DEVELOPMENT AND VALIDATION OF A RISK PREDICTION MODEL FOR CHILD STUNTING USING CROSS-SECTIONAL DATA

MASTER OF SCIENCE (BIOSTATISTICS) THESIS

JONATHAN THOMAS MKUNGUDZA

UNIVERSITY OF MALAWI

CHANCELLOR COLLEGE

AUGUST 2023

DEVELOPMENT AND VALIDATION OF A RISK PREDICTION MODEL FOR CHILD STUNTING USING CROSS-SECTIONAL DATA.

MSc. (BIOSTATISTICS) THESIS

BY

JONATHAN THOMAS MKUNGUDZA

BSc. (Mathematical Sciences Statistics and Demography) University of Malawi

Submitted to the Department of Mathematical Sciences, Faculty of Science, in partial fulfilment of the requirement for the Degree of Master of Science in Biostatistics

University of Malawi

AUGUST 2023

DECLARATION

I, the undersigned, declare that this thesis/dissertation is my original work, which has not been

submitted to any other institution for similar purposes. Where other people's work has been used, acknowledgements have been made.

_____JONATHAN THOMAS MKUNGUDZA_____

Full Legal Name

_____

Signature

_____

Date

CERTIFICATE OF APPROVAL

The undersigned certifies that this thesis represents the student's work and effort and has been submitted with our approval.

Signature_____ Date_____
Samuel Manda PhD (Professor)
Main Supervisor

Signature_____ Date_____
Halima Sumayya Twabi, MSc (Lecturer)
Co-supervisor

Signature_____ Date_____
Patrick Sawerengera, MSc (Lecturer)
Programme Coordinator

DEDICATION

To my late brother Pearson, my wife, Emma and my children Angelina, Jonathan, and Joana

ACKNOWLEDGMENTS

I would like to thank whole-heartedly my supervisors: Professor Samuel Manda and Halima

Sumayya Twabi, for their guidance, support, constructive ideas, comments, time, and knowledge of my thesis work. I appreciate the encouragement and their critiques of my work which have helped me structure the entire thesis with a clear understanding.

I sincerely thank God for guiding me through this work. Special thanks to my wife and children for your support throughout my work. Thank you for giving me strength and encouragement.

ABSTRACT

Child stunting, defined as impaired height for age, is a major indicator of severe undernutrition and is more prevalent in Sub-Saharan Africa. Individual child stunting risk factors for childhood stunting are well-studied and known. This study aimed at assessing the viability of combining individual child stunting risk factors into a simple risk factor prediction model that could be used to predict stunting among children aged 5 years or lower. Firstly, a systematic review of risk factors for childhood stunting was conducted. Secondly, using stunting data on nearly 5,000 children aged 5 years or below in the Malawi Demographic Health Survey (MDHS 2015-16) we identified risk factors that were used in the primary multivariate logistics model for child stunting. Thirdly, several reduced models were then obtained depending on the variable selection algorithm that included backward, forward, stepwise, random forest, Least Absolute Shrinkage and Selection Operator (LASSO), and own subjective judgment. Finally, from each reduced multivariable logistic model, a stunting risk score, based on its coefficients, was calculated for each child. The stunting risk prediction models were assessed using discrimination measures including area under-receiver operator curve (AUROC), sensitivity and specificity. The systematic review produced 68 predictor variables of child stunting, of which 67 were available from the 2016 MDHS dataset, and 27 had complete information. The common risk factors selected by all the variable selection methods include household wealth index, age of the child, household size, type of birth (singleton/multiple births), and birth weight. The best cut-off point on the child stunting risk prediction model was 0.37. The best predictive model was based on risk factors determined by the judgment method, which had AUROC 64% (95% CI: 60%-67%) in the test data. For children residing in urban areas, the AUROC was 67% (95% CI: 58-76%) as opposed to those in rural areas, AUROC =63% (95% CI: 59-67%). The derived child stunting risk prediction model could be useful as a first screening tool to identify children more likely to be at risk of stunting. The identified children could then receive necessary nutritional interventions.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS AND ACRONYMS

