



ISSN: 2410-1397

Master Project in Biometry

Comparison of Elastic Net and Random Forest in identifying risk factors of stunting in children under five years of age in Kenya

Research Report in Mathematics, Number 51, 2020

Rachael Mburu

October 2020



**Comparison of Elastic Net and Random Forest in
identifying risk factors of stunting in children
under five years of age in Kenya**

Research Report in Mathematics, Number 51, 2020

Rachael Mburu

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 Biometry

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

Abstract

Background: Children with a Height-for-Age (HAZ) below -2 Standards Deviations based on the World Health Organization (WHO) child growth standards median are said to be stunted. Most stunted children are too short for their age. Stunting is determined by calculating the number of under-five children whose z-score is below -2 SDs from the median HAZ of the WHO child growth standards divided by total number of under five children who are measured. According to Kenya Demographic Survey (KDHS, 2014), the national prevalence of stunting among the under-five children was 26% which was relatively higher than the average prevalence of developing countries which is 25%.

Objective: This work compares Random Forest and Elastic Net in identifying determinants of under five childhood stunting with Variable Importance as the key outcome.

Methods: The Kenya Demographic Health Survey (KDHS) women and children data was used for analysis. This data was cleaned using STATA and analyzed with R software. Due to the variance in the classes of the response variable, Synthetic Minority Oversampling Technique (SMOTE) was employed to obtain a balanced class data. Missing observations were imputed using *rfimpute* function from library *randomForest* in R software. Random Forest and Elastic Net algorithms were used to obtain determinants of stunting while Area Under the Curve (AUC) of the Receiver Operating Characteristic (ROC) Curve was used to compare the models.

Results: The top 5 factors in terms of importance according to Random Forest are: underweight status, region, child's age, ethnicity, and mother's current age. According to the Elastic Net algorithm, the top 5 important coefficient variables are: underweight children, Nairobi region, 60+ months preceding birth interval, 12-23 months old children, and children from Luhya ethnicity. In terms of the ROC values, Random Forest had an AUC of 0.92 while Elastic Net had an AUC of 0.86.

Conclusion: Based on our findings, most of the top ranked important variables selected by Random Forest and Elastic Net are similar. Nevertheless, Random Forest performed better than the Elastic Net algorithm in determining the factors of under five childhood stunting.

Keywords: Stunting, Random Forest, Elastic Net, Variable Importance, Gini Index, Area Under the Curve (AUC), Receiver Operating Characteristic Curve (ROC), Missing values

Master Thesis in Mathematics at the University of Nairobi, Kenya.
ISSN 2410-1397: Research Report in Mathematics
©Rachael Mburu, 2020
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.

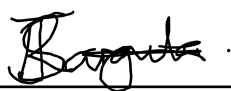
Signature

Date

RACHAEL MBURU

Reg No. I56/24855/2019

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



19-11-2020

Signature

Date

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



NOVEMBER 18, 2020

Signature

Date

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

Dedication

This Masters project is dedicated to the Almighty God for His protection, grace, wisdom and favour. Special dedication to my supportive parents, Mr. & Mrs. Mburu for their love and prayers. I would also want to dedicate this work to the church I fellowship in, Defenders Ministries International and Joan Abucheri, my mentee and friend. Thanks all for your encouragements and support.

Contents

Abstract	iii
Declaration and Approval	vi
Dedication	ix
Acknowledgments	xii
1 General Introduction	1
1.1 Background	1
1.2 Literature Review.....	1
1.3 Statement of the Problem.....	4
1.4 Objectives of the study.....	5
1.4.1 Overall Objective	5
1.4.2 Specific Objective.....	5
1.5 Significance of the study	5
2 Methodology	6
2.1 Data.....	6
2.2 Statistical methods	6
2.2.1 Balancing imbalanced response data.....	6
2.2.2 Missing data imputation	7
2.2.3 Random Forest data imputation.....	8
2.3 Random Forest	9
2.3.1 Random Forest Variable importance	11
2.4 Elastic Net	12
2.5 Model Performance Analysis	15
3 Results	17
3.0.1 Random Forest Classifiers.....	22
3.0.2 Confusion matrix from Random Forest algorithm.....	22
3.0.3 Variable Feature selection from Random forest	22
3.1 Results from Elastic Net results	24
3.1.1 Confusion matrix from Elastic Net	24
3.1.2 Variable Feature selection from Elastic Net	24
3.1.3 AUC-ROC Curves	25
3.2 Comparing the Random Forest and Elastic Net model	26
4 Discussion	27
4.1 Conclusion	27
4.2 Future Research	27
4.3 Study Limitations	27

Bibliography..... 29

Acknowledgments

I would first like to thank the Almighty God for His protection, wisdom and for enabling me to finish this research project and my Masters study successfully.

I want to express my profound gratitude to the Deltas Africa Initiative [grant 107754/Z/15/Z-DELTAS Africa SSACAB programme] for the funding and support offered to me throughout my Masters Study programme.

I would also like to thank my supervisors Dr. Rachel Sarguta and Dr. Nelson Owuor for their invaluable insights and input towards this project. God bless you.

To all the Biometry class members, am indebted to you for your willingness to share ideas. You were the best classmates.

Finally, I acknowledge anyone else who encouraged me and contributed to the success of this research.

Rachael Waithira Mburu

Nairobi, 2020.

1 General Introduction

1.1 Background

According to De Onis & Branca (2016) children with a Height-for-Age (HAZ) below -2 Standards Deviations based on the World Health Organization (WHO) child growth standards median are said to be stunted. Stunting in children is often referred to as impaired growth. The impairment develops over time, especially for children aged below five years. The reduced growth rate is due to limited access to proper nutrition, health and proper care. Stunting is characterized by slower growth rate than normal in a child. Most stunted children are too short for their age. Stunting is determined by calculating the number of under-five children whose z-score is below -2 sds from the median HAZ of the WHO child growth standards /total number of under-five children who are measured. Stunting has a wide range of negative impacts on children in their physical, emotional and cognitive development. Some of the effects include poor immunity, slow motor growth, impaired brain function, poor education performance and higher likelihood to suffer chronic diseases. The effects of stunting are irreversible hence causing long lasting impact on the child, the family and the country at large.

According to worldwide statistics in 1990, the total number of children aged five years and below who had the stunted growth was 255 million. In 2014, the number decreased to 159 million. Regardless of the decrease globally, there was an increase in Africa from 47 million in 1990 to 58 million in 2014. In order to curb these worrying trends WHO in collaboration with various governments has initiated interventions to curb childhood stunting prevalence for instance; nutritional education for pregnant women, zinc supplementation for pregnant women, macronutrient and micronutrient supplementation in children. The Kenyan government has further taken the initiative of encouraging exclusive breastfeeding, deworming, timely complementary diet and handwashing habits.

