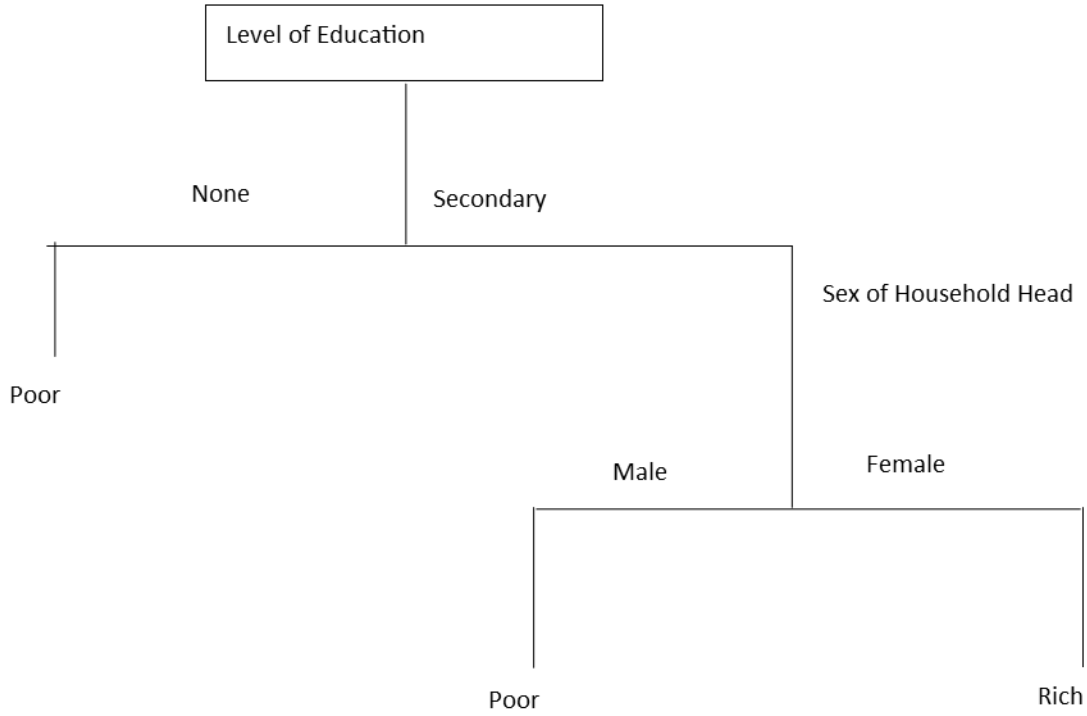

Figure 1. Decision Tree

In a data set perspective, every tree provides a repeated partitioning of the subset of the data used to train the algorithm. Once a subset that is less heterogenous is reached, a classification is made. An if then condition is used to partition the more heterogenous subsets into further subsets, to ensure homogeneity is reached. The predictor used in the final decision tree is one that has less heterogeneity. Heterogeneity, or homogeneity in the whole classification problem is measured by Entropy, denoted by E, and is given as the negative logarithm of the proportion of a certain level of the outcome variable, compared against another level of the outcome variable, given as;

$$E = -p\log p - (1-p)\log(1-p) \tag{1}$$

Where p represents the proportion of an event of interest, say in a binary outcome setup, being rich, while $1-p$ is the proportion of being rich.

Consider two desired outcomes A and B dependent on two predictor variables x and y, and partitions are done as shown in figure 2, the first partition is done at $x = 3$ while

the second partition is done at $y = 6$, and the final and last partition done at $y = -4$. A decision to classify A is reached after the first two partitions of x and y while the decision to classify B is reached after the third partitioning. If there is no homogeneity, then it means the entropy levels are still high, and partitioning should continue using the information available, until there are no further ways to classify, then the algorithm stops.