AIC Akaike's Information Criteria

AUROC Area under the receiver operating curve

BIC Bayesian Information Criteria

HIV Human Immuno-deficiency Virus

NSO National statistics office

WHO World Health Organization

CHAPTER 1

INTRODUCTION

1.1. Introduction

child malnutrition persists to pose a significant public health challenge in the sub-Saharan African region. About 3.1 million children are projected to lose their lives each year, either directly or indirectly, due to malnutrition. Moreover, around 165 million children experience the long-term effects of stunted growth, limiting their full growth potential (Shinsugi , et al., 2015). According to the 2016 national data on childhood undernutrition in Malawi, it was found that 37% of young children experienced stunting, 3% were affected by wasting, and 12% were underweight (National Statistical Office, 2017). In 2019, the prevalence levels of stunting, underweight, and wasting in Malawi were documented as 39%, 12%, and 2%, respectively (Mtenda, 2019). The prevalence of stunting, greater than 20% in under five children is regarded as a public health concern by the World Health Organization (WHO) (Uwiringiyimana , Veldkamp, & Ocke, 2019). Stunting (child's height-for-age z-score) serves as a marker of inhibited linear growth and the aggregate deficit in growth faced by children (Akombi, Agho, Astell-Burt, Hall, & Renzaho, 2017). Stunting affects child health adversely by causing continued consequences, such as damaged cognitive abilities and educational performance during childhood, which could have harmful implications for adult health and economic productivity (Aguayo V. M., Nair, Badgaiyan, & Krishna, 2016).

The framework established by the World Health Organization (WHO) outlines comprehensively the factors that influence stunting. These factors are classified into four primary closely related factors in the WHO framework: household and family-related factors, insufficient practices regarding complementary feeding, inadequate practices concerning breastfeeding, and infections (Stewart, Iannotti, Dewey, Michaelsen, & Onyango, 2013). Numerous research studies have studied several factors influencing stunting as discussed in the subsequent section.

1.2. Factors Associated with Stunting.

1.2.1. Predictor Search Strategy

The study conducted a search of PubMed and Google Scholar databases for relevant articles between August and December 2021. Several searches were performed with the search terms "Determinants of stunting AND Africa" or "Risk factors of stunting AND Africa" or "Predictors of stunting AND Africa". All duplicate articles were eliminated from the results. In total, 28 articles were considered for the final identification of predictors of stunting in the sub-Saharan African region.

1.2.2. Demographic Factors

According to the study done by Mtambo et al. (2018), the primary factors influencing child stunting in Malawi were identified as child sex, sex of the household head, type of residence, maternal education, ethnicity, child age, and maternal height. Child sex was indicated as a significant determinant of child stunting (McDonald, et al.; Akombi, Agho, Astell-Burt, Hall, & Renzaho, 2017; Nshimyiryo , et al., 2019; Woldeamanuel & Tesfaye, 2019; Uwiringiyimana , Veldkamp, & Ocke, 2019; Bukusuba, Kaaya, & Atukwase, 2017; Chirande, et al., 2015; Nkurunziza , Meessen, Van geertruyden , & Korachais, 2017; Dake, Solomon, Bobe, Tekle, & Tufa, 2019). The geopolitical zone was reported to be an important feature of stunting in a study that was done in Nigeria (Akombi, Agho, Astell-Burt, Hall, & Renzaho, 2017)

Research conducted in Tanzania observed that the age of a household head <35 years was a significant feature of stunting (Semali, 2015). The analysis that was done by Haile and others in 2016, the results showed that being male and belonging to a household with a male head were identified as factors that raised the probability of being stunted. Furthermore, the study observed that children in the age bracket of 24 and 35 months had higher chances of experiencing stunting in comparison with children who were under one year old. The educational qualification of the father, and the mother's body mass index (greater or equal to 25.0kg/m2) were identified as some of the community-level factors linked with stunting (Haile , Azage , Mola , & Rainey,