One of the key indicators of household food security is the nutritional status of children under-five. According to Matanda et al. (2014) the major indicators of childhood malnutrition are stunting, wasting, underweight and obesity. Stunting has become the main indicator of childhood undernutrition due to its high prevalence in developing countries (Black et al., 2013). Kenya Demographic Household Survey(KDHS) 2014 report shows high prevalence of stunting with 26% children being stunted and 8% severely stunted. Despite the government's interventions, malnutrition remains one of the major concerns in Kenya and has been reducing at a slow rate.

1.2 Literature Review

Kismul et al. (2018) performed logistic regression to determine how determinants of stunting operates at different levels. The study used the UNICEF conceptual framework in deciding the factors to include in the analysis. They further grouped the risk factors of stunting into:

- Distal factors; Type of residence, ethnicity, wealth quintile, province and mother's education.
- Intermediate factors comprised of environmental and maternal factors.
- Environmental factors; Hygienic toilet, family size, number of children in the family and availability of safe water supply.
- Maternal factors; Mother's age at delivery, preceding birth interval, mother's BMI and height.
- Proximal factors; Breastfeeding initiation time after birth, Birth order and had diarrhea within the past 14 days.
- Other factors; Sex and age of the child.

The results showed that the sex and age of a child, urban/rural residence, preceding birth interval, wealth quantile, province, early initiation of breastfeeding, age of the mother at delivery, mother's height and BMI, access to hygienic toilet and to safe water and mother's education were statistically significant in bivariate logistic regression. However, in multivariate logistic regression number of children in a family, mother's BMI, mother's education, access to hygienic toilet, access to clean/safe water and place of residence lost statistical significance. This study used three hierarchical logistic regression models to include different determinants of stunting. The order in which factors are entered into the model is determined by the researcher based on theory and past studies. This might be limited to the researcher's knowledge, hence the need for a more improved model in feature selection.

Takele et al. (2019) employed Generalized Linear Mixed Models (GLMM) to identify environmental, demographic, socioeconomic and health related risk factors associated with stunting for under-five children in Ethiopia. The results showed that the major determinants of childhood stunting include; sex and age of child, preceding birth interval, educational level of the mother, household wealth index, mother's BMI, toilet type, breastfeeding, use of internet and drinking water source.

Habimana & Biracyaza (2019) conducted a study to investigate risk factors for stunting in children under five years in the Western and Eastern provinces of Rwanda using

univariate and multivariate logistic regression. According to the univariate logistic regression, maternal education, sex of child, maternal occupation and age, wealth index of household, giving child fortified food, antenatal care visits and sharing a toilet had a significant association with stunting. Multiple logistic regression indicated that household wealth index, breastfeeding and gender of a child were the common risk factors of stunting in Western and Eastern Provinces. The prevalence of stunting was higher in Eastern provinces compared to Western Province. In the two provinces the prevalence was high in rural residences.

Birhanu et al. (2017) identified factors linked with stunting for 6-59 months aged children in North East Ethiopia using binary logistic regression. Bi-variable logistic regression was used to determine the factors that had a significant association with stunting. After which only the significant factors were entered into the multivariable logistic regression. This was important in controlling the possible effect of confounders. The results showed that sex and age of child, family size, literacy status of parents, rural/urban residence, frequency of feeding, giving leftover food for child and wealth index were statistically associated factors with stunting at $p - value \leq 0.05$. Initiation of complementary feeding, Dietary diversity score(food groups), pre-lacteal feeding, water treatment, child's birth order, breast feeding duration, methods of feeding, washing hand, household head, time span of exclusively breast feeding and major source of income were associated with stunting in Bi-variable logistic regression analysis but not statistically associated in multivariable logistic regression.

Chirande et al. (2015) conducted a study to investigate determinants associated with stunting and severe stunting for under-fives children in Tanzania. The study used simple and multiple logistic regression analyses. From the results, the prevalence of stunting was 35.5% and severe stunting had 14.4% for children aged 0-23 months. On the other hand, the prevalence of stunting and severe stunting for children aged 0-59 months was 41.6% and 16.1% respectively. According to the multivariable analysis the significant risk factors for severely stunted and stunted children were maternal education, gender of the child, child size at birth and source of drinking water for children aged 0-23 months and 0-59 months. A manual stepwise backward elimination method was used to identify the factors that were significantly associated with stunting. The use of Random Forest and Elastic Net would improve the feature selection process.

García Cruz et al. (2017) performed an analysis to identify the main socio-demographic, health and environmental factors of stunting for children aged 0-59 months from the Tete province in Mozambique. The analysis involved univariate and multiple logistic regression analysis. The results from univariate logistic regression showed that child age, birth weight, family size, maternal education, rural residence, maternal occupation, number of children under-five in the household, cooking fuel used, wooden or straw housing and soil floor were significant determinants of stunting. In the multiple logistic regression

the factors that remained statistically significant are birth weight and sex of the child, rural area residence, soil type of floor, presence of siblings under-five, living in houses made of straw and wood and homes where other relatives lived. The factors that were not significantly associated with stunting in the uni-variate model were not included in the joint multivariate model.

1.3 Statement of the Problem

Kismul et al. (2018) suggest a need for further studies to establish how stunting operates at different levels of determination and the main factors contributing to the development of stunting. In addition, most studies on determinants of childhood stunting have always relied on literature review and UNICEF conceptual framework to determine which factors to include in the study and afterwards determine which ones are significant. The risk factors of stunting are multi-factorial and interdependent (Habimana & Biracyaza, 2019) hence the need for machine learning models which can identify at what level the factor impact stunting.

Most of the past studies have used bi-variable and multi-variable logistic regression to investigate the determinants of childhood stunting. This model has limitations on feature selection and hence the need for a better model to select the factors associated with stunting. It is also limited in that a variable is tested for significance using bi-variable logistic regression then depending on the significance of the results it is included in the multivariable logistic regression or dropped. This means a variable that would be significant in the joint model maybe eliminated. Machine learning models are better in feature selection and can rank the variables in level of importance. Linear regression is prone to overfit with many predictor features hence the need of machine learning models which are good in preventing overfitting. In this study we will use Elastic Net and Random Forest machine learning model to determine factors associated with stunting.