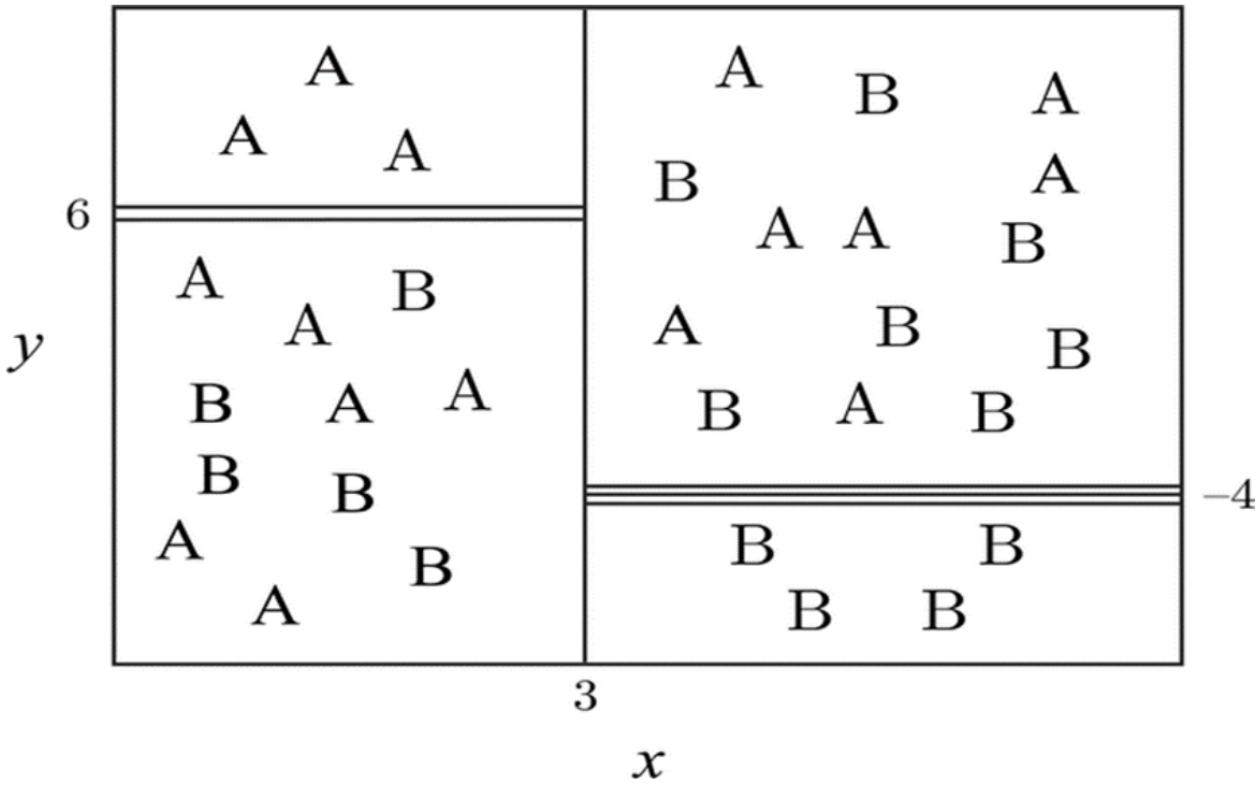

Figure 2. Partitioning Classes

The final set is then assigned a class, in our case poor or rich, or in the presence of more levels of the outcome variable, the assignment is done on all the levels. This assignment of the class is based on a majority vote; the proportion which has majority of observations in that set. In case of a new observation, or data point, one requires to determine the set in which the observation lies and its class [Deloncle et al., 2007].

3.2.1 Random Forests

A Random Forest is an algorithm used for classification and regression usually constructed from a set of Classification and Regression Trees as explained above, each tree is created using a sample chosen at random from the training data set, and used at each partition within a tree, a sample at random of the predictors is also used [Deloncle et al., 2007]. One limitation of the decision tree is the instability in prediction when a new data set is introduced, which usually results in overfitting, this instability is solved by using a

combination of decision trees such that each decision tree is constructed from a random independent sample. The advantages of Random Forests is that they perform well in large feature setups, and is effective for handling complex data structures for which no visible relationship exists and it also allows for selection of important variables through variable importance.

The tree predictors; such that each of the trees depends on a randomly sampled vector sampled independently and identically therefore makes the trees have the same distribution. A random forest classifier is therefore a classifier with a collection of various tree-like classifiers $h(x,\theta_k), k = 1,\cdots B$ where the (θ_k) are independent and identically distributed. Each one of these trees votes on the popular class, and the majority votes are used to classify. In the case of regression, the tree predictor $h(x,\theta_k)$ takes on numerical values.

The trees therefore generated are identically distributed and independent; the expectation of each tree is similar with any other tree. To improve the algorithm performance we aim to reduce the variance. The B identically independent random variables each with a variance σ^2, has variance $\frac{1}{B}\sigma^2$. Assuming the variables are identical but not independent, then the trees are correlated and the pairwise correlation between them affects the variance and is given as;

$$\rho\sigma^2 + \frac{1-\rho}{B}\sigma^2 \tag{2}$$

As B increases the last term in the equation vanishes, and therefore the variance reduction is left to be a function of the correlation between pairs of trees. This reduction in correlation is achieved by considering random samples of the features used in any tree. Therefore although a subset of the observations is considered for each tree, a subset of the covariates is also considered, the general rule is $m \leq p$ where m denotes the number of independent variables chosen, and p is the total number of variables.