2016)Various studies have reported mother's education is a significant predictor of child stunting (Akombi, Agho, Astell-Burt, Hall, & Renzaho, 2017; McDonald, et al., 2012; Nshimyiryo , et al., 2019; Chirande, et al., 2015; Nkurunziza , Meessen, Van geertruyden , & Korachais, 2017; Cruz, Azpaitia, Suarez, Rodriguez, & Ferrer, 2017; Habimana & Biracyaza, 2019; Haile , Azage , Mola , & Rainey, 2016; Kofi, 2018). Nkurunziza found that marital status was a contributing factor of child stunting among children in the age category of 6-23 months. Some studies found that residence was an important predictor of child stunting (Kismul , Acharya , Mapatano , & Hatløy, 2018; Woldeamanuel & Tesfaye, 2019; Cruz, Azpaitia, Suarez, Rodriguez, & Ferrer, 2017; Mehta, Suchdev, Rhodes, & Williams, 2018). Chirande (2015) conducted an in Tanzania and found that the birth order of the child and geographical region were important determinants of stunting. It was also found in other studies that the mother's height was a factor affecting child stunting (Kismul , Acharya , Mapatano , & Hatløy, 2018; Berhe, Seid, Gebremariam, Berhe, & Etsay, 2019)

Studies conducted in Africa found that the count of under-five children in the household was associated with child stunting (Fikadu , Assegid, & Dube, 2014; Berhe, Seid, Gebremariam, Berhe, & Etsay, 2019; Cruz, Azpaitia, Suarez, Rodriguez, & Ferrer, 2017; Nkurunziza , Meessen, Van geertruyden , & Korachais, 2017; Kofi, 2018). Studies conducted in Ethiopia and the Democratic Republic of Congo revealed that maternal age was an important variable affecting stunting (Woldeamanuel & Tesfaye, 2019; Kismul , Acharya , Mapatano , & Hatløy, 2018); Additional case-control research was done in Tigray, North Ethiopia, which concluded that the body mass index (BMI) of the mother played a significant role in detecting child stunting within the region (Berhe, Seid, Gebremariam, Berhe, & Etsay, 2019). It was reported that family size was one of the predictors of stunting among under-five children (Fikadu , Assegid, & Dube, 2014; Cruz, Azpaitia, Suarez, Rodriguez, & Ferrer, 2017). Child age was also found to be a significant determinant of child stunting (Berhe, Seid, Gebremariam, Berhe, & Etsay, 2019; Shinsugi , et al., 2015; Chirande, et al., 2015; Uwiringiyimana , Veldkamp, & Ocke, 2019; Dake, Solomon, Bobe, Tekle, & Tufa, 2019; Sema , Azage , & Tirfie, 2021). A study implemented in Ghana revealed that ethnicity was among the factors associated with child stunting (Kofi, 2018). Using standard regression methods on the Rwandan Demographic and Health Survey, Habibana and Biracyaza indicated that maternal age was a factor influencing stunting among children

under-five 5 years of age in Rwanda (Habimana & Biracyaza, 2019). Low maternal height and mother's literacy were also reported to be predictors of stunting among children under-five years (Nshimyiryo , et al., 2019). Another study with a specific emphasis on the practices related to complementary feeding was rolled out in Rwanda and the results showed that the child's age and the caretaker's BMI were important predictors of child stunting (Uwiringiyimana , Veldkamp, & Ocke, 2019). In Ethiopia, children belonging to Muslim, Orthodox, and other traditional religious communities were observed to have increased chances of facing stunting when contrasted to children from the Protestant community (Gebru , Haileselassie, Temesgen, Seid, & Mulugeta, 2019).

1.2.3. Economic Factors.

Several studies, particularly those conducted in Africa, have found an association between economic factors and stunting. One study in Malawi revealed that the working status of the mother and the availability of radio/TV were significant predictors of child stunting (Mtambo et al., 2018). Several authors concurred with each other that the household wealth index was an important determinant of stunting among children in sub-Saharan countries (Akombi, Agho, Astell-Burt, Hall, & Renzaho, 2017; Nkurunziza , Meessen, Van geertruyden , & Korachais, 2017; Chirande, et al., 2015; Habimana & Biracyaza, 2019). Residing in houses constructed with wood or straw, or lacking a proper floor, as well as cooking with charcoal, were identified as factors influencing stunting among children less than five years old in Mozambique. (Cruz, Azpaitia, Suarez, Rodriguez, & Ferrer, 2017). Mother's occupation was reported to be an important determinant of child stunting (Keino, Ettyang, & Borne, 2014; Keino, Ettyang, & Borne, 2014). Bakasuba reported that food insecurity and type of housing were also important determinants of child stunting (Bukusuba, Kaaya, & Atukwase, 2017). Household income had also been reported to be a predictor of stunting among children less than five years old (Keino, Ettyang, & Borne, 2014). It was found that children from households that were rearing animals were less likely to be stunted than those from households that were not rearing animals (Shinsugi , et al., 2015). It had been revealed that household poverty was also a determinant of child stunting among children in poor countries (Nshimyiryo , et al., 2019; Kismul , Acharya , Mapatano , & Hatløy, 2018). In Malawi, it was learned that children from mothers who were on health insurance were less likely to be stunted than children from mothers who were not on