1.4 Objectives of the study

1.4.1 Overall Objective

To determine the risk factors of stunting in under five children.

1.4.2 Specific Objective

1. To determine the risk factors of childhood stunting using Random Forest.
2. To determine the risk factors of childhood stunting using Elastic Net models.
3. To compare Elastic Net and Random Forest in determination of stunting risk factors.

1.5 Significance of the study

The Elastic Net and random forest methods accommodate a wide range of independent factors regardless of their significance in relation to the dependent variables. In addition, these methods can measure the variable importance in relation to the dependent variable hence will help in determining the factors that heavily cause stunting. The knowledge of the factors that have a high impact on childhood stunting prevalence will aid the government in decision making of the interventions to undertake. This will in turn create strategic techniques in the government interventions of curbing childhood stunting.

2 Methodology

2.1 Data

Secondary data obtained from Kenya Demographic Health Survey (KDHS,2014), particularly the data for under-five children and women aged 15-49 years was used for this study. The data set was taken from the main sampling frame, the Fifth National Sample Survey and Evaluation Programme (NASSEP V). This survey data is a representation of all the 47 counties in Kenya. Stratified sampling was used with the 47 counties being stratified into rural and urban. 40,399 households were considered from 1,612 clusters in the whole country with 995 of these being in rural areas and 617 clusters in urban areas. A two-stage sample design was used to select samples independently from each sampling stratum. This data set had a total of 1099 variables and 20964 observations. In this study the data was cleaned using STATA 14.1 College Station software after which we exported to R software version 3.6 and R studio 1.2.1335 for analysis. The data cleaning process involved dropping all variables that were flagged variables ,merged variables, variables used to calculate other variables, interview and sampling variables, repeated variables, variables that don't have impact over health of a child such as decision maker for using contraception, date of first marriage, reason for not having sex, index birth history, etc. Also variables that were totally missing and 90% missing. We also regrouped the variables with many categories into fewer categorical classes and renamed them accordingly.

2.2 Statistical methods

2.2.1 Balancing imbalanced response data

In cases where the response variable is binary and the ratio of one class is higher than the other, balancing is imperative. Imbalanced data degrades specific standard classifiers. The response variable (stunting) used in this study is imbalanced since the covariate have a ratio of 3:7 hence the need for balancing. Accuracy of machine learning methods is to some extent affected by unequal distribution of dependent variable in that the performance of the classifier is biased towards the class with majority. There are different methods of dealing with imbalanced data. These methods are referred to sampling methods. They modify an imbalanced data by adjusting the size of the original data to produce the same proportion of balance. Some methods of balancing imbalanced data sets are:

- Synthetic data generation

- Under-sampling
- Cost sensitive learning
- Over-sampling

In this study the Synthetic data generation has been used, specifically the Synthetic Minority Over-sampling Technique (SMOTE). SMOTE is a type of over-sampling which generates artificial data using bootstrapping and K-nearest neighbors (KNN). This mechanism works as follows:

- Calculate the difference between nearest neighbor and the vector under consideration. The distance is calculated by KNN using the euclidean distance formula below;

$$d(p, q) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + \dots + (q_n - p_n)^2} = \sqrt{\sum_{i=1}^n (q_i - p_i)^2} \quad (1)$$

where

p and q are two points in a feature space.

$p_1, p_2 \dots p_n$ are feature vectors of point p

$q_1, q_2 \dots q_n$ are feature vectors of point q

n is the dimension of the feature space.

- The difference is multiplied by a random number between zero and one.
- The outcome is added to the defined vector.
- This is followed by the selection of a random point between the current data point and one of the k neighbors.

2.2.2 Missing data imputation

Missing data is defined as unobserved values. This occurs when the actual data intended to be measured is not measured for reasons which sometimes can be controlled by the researcher while others are beyond the control of a researcher. It is a common challenge in huge data sets and especially in health and social demographic fields. Dibal et al. (2017) suggests the importance of understanding reasons for missing data and pattern of missingness before applying the respective methods of imputing missing data. According to Pedersen et al. (2017) there are three categories of missing data namely: Missing not at random - MAR, Missing completely at random -MCAR and missing at random - MNAR. Some of the most common methods of handling data with missing values are;

- Missing indicator techniques
- Complete case analysis
- Single and multiple value imputation
- Bayesian simulation methods

Most studies use the multiple value imputation techniques of imputation.

2.2.3 Random Forest data imputation

In this study, *rfimpute* function from library *randomForest* in R software is used in the imputation of data. For the categorical variable, the method imputes the missing cases by the largest average proximity while for continuous variables missing observations are imputed by the weighted average of the non-missing values k-nearest neighbors. The imputation process is repeated the number of times specified in the iteration command function. In a glimpse, if $y(p,q)$ is a missing categorical variable the equation below is used;

$$\hat{y}(p, q) = \underset{C_q}{\operatorname{argmax}} \sum_{i \neq p} \operatorname{prox}(j, p) \quad (2)$$

while in the case of a missing continuous variable the equation becomes;

$$\hat{y}(p, q) = \frac{\sum_{\substack{i \neq p \\ i \in \text{neighbor}}} \operatorname{prox}(j, p) y(j, q)}{\sum_{\substack{i \neq p \\ i \in \text{neighbor}}} \operatorname{prox}(j, p)} \quad (3)$$

In equation 2 and 3

- j = the class of categorical variable.
- p = the p_{th} observation
- q = the q_{th} variable.
- $\operatorname{prox}(j, p)$ = the proximity
- C_q = the q^{th} categorical variable.

2.3 Random Forest

Random forest is a classifier based on random family of decision classification trees. It is a machine learning algorithm proposed by Breiman (2001) which combines both the bootstrap aggregation and random subspace method to build a set of decision trees. Bootstrap aggregation is also known as bagging. According to Breiman (2001), RF contains multiple decision tree classifiers, each decision tree in the collection is formed by randomly selecting training samples and the feature attributes at each node. Afterwards, subsets from the training data sets are repeatedly drawn using the bagging method. Finally, the equal-weight voting method is used to calculate the final prediction based on the average from bootstrapped training subsets of all the decision trees.

Random forest is a bagging algorithm. Bagging is a statistical resampling method which involves random sampling of a data set with replacement. It helps in eliminating overfitting by reducing variance.

If we have a random forest model as $y \approx \hat{f}(x)$ for a data set $(y_i, x_i) \in \mathbb{R}^{p+1}$ then bootstrap aggregation work as follows;

- Generate many random sub-samples from the original data set with replacement, where $B \in \mathbb{N}$.
- Train the random forest model on each b^{th} bootstrap sample to get $\hat{f}^b(\mathbf{x})$.
- Calculate the average prediction from each model for a given data set.