3.2.2 Random Forest Algorithm

This section outlines the steps for construction of trees for the random forests. For each $b = 1,\cdots,B$, a random sample, n is drawn from the training data set. This sample is selected with replacement, usually two thirds of the training data. Next step involves taking a random sample of the set of predictors without replacement consisting of all the possible choices of the predictors. Using these two samples the first random forest tree T_b is developed as in the decision trees above using the randomly selected subset of the

training data set, the steps for classification trees are repeated until terminal node size is reached.

The steps are repeated for a maximum number of trees, and the output of the ensemble of decision trees made are T_{b1}^{B}. In order to make a prediction, incase of a classification, the prediction is given as follows;

Let $\hat{C_b}(x)$ be the class predicted by the bth tree; then

$$\hat{C_{rf}^{B}}(x) = majorityvote\hat{C_b}(x)_1^B \tag{3}$$

While regression is given as follows;

$$\hat{f_{rf}^{B}}(x) = \frac{1}{B}\sum_{b=1}^{B} T_b(x) \tag{4}$$

The equations above represents the class choice and the regression choice as an average, a majority vote means the class mostly classified is voted as the final classification from a set of decision trees in the random forest. In the random forest tree branching problem, the Gini Index is used to determine the branching of the decision trees, right from the root nodes, to the daughter nodes. It is given by;

$$GI = 1 - \sum_{c=1}^{C} (pi)^2 \tag{5}$$

Where GI is the Gini Index, pi is the frequency of the class observed in the data set, and C is the number of classes under the classification problem. This index on each tree uses the class information to determine the most likely tree to occur.

Data Imputation

To use Random Forest for multiple imputations, the following steps are followed; more specifically to impute Y using observed values $X_1, \cdots, X_p$, random forests is applied to the observed values (y^{obs}, x^obs) using k boostrap samples. When a subject is missing Y with

some set of predictors $X_1, \cdots, X_p$, take the values observed of Y in the terminal nodes of the k trees used in the creation of the random forest. To impute the missing value, one observed value of Y from these terminal node values is used to impute the missing Y. Repeating this process multiple times creates multiple imputations for the missing values.

3.2.3 Variable Importance

The variable importance is a measure of how the variance is reduced, or the impurities reduction on each decision tree brought about by the gain in information or the Gini coefficient index. Variable importance is calculated through a Mean Decrease Impurity which sums up the Gini Index decrease of the variables and averages to obtain a list of important variables. It is given as

$$V_{imp}(x_i) = \frac{1}{n_{trees}}[1 - \sum_{j=1}^{n_{tree}} GI(i)^j] \tag{6}$$

Model Performance

To fit the model, the data is separated in a test and training sets. The train data set is a subset representing about 75% of the original data set, while the test data contains the remaining amount of the original data set. Once the model is fitted on the train data set, then the prediction is done on the test data set, with Wealth Index as the response variable.

The model performance is analyzed using the Receiver Operating Characteristic (ROC) curve, which plots the diagnostic ability of a classifier. The Area Under a Curve (AUC) is also used to assess the models classification accuracy. AUC also distinguishes between the classes, and the higher the value of the AUC the better the ability to differentiate between the classes.

A confusion matrix is also used to assess the model's performance. It is a contingency table representing the actual classified values as true in the model compared to the test data, and the values classified as actually negative when they are negative, while the other cells of the matrix represent the misclassified values. More specifically, the matrix represents True Positive values classified as positive, and the True Negative values on the main diagonal, the off diagonal values represent False Negatives and False Positives. A 2 by 2 contingency table is shown below;

Table 2. Confusion Matrix

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

This matrix is used to calculate the major indicators of the model performance as shown below;

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{7}$$

while the error rate is given as

$$Errorrate = \frac{FP+FN}{TP+TN+FP+FN} \tag{8}$$

Specificity referes to the proportion of actual negative classes classified as negative. These are the number of poor people actually classified as poor from the model.

$$Specificity = \frac{TN}{TN+FP} \tag{9}$$

Sensitivity refers to the number of actual positives classified as positive. This is the proportion of people who are rich and actually classified as rich by the model.

$$Sensitivity = \frac{TP}{TP+FN} \tag{10}$$

The problem at hand is a multi class classification problem, and instead of two classes in a binary classification, we have 5 classes. Therefore, we compute a macro-average to compute the precision for each class, and then assume the average. This average works well in classes whose sizes of observations are almost similar. If they are not similar then a micro-average is computed, this average aggregates the contributions of each class together and computes an average metric. For imbalanced classes, this average is more reliable.

Analysis is carried out using the Random Forest package in R software, while imputation is carried out in missForest package. Inference is made on the model's classification accuracy, model diagnostics and interpretation.

4 RESULTS

4.1 Introduction

This section deals with reporting the descriptives of the data for the various variables under consideration, presenting the findings of the random forest classification model, model diagnostics and results interpretation.