health insurance (Afolabi & Palamuleni, 2019). Children from households that were depending on food from the farms (own production) were found to have decreased chances of being stunted than those who were depending on purchased food from the market (Tariku , Biks , Derso , Wassie, & Abebe, 2017)

1.2.4. Child Caring Practices and Environmental Health Factors

Practices related to childcare and environmental health factors have been reported as determinants of child stunting in studies conducted in some countries around Africa. In Malawi, it was shown that vitamin A supplementation, vaccination coverage and period of breastfeeding were associated with child stunting. (Mtambo , Masangwi, & Kazembe , 2014) Akombi and others conducted a study in Nigeria and found that prolonged duration of breastfeeding (>12 months) was a determinant of child stunting (Akombi, Agho, Astell-Burt, Hall, & Renzaho, 2017). One study reported that an unimproved water supply and vitamin A deficiency were associated with stunting (Mehta, Suchdev, Rhodes, & Williams, 2018). Another study revealed that the duration of exclusive breastfeeding, period of breastfeeding and method of feeding supplementary food were predictors of child stunting (Fikadu , Assegid, & Dube, 2014). Where one gets drinking water was also reported to be a significant predictor of child stunting in various research conducted in Africa (Mtambo , Masangwi, & Kazembe , 2014; Fikadu , Assegid, & Dube, 2014; Kismul , Acharya , Mapatano , & Hatløy, 2018). It was shown that early initiation of breastfeeding was an important determinant of child stunting among under-five children (Kismul , Acharya , Mapatano , & Hatløy, 2018; Kofi, 2018). Consuming fortified food, visiting antenatal care facilities, sharing toilets and breastfeeding were given as crucial determinants of child stunting in a study carried out in Rwanda (Habimana & Biracyaza, 2019). A separate study carried out in Rwanda demonstrated that exclusive feeding during the preceding six months and dietary intake of zinc were identified as predictors of child stunting (Uwiringiyimana , Veldkamp, & Ocke, 2019). The results of some studies in East Africa showed that deworming tablet use was also a significant predictor of child stunting (Uwiringiyimana , Veldkamp, & Ocke, 2019; Nshimyiryo , et al., 2019). Kofi conducted a study in Ghana and concluded that exposure to a proper toilet facility and visiting a health centre were some of the predictors of stunting among children less than five years old (Kofi, 2018). The use of family planning

methods and pre-breastfeeding were also observed to be determinants of stunting in children (Dake, Solomon, Bobe, Tekle, & Tufa, 2019). Like in other studies in Africa, caregivers' knowledge of stunting and initiation time to complementary food were reported to be determinants of child stunting in a study conducted by Bakasuba (Bukusuba, Kaaya, & Atukwase, 2017). Nkurunziza and others found that distance to a health facility can also predict child stunting (Nkurunziza , Meessen, Van geertruyden , & Korachais, 2017). The availability of improved latrine facilities was also reported to determine child stunting (Haile , Azage , Mola , & Rainey, 2016). Continued breastfeeding for 1 year was revealed to be a significant factor associated with child stunting (Nsereko, et al., 2018). Children whose mothers did not consistently use water and soap for handwashing had higher odds of childhood stunting (Sema , Azage , & Tirfie, 2021). Krasevec and others found that children who were not given food from animal source on the previous day had elevated chances of being stunted contrasted to children who were given all three groups of food from animal sources (eggs, meat, and dairy) (Krasevec, An, Kumapley, Bégin, & Frongillo, 2017). WHO dietary diversity score was reported to be associated with stunting (Berhe, Seid, Gebremariam, Berhe, & Etsay, 2019; Krasevec, An, Kumapley, Bégin, & Frongillo, 2017). A study conducted in Ethiopia had shown that feeding powdered or fresh milk, feeding formula, eating organ meat, and taking fruits high in beta-carotene as part of the diet, and vegetables were significant factors linked to stunting (Ayelign & Zerf, 2021).