In a population one would take many training sets and calculate estimators of B separate bootstrapped samples such that

$$\begin{bmatrix} \text{Sample}_1 & \text{Sample}_2 & \text{Sample}_3 \\ (y_{i1}, x_{i1}) & (y_{j1}, x_{j1}) & (y_{k1}, x_{k1}) \\ (y_{i2}, x_{i2}) & (y_{j2}, x_{j2}) & (y_{k2}, x_{k2}) \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ (y_{in}, x_{in}) & (y_{jn}, x_{jn}) & (y_{kn}, x_{kn}) \end{bmatrix}$$

Each of the above bootstrap sample imitates statistical properties of the original data. Averaging them results to a low-variance estimator. The standard error of the bootstrap estimators is

$$SE_B(\hat{\alpha}) = \sqrt{\frac{1}{B-1} \sum_{r=1}^B (\hat{\alpha}^{*r} - \frac{1}{B} \sum_{r'=1}^B \hat{\alpha}^{*r'})^2} \quad (4)$$

In the case of bootstrap, the model is trained on the b th bootstrapped training set and then average all the predictions to acquire:

$$\hat{f}_{\text{bag}}(\mathbf{x}) = \frac{1}{B} \sum_{b=1}^B \hat{f}^* b(\mathbf{x}) \quad (5)$$

This procedure is called bagging or bootstrap aggregation.

The equations below show how RF works;

$$f = f_1(x), f_2(x) \dots, f_k(x) \quad (6)$$

where:

$f_k(x)$ = decision tree while the ensemble is rf.

The decision tree parameters are defined as

$$\theta_k = \theta_{k1}, \theta_{k2}, \dots, \theta_{kp} \quad (7)$$

Hence the classifier equation below;

$$f_k(X) = f(X|\theta_k) \quad (8)$$

Each decision tree casts votes for the most common class at input X and the class with majority votes wins. The final classification $h(X)$ combines the classifiers $f_k(X)$.

Figure 1 is a summary of random forest algorithm.

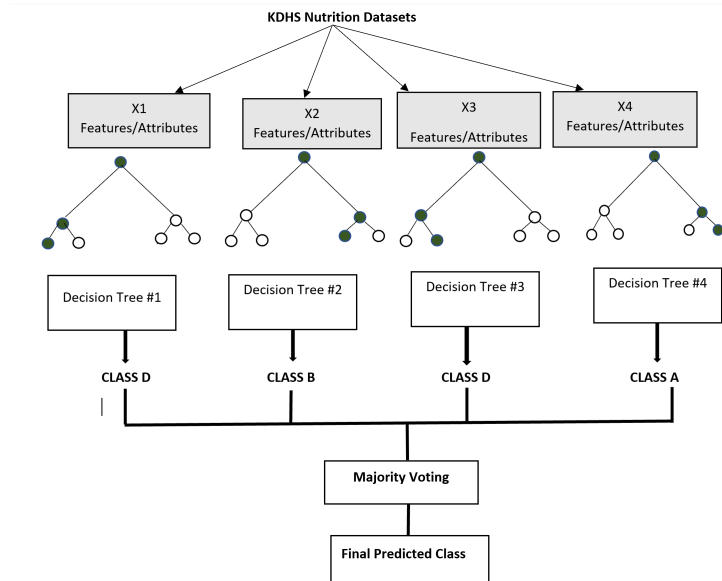


Figure 1. Random Forest Summary

In random forest classification problem, the Gini Index is an important index used to determine how nodes of a decision tree branch.

$$GI = 1 - \sum_{i=1}^C (p_i)^2 \quad (9)$$

where:

GI = Gini Index

p_i = The relative frequency of the class observed in the data set.

C = The number of classes.

The Gini Index of each tree on a node uses probability and class to determine the most likely tree to occur.

2.3.1 Random Forest Variable importance

Variable importance is the total of the impurity reductions in all the decision trees. Information gain or Gini coefficient index is used in classification trees to calculate impurity reductions.

Mean Decrease Impurity (MDI) calculates the variable importance by summing up the

Gini index decrease of each variable from 1 to the total number of trees then gets the average. The MDI equation is defined as

$$V_{\text{imp}}(x_i) = \frac{1}{n_{\text{tree}}} \left[1 - \sum_{j=1}^{n_{\text{tree}}} GI(i)^j \right] \quad (10)$$

where:

GI = Gini Index

p_i = The relative frequency of the class observed in the data set.

C = The number of classes.

2.4 Elastic Net

Elastic Net is a regularization machine learning algorithm. Regularization models are important in preventing over fitting by artificially penalizing the model coefficients. There are three common regularization models namely Lasso regression, Ridge regression and Elastic regression. This study has used Elastic Net since it is a hybrid of Lasso and Ridge models.

Elastic Net regression combines the Lasso regression penalty and ridge regression penalty. Elastic Net groups and shrinks the parameters associated with the correlated variables and leaves them in the equation or removes them all at once. This model is highly applicable when the parameters are highly correlated.

The ridge regression shrinks the regression coefficients so that variables with minimal contribution to the response variable are close to zero values but none is equal to zero. This means that it includes all predictor variables into the final model. The penalty term used to penalize the regression model for achieving the shrinkage of the coefficients is called L2-norm.

The Least Absolute Shrinkage and Selection Operator (Lasso) shrinks the regression coefficients to zero using a penalty term called L1-norm for penalizing the regression model. Unlike the ridge regression, which retains all the predictor variables, Lasso regression drops some of the correlated variables. It helps in feature selection.

Assume a data set of n observations with q number of predictors. Then let the response variable $y = (y_1, y_2, \dots, y_n)$ and $X = (x_1, \dots, x_q)$ to be the matrix of the model, where $x_k =$

$x_{11}, \dots, x_{nk})^T, k = 1, \dots, q$ are the predictors. Suppose the predictors are standardized and the response variable is centered such that,

$$\sum_{j=1}^n y_j = 0, \sum_{j=1}^n x_{jk} = 0 \text{ and } \sum_{j=1}^n x_{jk}^2 = 0 \text{ for } k=1,2,\dots,q.$$

Zhang et al. (2019) defines Elastic Net model as follows

$$L(\lambda_1, \lambda_2, \beta) = |y - X\beta|^2 + \lambda_2 |\beta|^2 + \lambda_1 |\beta|_1, \quad (11)$$

For $\lambda_1 > 0, \lambda_2 > 0$

where $|\beta|^2 = \sum_{k=1}^p \beta_k^2$,

$|\beta|_1 = \sum_{k=1}^p \beta_k$.