4.2 Descriptive Statistics

4.2.1 Categorical Variables

Table 3. Summary Statistics of Key Variables

VARIABLE	FREQUENCY	PERCENTAGE
		REGION
Rift valley	10,534	29%
Eastern	6,261	17%
Nyanza	4,801	13%
Coast	4,476	12%
Central	4,041	11%
Western	3,220	9%
(Other)	3,097	9%
		TYPE OF RESIDENCE
Urban	13,914	38%
Rural	22,516	62%

Table 4. Summary Statistics of Key Variables contd'

EDUCATION LEVEL		
No education, preschool	7,631	21%
Primary	16,585	46%
Secondary	8,233	23%
Higher	3,968	11%
Don't know	13	0%
SEX OF HHH		
Male	23,860	65%
Female	12,570	35%
WEALTH INDEX		
Poorest	9,114	25%
Poorer	6,994	19%
Middle	6,849	19%
Richer	7,267	20%
Richest	6,206	17%
MARITAL STATUS		
Married/living together	25,445	70%
Divorced/separated	2,684	7%
Widowed	5,013	14%
Never married and never lived together	3,288	9%
Majority (62%, n=22516) of the respondents were from the rural areas.		
The total number of respondents were 36,430		

The total number of respondents were 36,340 out of which a majority of them, 29%, n=10,534 came from the Rift Valley region, 17% from Eastern and 13% from Nyanza. From the coast, there were 12%, with central with 11% and 9% from Western. The rest of Kenya constituted 9% of the total number of respondents. It was also found that rural dwellers had a lion's share of the respondents, taking 62% of the total number of respondents. On the level of education, respondents with primary level of education were the majority with 46% of the total number of respondents, while those with no education or preschool were 21% of the total number of respondents. Secondary school level had 23% of the respondents while 11% said they had higher education level. The survey had a vast majority of the respondents being male as the heads of the household at 65% of the total number of those interviewed. 44% of the respondents said they were poor, 19% were in the middle while 37% responded that they were rich. On marital status, majority of the respondents, 70%, were either married or living together while only 7% were either divorced or separated. The proportion of the widowed among the respondents was 14% and 9% of the respondents either never married nor lived together 5.

4.2.2 Continuous Variables

Table 5. Description of Continuous Variables

	Min/Max	Mean (SD)
Number of household members	1 and 23	4.22 (2.51)
Age of head of household	14 and 95	44.09 (16.09)

The minimum number of household members was one while the maximum was 23. The household in this case meaning members who eat from the same house. The average number of members in each household was 4.22 with a standard deviation of 2.51. The lowest age of the household head was 14 years while the highest was 95 years, with a mean of 44.09 years and a standard deviation of 16.09.

4.3 Imputation

Imputation is done to the missing observations, the missingness levels are shown in 6. Only 3 variables have missing values, with very small percentages of missingness.

Table 6. Level of Missingness per Feature

Level of Missingness	
Feature	Percent
Number of household members	0
Region	0
Type of place of residence	0
Highest educational level attained	0.003046939
Sex of head of household	0
Age of head of household	0.000164699
Wealth index	0
Marital status	0.004199835

4.4 Random Forest Classifiers

The number of trees are assessed for the optimal number that reduces the amount of error for each classifier. 3 shows the reduction in error for the classifiers as the trees input in the classification is increased. The optimal number of trees that reduces the errors computed for the whole model is 1502.