1.2.5. Obstetric Conditions, Child Illness and Additional Maternal Factors.

Stunting in children is also linked to obstetric conditions, child illness and additional maternal-related factors as observed in some studies conducted in sub-Saharan Africa. It was revealed that infectious diseases were important predictors of stunting in children less than five years in Malawi (Mtambo , Masangwi, & Kazembe , 2014). Macdonald and others found that child HIV infections and low Apgar score at birth were important predictors of child stunting (McDonald, et al., 2012). It was revealed that the birth size of the child, place of delivery and low birth weight were some of the predictors of child stunting ((Akombi, Agho, Astell-Burt, Hall, & Renzaho, 2017; Nkurunziza , Meessen, Van geertruyden , & Korachais, 2017; Chirande, et al., 2015). Chirande and others concluded that the type of delivery assistance was one of the factors

affecting child stunting (Chirande, et al., 2015). Diarrhoea episodes were reported in various studies as a predictor of child stunting among under-five children in sub-Saharan Africa (Dake, Solomon, Bobe, Tekle, & Tufa, 2019; Woldeamanuel & Tesfaye, 2019; Akombi, Agho, Astell-Burt, Hall, & Renzaho, 2017; Berhe, Seid, Gebremariam, Berhe, & Etsay, 2019). Haile and others reported that short birth intervals and severe anaemia were factors associated with child stunting (Haile , Azage , Mola , & Rainey, 2016). A study that was done in Ethiopia showed that having a fever influenced stunting (Sema , Azage , & Tirfie, 2021). Various studies had shown that being multiple births also significantly increases the odds of childhood stunting (Gebru , Haileselassie, Temesgen, Seid, & Mulugeta, 2019; Afolabi & Palamuleni, 2019; Ayelign & Zerf, 2021).

As seen in the preceding review, several potential predictors of stunting in children less than five years old have been studied. These are summarized in **Table 1**

Table 1. Characteristic of the selected studies on child stunting and associated risk factors (no of studies =28)

Title	Authors, Year	Country	Type of data	Number of children	Method	Risk factor identified
Analysis Of Childhood Stunting in Malawi Using Bayesian Structured Additive Quantile Regression Model	Owen P. L. Mtambo, Lawrence Kazembe and Salute Masangwi (2014)	Malawi	DHS	2138	Bayesian Structured Additive Quantile Regression Model	Sex of child, Sex of household head, Type of residence, mother's working status, Vitamin A supplementation, Vaccination coverage, Availability of radio/TV, Source of drinking water, infectious diseases, maternal education, ethnicity, child age and duration of breastfeeding
Prevalence and Determinants of Stunting in Under-five Children in Central Tanzania: a remaining threat to Achieving Millennium Development Goal 4	Innocent Antony Semali, Anna Tengia-Kessy, Elia John Mubanga, and Germana Leyna (2015)	Tanzania	Survey data	678	multivariate logistic regression	Age of household head <35yrs, Maternal education, and ownership of the mobile phone