$\hat{\beta}$ is the estimator of Elastic Net which minimizes equation 11 as follows;

$$\hat{\beta} = \arg \min_{\beta} L(\lambda_1, \lambda_2, \beta) \quad (12)$$

In simple terms Elastic Net regularization equation is as follows.

$$SSE_{EN} = \sum_{i=1}^N (Y_i - \hat{Y})^2 + \lambda [(1 - \alpha) \sum_{i=1}^N \beta_i^2 + \alpha \sum_{i=1}^N |\beta_i|] \quad (13)$$

When alpha is equal to zero the equation becomes Ridge regression.

$$SSE_{Ridge} = \sum_{i=1}^N (Y_i - \hat{Y})^2 + \lambda \sum_{i=1}^N \beta_i^2 \quad (14)$$

When alpha is equal to one the equation becomes Lasso regression.

$$SSE_{Lasso} = \sum_{i=1}^N (Y_i - \hat{Y})^2 + \lambda \sum_{i=1}^N |\beta_i| \quad (15)$$

In the equations above the regularization parameters are; $\lambda = \lambda_1 + \lambda_2$ while $\alpha = \frac{\lambda}{\lambda_1 + \lambda_2}$

2.5 Model Performance Analysis

The Area Under the Curve of Receiver Operating Characteristic (AUC-ROC) is a curve that plots sensitivity (TPR) against FPR (1-Specificity). Confusion matrix is also important in evaluating the performance of the model. The values on the diagonal represents the True Positives (TP) and True Negatives (TN) which refer to the correct predictions whereas the values off the diagonal corresponds to the False Negatives (FN) and False Positives (FP).

Table 1. Confusion matrix summary

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

From the confusion matrix some of the major functions that can be calculated are; Accuracy is the proportion of true positives and true negatives correctly classified.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (16)$$

$$Error\ rate = \frac{FP + FN}{TP + TN + FP + FN} \quad (17)$$

Specificity is the proportion of negative classes classified correctly. In this study the number of non-stunted children that were correctly identified by the model.

$$Specificity = \frac{TN}{TN + FP} \quad (18)$$

Sensitivity is the proportion of positive classes identified correctly. In this study it would mean the proportion of stunted children that were correctly classified by the model.

$$Sensitivity = \frac{TP}{FN + TP} \quad (19)$$

Where:

True Positives (TP) : Predicted children to be stunted and it is true they are stunted.

True Negatives (TN) : Predicted children to be non-stunted, and they are non-stunted.

False Positives (FP) : Predicted they are stunted, but they are not stunted (Type I error).

False Negatives (FN) : Predicted children are not stunted, but they are stunted (Type II error).

AUC is a summary of ROC curve that measures the ability of a classifier to distinguish between classes. The higher the AUC the better the model. If $0.5 < \text{AUC} < 1$ then it is in a better position to distinguish the negative and the positive classes. However, if $\text{AUC} = 0.5$ then it means the model cannot distinguish between the positive and the negative classes. In such cases the model predicts constant or random classes for all data points.

3 Results

The summary descriptive of the data used in the analysis is as shown below.

Table 2

Variables	Levels	N (%)
Stunting status	Non stunted	10,100 (50%)
	Stunted	10,100 (50%)
Mother's current age	15-19	896 (4.4%)
	20-24	6,740 (33%)
	25-29	5,311 (26%)
	30-34	3,578 (18%)
	35-39	2,402 (12%)
	40-44	1,033 (5.1%)
	45-49	240 (1.2%)
Region	coast	2,233 (11%)
	north eastern	1,234 (6.1%)
	eastern	2,685 (13%)
	central	1,091 (5.4%)
	rift valley	5,983 (30%)
	western	1,683 (8.3%)
	nyanza	2,410 (12%)
	nairobi	2,881 (14%)

Table 3

Variables	Levels	N (%)
Number of household members	1-5 household members	11,207 (55%)
	6-10household members	8,130 (40%)
	11-15household members	793 (3.9%)
	16+ household members	70 (0.3%)
Number of under five children	0-3children	19,543 (97%)
	4-7children	657 (3.3%)
Mother's education attainment	No education	3,780 (19%)
	Primary	10,972 (54%)
	Secondary	4,474 (22%)
	Higher	974 (4.8%)
Household head age	15-25yrs	2,911 (14%)
	26-36yrs	9,420 (47%)
	37-47yrs	4,608 (23%)
	48-58yrs	1,853 (9.2%)
	59yrs and above	1,408 (7.0%)
Wealth_index	poorest/poorer	10,241 (51%)
	middle	2,924 (14%)
	richer/richest	7,035 (35%)
Total children ever born	1-5children	16,338 (81%)
	6-10children	3,658 (18%)
	11-15children	204 (1.0%)
Wasting status	Non wasted	17,960 (89%)
	Wasted	2,240 (11%)
Mother's height	150 cm	67 (0.3%)
	>=200 cm	20,133 (100%)
Main wall material	natural walls	9,003 (45%)
	rudimentary walls	4,539 (22%)
	Finished walls	5,885 (29%)
	Others	426 (2.1%)
	not a resident	347 (1.7%)

Table 4

Variables	Levels	N (%)
Mother's age at 1st birth	5-15yrs	3,431 (17%)
	16-26yrs	16,130 (80%)
	27-37yrs	629 (3.1%)
	38-48yrs	10 (<0.1%)
Child twin status	single birth	19,667 (97%)
	1st of multiple	257 (1.3%)
	2nd of multiple	276 (1.4%)
	3rd of multiple	0 (0%)
	4th of multiple	0 (0%)
	5th of multiple	0 (0%)
Child had diarrhea recently	no	16,245 (80%)
	yes, last 24 hours	0 (0%)
	yes, last two weeks	3,953 (20%)
	don't know	2 (<0.1%)
Child's age	0-6 months	1,747 (8.6%)
	7-11 months	1,366 (6.8%)
	12-23 months	6,321 (31%)
	24-35 months	3,765 (19%)
	36-47 months	3,715 (18%)
	48-59 months	3,286 (16%)
Underweight_status	Non underweight	15,678 (78%)
	Underweight	4,522 (22%)
Type of residence	Urban	7,804 (39%)
	rural	12,396 (61%)
In charge prenatal assistance	Doctor	3,552 (18%)
	Nurse/midwife	7,058 (35%)
	No one	801 (4.0%)
	Some other	8,789 (44%)