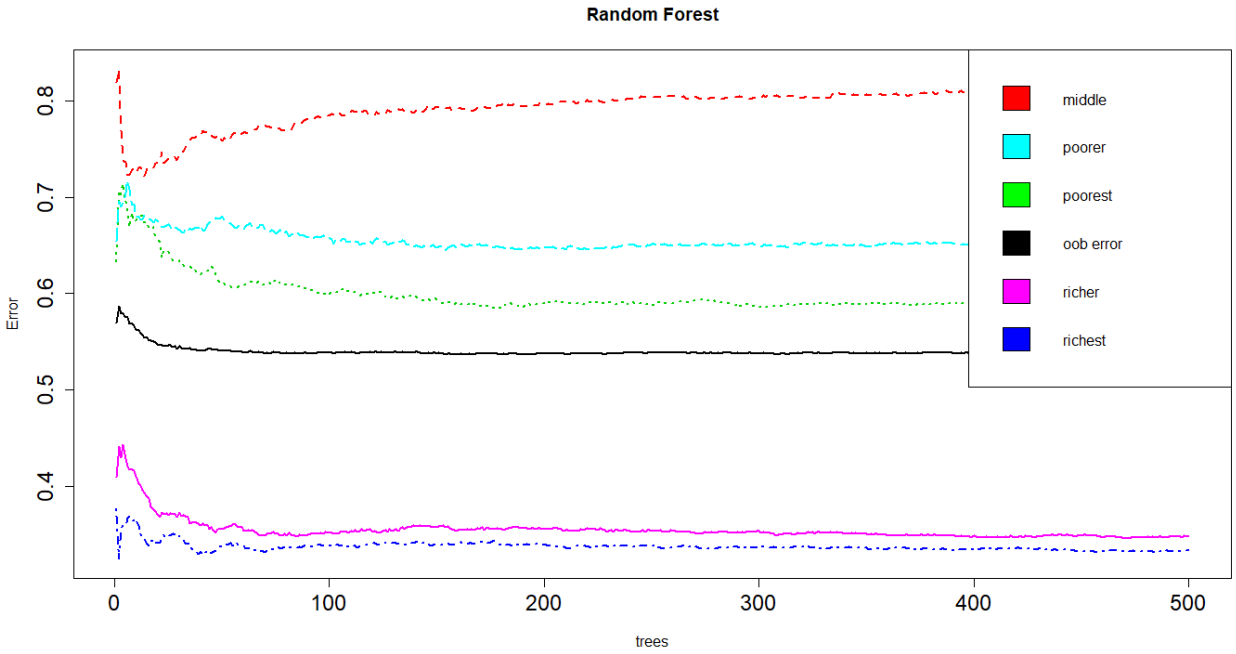

Figure 3. Number of Trees by Class for Error Reduction

4.5 Variable Importance

The results in 4 indicate the variable highest education level attained was the variable with the highest importance with a score of 2,728, followed by type of residence at 1,546. Region followed closely at 1,507 with age of the household head and number of household members at 1,441 and 791 respectively. Marital status and sex of the household head were, the variables with least importance at 399 and 205 in that order.