Determinant of Stunting among Preschool Children in the 2015-2016 Malawi Micronutrient Survey	Rukshan Mehta, Anne Williams (2018)	Malawi	Survey data		block stepwise logistic regression	underweight status, Vitamin A deficiency, Unimproved water supply, rural residence, household hunger
Determinant of Stunting and Severe stunting among Burundian Children Aged 6-23 months: Evidence from a national cross-sectional household survey	Sandra Nkurunziza, Bruno Meesen, Jean-Piere Van Geertruyden, Catherine Karachais (2014)	Burundi	Survey data	6199	Binary and multivariate logistic regression	low birth weight, size of the baby at birth, sex of the child, Place of delivery, family wealth index, maternal education, marital status, distance to a health facility, severe food insecurity, and number of under-five years children in the household.
Determinants of stunting and Severe Stunting among Under-Five in Tanzania: evidence from the 2010 cross-sectional household survey.	Lulu Chirande, Debora Charwe, Hadijar Mbwana, Rose Victor, Sebas Kimboka, Abukar Ibrahim Isaka (2015)	Tanzania	Survey data	7324	multiple logistic regression	maternal education, sex of child (male), age of the child, household wealth index, place of delivery, type of delivery assistance, birth order of the child, the perceived size of the baby at birth, source of drinking water, geographical region
Determinant of Childhood Stunting in the Democratic Republic of Congo: Further Analysis of Demographic and Health Survey 2013-14	Hallgeir Kismul, Pawan Acharya, Mala Ali Mapatamo and Anne Hatloy (2018)	Democratic Republic of Congo	DHS	9030	logistic regression	Province, poverty, residence (rural), mother's height, source of drinking water, early initiation of breastfeeding, childbirth intervals, mother's age >20yrs
Risk factors of Stunting (Chronic Undernutrition) of Children Aged 6 to 24 Months in Mekelle City, Tigray Region North Ethiopia: Unmatched case-control Study	Kidanemay Berhe, Omer Seid, Yemane Gebremariam Almez Berhe, Natnael Etsay (2019)	Ethiopia	Survey data	330	logistic regression	mother's height, mother's body mass index, Childbirth weight, number of under-five children in the household, repeated diarrhoea episodes and WHO dietary diversity score
Factors Associated with Under-five Stunting, Wasting and Underweight Based on Ethiopian Demographic Health Survey Dataset in Tigray Region Ethiopia	Berhann Teshome Woldeamanuel, and Tigist Tigabie Tesfaye (2019)	Ethiopia	DHS	1077	multivariate binary logistic regression	maternal age, source of drinking water, sex of the child, antenatal follow-ups, diarrhea episodes, household wealth, birth weight, and residence(rural)

Factors Associated with Stunting Among Children According to the Level of Food Insecurity in the Household: a cross-sectional Study in a rural community of Southern Kenya	Chisa Shinsugi, Masaki Matsumura, Mohamed Karama, Junichi Tanaka, Mwatasa Changoma and Satoshi Kaneko (2015)	Kenya	Survey data	404	multivariate logistic regression	age of the child, animal rearing, number of siblings younger than school age
Factors Associated with Stunting Among Children Aged 0 to 59 Months from the Central Region of Mozambique	Loida M. Garcia Cruz, Gloria Gonazlez Azpeitia, Desderio Reyes Suarez, Alfredo Santana Rodriguez, Juan Francisco Loro Ferrer and Lluis Serra-Majem (2017)	Mozambique	Survey data	282	multiple logistic regression	birthweight, maternal education status, maternal occupation, living in rural areas, family size, number of children under-five years of age in the household, cooking withcharcoal, inhabiting wooden or straw housing or housing without a proper floor, duration of breastfeeding complementary feeding
Early feeding practices and stunting in Rwandan Children; a cross-sectional study from the 2010 Rwanda Demographic and Health Survey	Etienne Nsereko, AssumptaMukabutera, Damien Lyakaranye, Yves Didier Umwungerimwiza, Valens Mbrushimana, Manasse Nzayirabaho (2018)	Rwanda	DHS	1634	multivariate logistic regression	continued breastfeeding for 1 year
Predictors of Childhood Stunting in Ghana: A cross-sectional survey of the Association Between Stunting among children under age five and maternal bio-demographic and socioeconomic characteristics in Ghana 2014	Janet Oyedi Kofi (2018)	Ghana	DHS	2759	logistic regression	early initiation of breastfeeding, access to proper toilet facility, mother's level of education, ethnicity, access to health care, number of under-five children in the household (>20
Predictors of Stunting among Children 6-59 months of Age in Sodo Zuria District, South Ethiopia: a community-based cross-sectional study	Samson Kastro Dake, Fithamlak Bisetegen Solomon, Testahun Molla Bobe, Habtamu Azene Tekle and Efrata Girma Tufa (2019)	Ethiopia	Survey data	342	multivariate logistic regression	sex of the child, Age of the child, use of family planning, diarrhea morbidity, Pre-lacteal feeding