Table 5

Variables	Levels	N (%)
Source of drinking water	Piped	7,740 (38%)
	borehole	1,565 (7.7%)
	well	2,907 (14%)
	spring	2,279 (11%)
	river/dam/lake/ponds/stream	4,303 (21%)
	others	1,406 (7.0%)
Time to get to water source	less than 30 minutes	6,054 (30%)
	30 minutes or longer	7,050 (35%)
	water on premises	6,663 (33%)
	not a resident	352 (1.7%)
	other	81 (0.4%)
Type of toilet	flush toilet	1,204 (6.0%)
	pit latrine	14,132 (70%)
	no facility	4,864 (24%)
Roof_material	natural roofing	19,212 (95%)
	rudimentary roofing	331 (1.6%)
	Others	657 (3.3%)
Household head gender	male	14,985 (74%)
	female	5,215 (26%)
Household phone ownership	no	19,775 (98%)
	yes	71 (0.4%)
	not a de jure resident	354 (1.8%)
Child's birth order number	1-3	15,276 (76%)
	4-6	3,845 (19%)
	7-10th born	956 (4.7%)
	above 10th	123 (0.6%)
Child's sex	male	10,413 (52%)
	female	9,787 (48%)

Table 6

Variables	Levels	N (%)
Ethnicity	embu	124 (0.6%)
	kalenjin	2,874 (14%)
	kamba	1,468 (7.3%)
	kikuyu	1,935 (9.6%)
	kisii	915 (4.5%)
	luhya	4,530 (22%)
	luo	1,755 (8.7%)
	maasai	591 (2.9%)
	meru	645 (3.2%)
	mijikenda/ swahili	1,070 (5.3%)
	somali	1,315 (6.5%)
	taita/ taveta	179 (0.9%)
	turkana	612 (3.0%)
	samburu	571 (2.8%)
other	1,616 (8.0%)	
Type of cooking fuel	electricity/lpg	1,839 (9.1%)
	biogas/kerosene	1,695 (8.4%)
	coal,agri.crops/animal dung	16,308 (81%)
	no food cooked in house	15 (<0.1%)
	other	5 (<0.1%)
	not a resident	338 (1.7%)
Mother's current marital status	single	1,016 (5.0%)
	married	17,671 (87%)
	widowed	419 (2.1%)
	seperated/divorced	1,094 (5.4%)
Preceding birth interval	7-17 months	2,666 (13%)
	18-23 months	2,076 (10%)
	24-35 months	6,980 (35%)
	36-47 months	2,830 (14%)
	48-59 months	1,861 (9.2%)
	60+ months	3,787 (19%)

3.0.1 Random Forest Classifiers

Figure 2 shows the number of trees versus the error. The black line represents the out of bag samples over the amount of trees while the coloured lines indicate the error for the stunted and non stunted classes of the response variable. The number of trees producing the lowest error rate is 443. The error rate is decreasing with the increase in the number of trees.

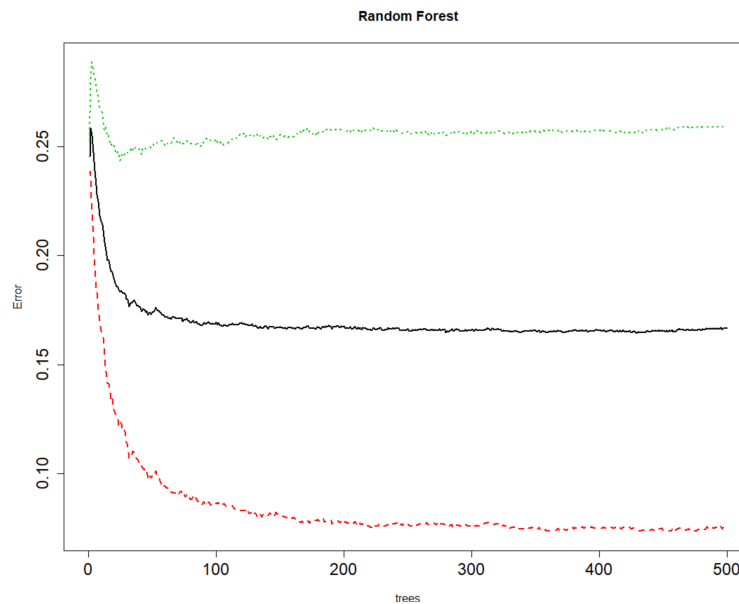


Figure 2. number of trees

3.0.2 Confusion matrix from Random Forest algorithm

Table 7 shows the confusion matrix results of Random Forest algorithm. It is a summary of the prediction results of this algorithm in classifying stunted and non-stunted children. Both correct and incorrect classes are represented in the table below.

Table 7. RF confusion matrix

	Reference	
Prediction	Non stunted children	Stunted children
Non stunted children	2,820	772
Stunted children	206	2,262

3.0.3 Variable Feature selection from Random forest

The mean decrease in impurity (MDI) and mean decrease accuracy are the two methods for checking the variable importance in random forest technique. The mean decrease accuracy also known as the gini importance evaluates how much the accuracy of the model

decreases when a variable is dropped. The greater the decrease the more significant the variable is. The MDI is used to calculate the feature importance. Figure 3 shows the MDI results from our study. The variables with a high importance have a high impact on childhood stunting.