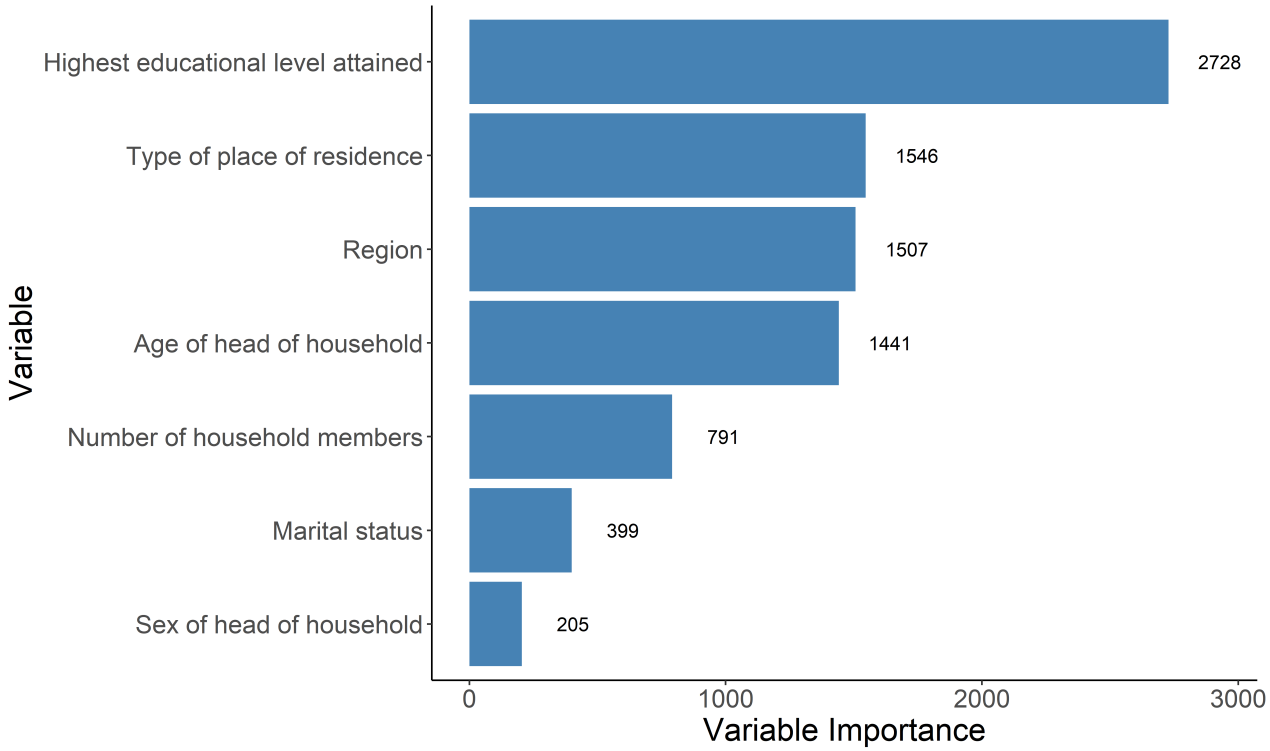

Figure 4. Variables by Importance

4.6 Model Accuracy

Accuracy is a measure of the correctly predicted observations over the total number of the predictions. It helps in determining the extent to which the model classifies the positive classes as positive and the negative classes as negative. Table 7 shows an accuracy of the classification model as 47.73%. One limitation of the accuracy in the multi class classification problem is the lack of clarity on how the classification breaks down across the classes, hence not a good measure of accessing classification accuracy. Therefore a confusion matrix is important to enable an understanding of the process of classification for each class.

Table 7. Model Accuracy results

Overall Statistics	
Accuracy	0.4773
95% C.I	(0.467,0.4876)
No Information Rate	0.2501
P-Value	2.20E-16
Kappa	0.3434
Mcnemar's Test P-Value	2.20E-16

This matrix helps in showing how the classification has been done for each class. Table 8 shows the confusion matrix for all classes. This classification has the correctly classified observations in the test data set in the main diagonal, while the off-diagonal elements represent misclassified observations for the classes. Poorest and Richest classes got the highest number of classifications (1532 and 1057) in the classification problem, while middle placed wealt indicator was least classified correctly as the count of observations correctly predicted as middle (377) was lowest among the classes, poorer and richer classes had (712 and 669) respectively. The model therefore classified the extreme ends of the wealth index more correctly as opposed to the other classes; poorer, middle and richer.

Table 8. Confusion Matrix by Class

	Reference				
Predictions	middle	poorer	poorest	richer	richest
middle	377	273	121	240	55
poorer	527	712	448	253	42
poorest	253	405	1532	171	54
richer	397	235	116	669	344
richest	158	123	61	484	1057

To determine the performance of the classification per class, an overall performance by class table was drawn, showing sensitivity and specificity; positive classes for each class

classified correctly against other classes, and negative classes classified correctly against each class respectively for the five classes. Once again the per class performance ranked richest and poorest as with high sensitivity (68% and 67%) and specificity (87% and 89%) respectively. Other classes performed poorly for sensitivity, while their specificity was high (middle-90%, poorer-87%, richer-85%), it means that their classification as not belonging to that class was high. It is important to note that specificity for classes with a low positive classification experienced classification negatively such that they were not classified as being middle, poorer, or richer. This is shown in 9.

Table 9. Classification Results by Class

	middle	poorer	poorest	richer	richest
Sensitivity	0.2202	0.4073	0.6725	0.3682	0.6811
Specificity	0.9068	0.8274	0.8707	0.8502	0.8907
Pos.Pred.Value	0.3537	0.3592	0.6344	0.3799	0.5613
Neg.Pred.Value	0.834	0.8546	0.8885	0.8437	0.9315
Precision	0.3537	0.3592	0.6344	0.3799	0.5613
Recall	0.2202	0.4073	0.6725	0.3682	0.6811
F1	0.2714	0.3818	0.6529	0.374	0.6154
Prevalence	0.188	0.1919	0.2501	0.1995	0.1704
Detection.Rate	0.0414	0.0782	0.1682	0.0735	0.1161
Detection.Prevalence	0.1171	0.2176	0.2652	0.1934	0.2068
Balanced.Accuracy	0.5635	0.6174	0.7716	0.6092	0.7859

4.7 ROC/AUC Curves

The Receiver Operating Characteristic/Area Under a Curve are plots for the classes that maximize the true positive rate and minimizes the false positive rate. The true positive rate is measured by sensitivity, while the false positive rate is the difference of the specificity from 1, which represents the proportion of false positives. As shown in figure 5 Richest and Poorest classes show a tendency towards one, which implies that their classification was highly predicted as positive among all the classes. This shows that random forest was good in identifying the correct classes among the Richest and Poorest classes. The Micro and Macro Average plots also show an average AUC for all the classes. These

averages are almost similar, although the Micro Average outperforms the Macro Average, indicating its suitability when there is a class imbalance as experienced in this study; the poorest class has more observations (9114), although slight in the data used, against middle (6849), poorer (6994) and richer (7267). Poorer and Richer classes have a similar curve, showing that correct classification between the two is relatively similar, while middle has the lowest amount of correctly classified observations as positive.