Predictors of Stunting with Particular Focus on Complementary feeding practices: a cross-sectional Study in the Northern Province of Rwanda	Vestine Uwiringiyimana, Marga C. Ockey, Sherif Amer, Antonie Veldkamp (2018)	Rwanda	Survey data	138	logistic regression	Age of child, exclusive breastfeeding, deworming tablet use in the previous 6 months, caretaker body mass index and dietary zinc intake.
Risk Factors for Stunting among Children Under-five Years: a cross-sectional population-based Study in Rwanda Using 2015 Demographic and Health Survey	Alphonse Nshimyryo, Bethany Hedt-Gauttier, Christine Mutaganzwa, Catherine M. Kirk, Kathryn Beck, Albert Ndayisaba, Joel Mubiligi Fredrick Kateera and Ziad El-Khatib (2019)	Rwanda	DHS	3594	logistic regression	Sex of child, age of the child, low birth weight, low maternal height, mother's education, mother's literacy, deworming tablet use, poverty of household.
Predictors of stunting in Children Aged 6 to 59 Months: a case-control study in Southwest Uganda	John Bakusuba, Archileo N. Kaaya, Abel Atukwase (2017)	Uganda	Survey data	168	multiple logistic regression	Sex of the child, food insecurity, initiation time to complimentary food, caregiver's knowledge about stunting and type of housing.
Risk Factors of Stunting Among Children Under-five 5 Years of Age in the Eastern and Western Provinces of Rwanda: Analysis of Rwanda Demographic and Health Survey 2014/2015	Samuel Habimana, Emmanuel Biracyaza (2019)	Rwanda	DHS		multiple logistic regression	maternal education, maternal age, maternal occupation, wealth index, sex of the child, fortified food intake, antenatal care visit, breastfeeding
Stunting and Severe Stunting among Children Under Five Years in Nigeria: a multilevel analysis	Blessings J. Akombi, Kingsley E. Agho, John J. Hall, Andre M.N Renzabo Thomas Astell-Burtand Dafna Merom (2017)	Nigeria	DHS	24529	multilevel logistic regression	sex of the child(male), mother's perceived birth size (small and average), household wealth index, duration of breastfeeding (more than 12 months) geopolitical zone, and diarrhea episodes prior to the survey.
Factors Associated with Stunting Among Children of 24 to 59 Months in Meskan district, Gurage Zone, South Ethiopia: a case-control study.	Teshale Fikadu, Sahilu Assegid, Lamessa Dube (2014)	Ethiopia	Survey data	242	logistic regression	family size, number of under-five children in the household, maternal occupation, duration of exclusive breastfeeding, duration of breastfeeding and method of feeding complementary food.

Exploring Spatial Variations and Factors Associated with Childhood Stunting in Ethiopia: Spatial and multilevel analysis	Damewoz Haile, Muluken Azage, Tegegn Mola, Rochelle Rainey (2016)	Ethiopia	DHS	9893	multilevel multivariate logistic regression	short birth interval, sex of the child, sex of household head, age of the child, severe anaemia, mother's education, father's education level, mother's body mass index, family wealth, and availability of improved latrine facilities
Predictors of Stunting, Wasting and Underweight among Tanzanian Children Born to HIV-infected Women	CM McDonald, R. Kupka, K.P Manji, J Okuma, R. J Bosch, S Abound, R. Kisenge, D. Spiegelman, W. W. Fawzi and C. P. Duggan (2012)	Tanzania	survey data	2387	Multivariate Cox proportional hazards	low maternal education, few household possessions, low infant birth weight, child HIV infection, sex of the child and low Apgar score at birth
Childhood Stunting and Associated Factors among Irrigation and Non-irrigation Users, northwest Ethiopia: A comparative cross-sectional Study	Balew Sema, Muluken Azage, Mulat Tirfie (2021)	Ethiopia	Survey data	1164	Multivariate logistic regression	Child age, ANC visit, fever, ways of hand washing habits
Diet Quality and Risk of Stunting among Infants and Young Children in Low-and Middle-Income Countries	Julia Krasevec, Xiaoyi An, Richard Kumapley, France Begin, Edward A. Froyginllo. ()	LMIC	DHS	74548	Multiple logistic regression	Dietary diversity, animal source food consumption (ASF)
Determinants of Stunting among Under-five Children in Ethiopia: A Multilevel mixed- effects analysis of 2016 Ethiopian Demographic and health survey data	K. Fantay Gebru, W.Mekoonnen Haileselassie, A.Hafton Temegen, A. Oumar Serd, B.Afework Mulugeta. (2019)	Ethiopia	EDHS	8855	Multilevel logistic regression	Child age, child size at birth, child sex, maternal education, poverty, multiple births, religion
Determinants of Stunting among Under-five Children in Malawi	Felix Afolabi, Martin E Palamuleni (2019)	Malawi	DHS	5707		Child sex, anaemia, location, wealth index, mothers' education, multiple births, child size at birth, mother's weight, health insurance,