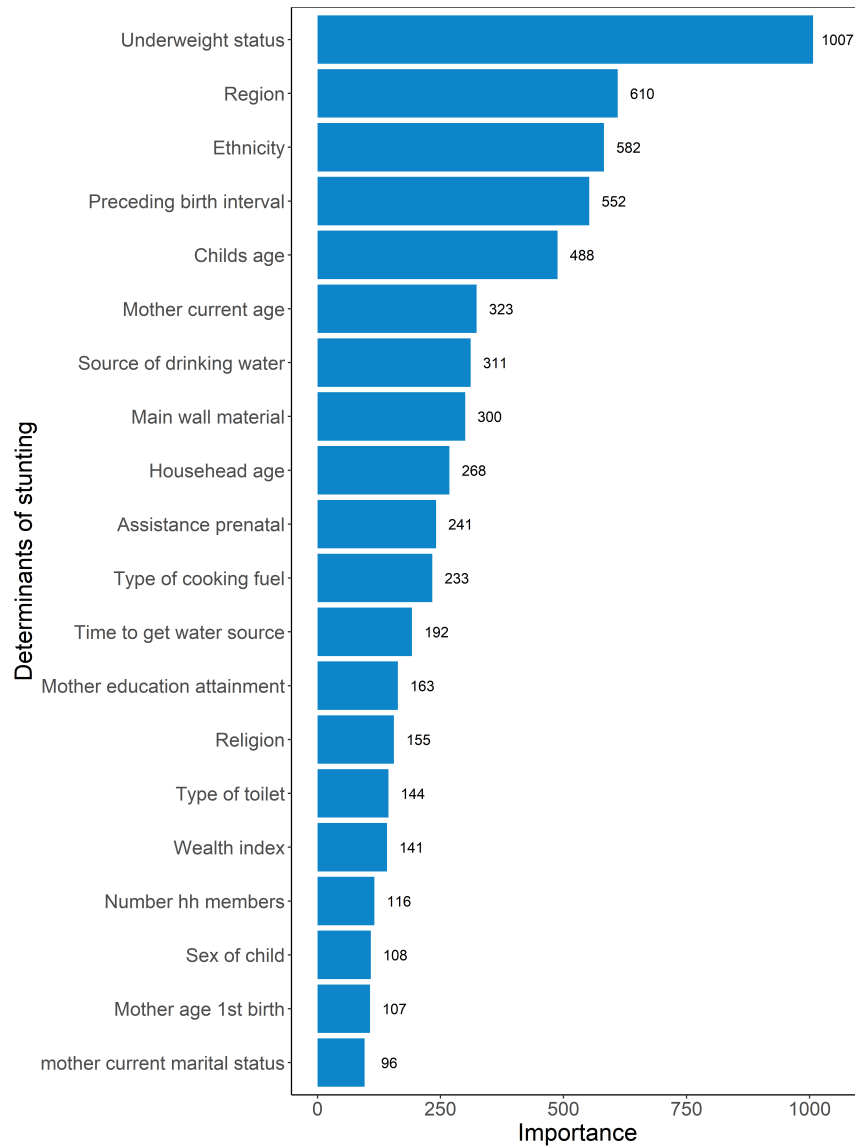


Figure 3. Random forest variable importance

3.1 Results from Elastic Net results

According to Elastic Net the prediction results of the correct and incorrect classes are shown in the confusion matrix table 8

3.1.1 Confusion matrix from Elastic Net

Table 8. EN confusion matrix

	Reference	
Prediction	Non stunted children	Stunted children
Non stunted children	6,285	1,932
Stunted children	789	5,134

3.1.2 Variable Feature selection from Elastic Net

The top 20 most important variables in determining stunting in under-five childhood are as figure 4 shows. Elastic Net specifies the covariate of that variable that is of high importance in impacting stunting.

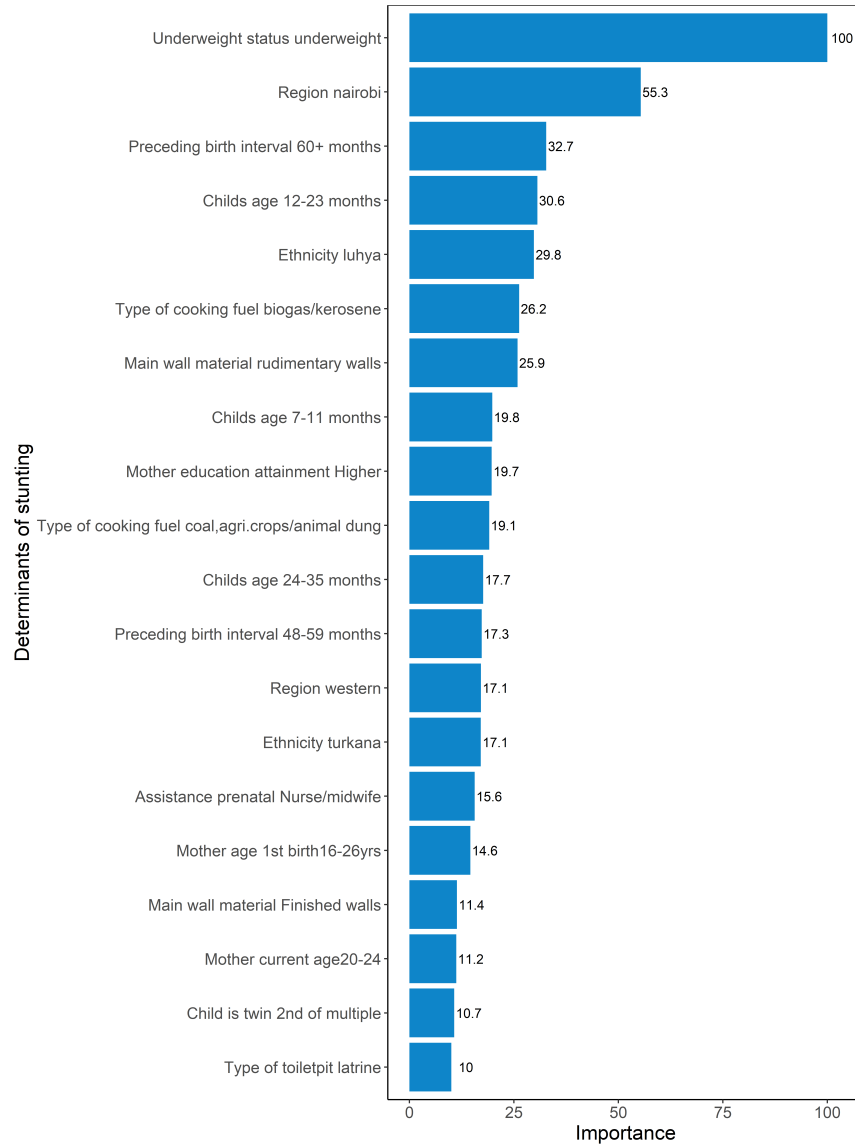


Figure 4. Elastic Net variable importance

3.1.3 AUC-ROC Curves

The results from rf algorithm shows that the model had a AUC of 0.92. This means that the algorithm was pretty good in identifying the correct classes. From the ROC curve, the AUC measure of Elastic Net is 0.86 which is relatively good but lower than that of random forest algorithm.