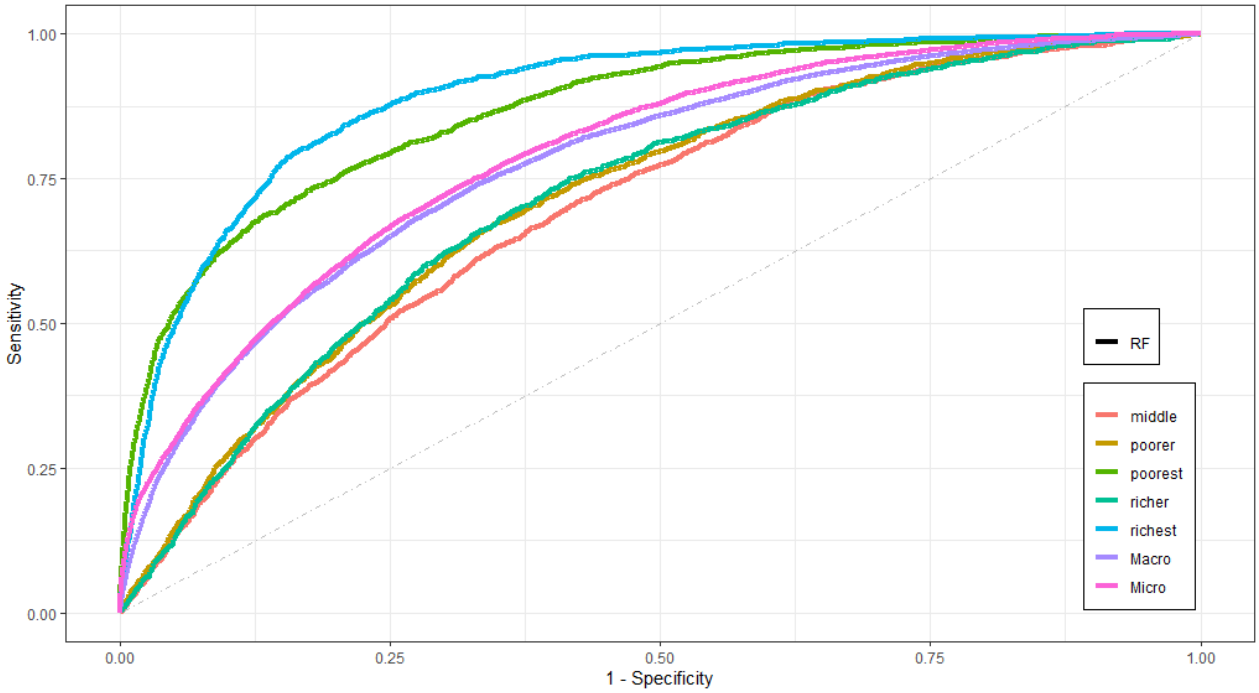

Figure 5. Receiver Operating /Are Under Curve

5 DISCUSSION AND CONCLUSION

5.1 Discussion

The study aim was to utilize random forests to classify the wealth index of individuals in the country using the KDHS data set 2014. Random forests was used to impute the missing observations, it was also used to determine the important variables that help in telling the particular class a person belongs to in terms of wealth. Finally it was used as a classification tool, classifying over and above the binary classification problem usually employed in logistic regression, random forests enabled a multi-class classification on the 5 levels of wealth index of individual household in the country. The accuracy and the diagnostics of the model were also determined.

The level of missingness in the variables was low. Highest level of education of an individual, followed by the type of place of residence and the region were more important in classifying an individual as poor or rich. Attending school to secondary or tertiary level improves the chances of one falling on the upper side wealth index, meaning middle, richer and richest. Level of education also presented a high margin compared to the next most important variable in determining classification of a household on one of the wealth index classes. The type of the place of residence, categorized as rural or urban also influenced the classification of an individual to the wealth index classification. One could argue that it is closely tied to the region one hails from, which was the next important variable on classification. However, it remains unclear as to what region characteristics influenced the classification as above, and it beyond the scope of this paper. A stand alone determinant of wealth index classification was also the age of the household head, a telling of the change in age that causes a change in classification.

The classes with the highest level of accuracy; representing classes that were more accurately classified were poorest and richest classes. Random forests were able to better classify the extreme ends of the wealth index. The middle classes consisting of middle, poorer, and richer classes were not as accurately classified. This is evidenced in the ROC/AUC curves. It is evident that the random forest model performed better in classifying the poorest households, which could be attributed to the variables, indicating an isolation in the characteristics of individuals present in both extremes of the classes. This means that the attributes identifying the two classes are so independent in their nature such that there are clear stark differences say for the level of education one individual with absolutely no education might be identified as poor while the richest counterpart would have gone through all possible levels of the education system in the country. The

poor performance in classifying the middle level classes of the wealth index might be attributed to the similarities in the attributes of the individuals belonging to these categories, as it were.

5.2 Conclusion

This study brings out the application of random forests to classify and predict the wealth index class of an individual in Kenya. The random forests are a significant improvement from classical regression techniques, although not at an exhaustive level, use of both for the suitable problem is appropriate. The multi-class classification problem was effectively captured in this study, regional, residence and level of education details should be taken into consideration when interventions are being considered for the improvement of livelihoods in the country. Any intervention should monitor to ensure the there is a significant reduction in the gap, as the ability of the model to capture the extreme classes very well tells there is an actual gap, that needs said monitoring.

Future work should consider the spatial effect on the classification problem to bring out the extent of the region effect, and consider using counties, and also how to effectively classify the middle level indices on the wealth index in the country.

Bibliography

[Achia et al., 2010] Achia, T., A, W., and N, K. (2010). A logistic regression model to identify key determinants of poverty using demographic and health survey data. *European Journal of Social Sciences*, 13(1).

[Akerele et al., 2012] Akerele, D., Momoh, S., Adewuyi, S. A., Phillip, B. B., and Ashaolu, O. F. (2012). Socioeconomic determinants of poverty among urban households in south-west nigeria. *International Journal of*.

[Beegle and Christiaensen, 2019] Beegle, K. and Christiaensen, L. (2019). Accelerating poverty reduction in africa.

[Christiaensen et al., 2012] Christiaensen, L., Lanjouw, P., Luoto, J., and Stifel, D. (2012). Small area estimation-based prediction methods to track poverty: validation and applications. *The Journal of Economic Inequality*.

[Dartanto and Nurkholis, 2013] Dartanto, T. and Nurkholis (2013). The determinants of poverty dynamics in indonesia: evidence from panel data. *Bulletin of Indonesian Economic Studies*.

[Deloncle et al., 2007] Deloncle, A., Berk, R., D'Andrea, F., and Ghil, A. M. (2007). Weather regime prediction using statistical learning. *Journal of Atmospheric Science*.

[Griggs et al., 2013] Griggs, D., Stafford-Smith, M., Rockstrom, G. O., Ohman, J. M. C., Shyamsundar, P., and Noble, I. (2013). Sustainable development goals for people and planet. *Nature*, 495(7441).

[Habyarimana et al., 2017] Habyarimana, F., Zewotir, T., and Ramroop, S. (2017). Structured additive quantile regression for assessing the determinants of childhood anemia in rwanda. *International Journal of Environmental Research and Public Health*.

[Lekobane and Seleka, 2017] Lekobane, K. R. and Seleka, T. B. (2017). Determinants of household welfare and poverty in botswana. *Journal of Poverty*.

[Mathiassen, 2013] Mathiassen, A. (2013). Testing prediction performance of poverty models: Empirical evidence from uganda. *Review of Income and Wealth*.

[Otok and Seftiana, 2014] Otok, B. and Seftiana, D. (2014). The classification of poor households in jombang with random forest classification and regression trees (rf-cart) approach. *International Journal of Science and Research (IJSR)*.

[Sahn and Stifel, 2003] Sahn, D. E. and Stifel, D. (2003). Exploring alternative measures of welfare in the absence of expenditure data. review of income and wealth. *Review of income and wealth*, 49(4).

[Sekhampu, 2017] Sekhampu, T. (2017). Association of food security and household demographics in a south african township. *Journal of Social Sciences and Humanity*.

[Sohnesen et al., 2014] Sohnesen, T., Christiaensen, L., and Carletto, C. (2014). Tracking poverty via consumption proxies. *Review of Income and Wealth*.

[Sumarto et al., 2007] Sumarto, S., Suryadarma, D., and Suryahadi, A. (2007). Predicting consumption poverty using non-consumption indicators: Experiments using indonesian data. *Social Indicators Research*.

[Tang and Ishwaran, 2017] Tang, F. and Ishwaran, H. (2017). Random forest missing data algorithms. statistical analysis and data mining. *The ASA Data Science Journal*, 10(6).

[Thoplan, 2014] Thoplan, R. (2014). Random forests for poverty classification", international journal of sciences: Basic and applied research (ijsbar). *International Journal of Sciences: Basic and Applied Research (IJSBAR)*.

[Varian, 2014] Varian, H. (2014). Big data: New tricks for econometrics. *Journal of Economic Perspectives*.

[Verikas et al., 2011] Verikas, A., Guzaitis, J., and Gelzinis, A. (2011). A general framework for designing a fuzzy rule-based classifier. *Knowledge and Information Systems*.

[Vyas and Kumaranayake, 2006] Vyas, S. and Kumaranayake, L. (2006). Constructing socioeconomic status indices: how to use principal components analysis. *The ASA Data Science Journal*, 21(6).

[Wambua, 2019] Wambua, S. (2019). Random forests application in missing data and predictive modelling for hierarchical routine clinical data: A case study of childhood pneumonia in kenya.