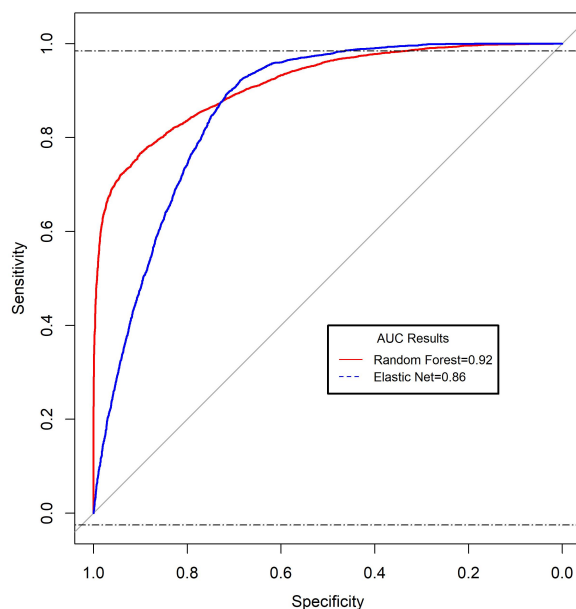


Figure 5. AUC ROC curves of RF and EN

3.2 Comparing the Random Forest and Elastic Net model

The random forest algorithm performs much better compared to Elastic Net algorithm. This is confirmed by AUC of RF (0.92) while that of Elastic Net is 0.86. However, most of the most important variables according to RF also appear in the list of those selected by Elastic Net.

Table 9. Model performance in analysing determinants of stunting in under five children

Classification	Elastic Net	Random Forest
Accuracy (95% CI)	0.8076 (0.801, 0.814)	0.8394 (0.8296, 0.8483)
Sensitivity	0.7266	0.7459
Specificity	0.8885	0.9326
AUC	0.8619	0.9168

4 Discussion

According to Random Forest algorithm the top 10 most important variables are underweight status of the child, Region, Ethnicity, Preceding birth interval, Child's age, Mother's current age, Source of drinking water, main wall material, House head age and the in-charge prenatal assistance. In evaluating variable importance, Elastic Net narrows down to the specific coefficients of a certain variable while random forest identifies only the variable without identifying the specific class level. Elastic Net identified the top 10 coefficients that have a high impact on under five stunting to be underweight children, children from Nairobi region, preceding birth interval of 60+ months, children of 12-23 months of age, children from Luhya ethnicity, children from homes that cooked using biogas/kerosene and the main wall material being rudimentary walls, the children within 7-11 months age group, mothers who had a high attainment education and homes that cooked using coal/agricultural crops /animal dung.

Random Forest had a higher AUC of 0.92 while Elastic Net had AUC of 0.86. Hence Random Forest had a 92% chance distinguishing between the positive and the negative class while Elastic Net had a capability of 86%. In addition, Random Forest had an accuracy of 84% which was slightly higher than Elastic Net which had 81% accuracy.

In a study by Sanchez-Pinto et al. (2018) on comparing different algorithms in variable selection, Elastic Net and Random Forest were among the methods compared. Hence these methods are essential in determining variables that are relatively important.

4.1 Conclusion

The goal of feature selection is that it increases the accuracy classification. Based on our results, most informative variables according to Random Forest and Elastic Net algorithm were similar. However, Random Forest performed better than Elastic Net algorithm. Incorporating the machine learning methods in nutrition studies will help greatly in revealing the factors that have a high impact on childhood malnutrition.

4.2 Future Research

Future research would consider comparing more machine learning algorithms and determine which among them performs best in this type of research.

4.3 Study Limitations

The study did not compare results from non imputed data with that from imputed data. In the future, I would recommend comparison of more than two algorithms and also results from both imputed and non-imputed data sets.

Bibliography

- Birhanu, A., Mekonen, S., Atenafu, A., & Abebaw, D. C. (2017). Stunting and associated factors among children aged 6-59 months in lasta woreda, north east ethiopia, 2015: A community based cross sectional study design. *J. Fam. Med*, *4*, 1112.
- Black, R. E., Victora, C. G., Walker, S. P., Bhutta, Z. A., Christian, P., De Onis, M., ... others (2013). Maternal and child undernutrition and overweight in low-income and middle-income countries. *The lancet*, *382*(9890), 427-451.
- Breiman, L. (2001). Random forests.
- Chirande, L., Charwe, D., Mbwana, H., Victor, R., Kimboka, S., Issaka, A. I., ... Agho, K. E. (2015). Determinants of stunting and severe stunting among under-fives in tanzania: evidence from the 2010 cross-sectional household survey. *BMC pediatrics*, *15*(1), 165.
- De Onis, M., & Branca, F. (2016). Childhood stunting: a global perspective. *Maternal & child nutrition*, *12*, 12-26.
- Dibal, N. P., Okafor, R., & Dallah, H. (2017). Challenges and implications of missing data on the validity of inferences and options for choosing the right strategy in handling them. *International Journal of Statistical Distributions and Applications*, *3*(4), 87-94.
- García Cruz, L. M., González Azpeitia, G., Reyes Suárez, D., Santana Rodríguez, A., Loro Ferrer, J. F., & Serra-Majem, L. (2017). Factors associated with stunting among children aged 0 to 59 months from the central region of mozambique. *Nutrients*, *9*(5), 491.
- Habimana, S., & Biracyaza, E. (2019). Risk factors of stunting among children under 5 years of age in the eastern and western provinces of rwanda: Analysis of rwanda demographic and health survey 2014/2015. *Pediatric health, medicine and therapeutics*, *10*, 115.
- Kismul, H., Acharya, P., Mapatano, M. A., & Hatløy, A. (2018). Determinants of childhood stunting in the democratic republic of congo: further analysis of demographic and health survey 2013-14. *BMC public health*, *18*(1), 74.
- Matanda, D. J., Mittelmark, M. B., & Kigaru, D. M. D. (2014). Child undernutrition in kenya: trend analyses from 1993 to 2008-09. *BMC pediatrics*, *14*(1), 5.
- Pedersen, A. B., Mikkelsen, E. M., Cronin-Fenton, D., Kristensen, N. R., Pham, T. M., Pedersen, L., & Petersen, I. (2017). Missing data and multiple imputation in clinical epidemiological research. *Clinical epidemiology*, *9*, 157.

Sanchez-Pinto, L. N., Venable, L. R., Fahrenbach, J., & Churpek, M. M. (2018). Comparison of variable selection methods for clinical predictive modeling. *International journal of medical informatics*, *116*, 10–17.

Takele, K., Zewotir, T., & Ndanguza, D. (2019). Understanding correlates of child stunting in ethiopia using generalized linear mixed models. *BMC public health*, *19*(1), 626.

Zhang, F., Sun, K., & Wu, X. (2019). A novel variable selection algorithm for multi-layer perceptron with elastic net. *Neurocomputing*, *361*, 110–118.