Household, Dietary and Healthcare Factors Predicting Childhood Stunting in Ethiopia	Abebe Ayelign, Taddese Zerfu (2021)	Ethiopia	EDHS	11023	Logistic regression	Maternal education, child sex, possession of refrigerator, possession of television, multiple births, type of cooking fuel, feeding powdered or fresh milk, formula feeding, consumption of organ meat, ANC follow-up, birth size deworming during pregnancy, feeding beta-carotene rich fruits and vegetables, house main floor materials
Stunting and its Determinant Factors among children aged 6-59 months in Ethiopia.	Amare Tariku, Gashaw Andargie Biks, Terefe Derso, Molla Mesele Wassie, Solomon Mekonnen Abebe (2015)	Ethiopia	Survey data	1295	Logistic regression	Mother's occupation, postnatal vitamin A supplementation, wealth index, Source of family food
Prevalence and Determinants of Concurrent Wasting and Stunting and Other Indicators of Malnutrition among Children 6-59 months old in Kersa, Ethiopia	Akhlu Abrham Roba, Nega Assefu, Yadeta Dessie, Abebe Tolera Kedir Teji, Hemler Elena,Lilia Bliznashka, Wafaise Fawzi.	Ethiopia	Survey data	1091	Logistic regression	Child age, child sex, cough, maternal, education maternal occupation, maternal BMI

About 28 articles were reviewed in this study. **Table 1** presents details of the studies that were reviewed. Many of the studies considered in the analysis were implemented in countries located in Eastern Africa. The sample sizes ranged from 138 to 74,548. The factors consistently linked to stunting, as indicated in **Table 1**, include the child's sex, maternal education, geographical location, the count of under-five children in the household, family size, wealth index, instances of diarrhea, birth weight, multiple births, and the age of the child. The study applies a series of variable selection methods using a multivariable logistic regression to identify the best predictive factors for child stunting.

1.3. Variable Selection

According to Heinze et al. (2017), statistical models can be described as straightforward

mathematical principles derived from empirical data that depict the relationship between an outcome and multiple explanatory variables. One of the problems in building a simple model is how to choose a subset of independent features to include in a model among the many variables that a researcher is presented within the dataset. Many scientists know the existence of several variable selection algorithms and their use, but they do not know that they produce poor-performing models (Ratner, 2010). Model building strategy is dependent on the purpose and the discipline of study. Some variable selection algorithms work well in other disciplines and do not perform well in other disciplines. The purpose of different variable selection methods in model fitting is to develop a simpler statistical model that is valid, provides predictions with acceptable accuracy, and is practically useful (Heinze, Wallisch, & Dunkler, 2018).

In most cases, the use of variable selection algorithms has been guided by the researcher's preference or experience. Automated techniques such as stepwise methods are commonly used and can be done using several statistical software packages that are on the market (Liao & Lynn, 2010). This has been the case even though the efficiency of stepwise selection of variables compared with other strategies such as all possible subsets, forward and backward elimination and LASSO in modelling health-related outcomes using logistic regression is not known. The weakness of stepwise in binary logistic regression is well documented. The R-squared values that are provided by the stepwise method of selecting variables tend to exhibit a strong bias towards higher values, and the regression coefficients derived from them are biased and require adjustment (Ratner, 2010).

A balance must be struck between model complexity and its usefulness when building a model. Furthermore, it is important to apply judgment based on the researcher's expertise in the subject area so that variable selection is not only driven by statistical significance. Without this balance, one runs the risk of having a model with covariates without any predictive significance.

The process of selecting variables has gathered significant attention in various fields of research, including the sector of health, and has become a focal point of extensive research. The variable selection offers numerous advantages, including augmenting the predictive performance of a model, providing a more concise and cost-effective set of variables by reducing training and

utilization time, enabling data visualization, and providing a complete comprehension of the fundamental data generation process (Chowdhury & Turin, 2020). There are many determinants of stunting, however, it is difficult to use all the determinants to predict stunting. Hence, it would be more robust to select the best predictors that will help predict stunting. This can be achieved by applying the various statistical methods that have been developed for variable selection. The main problem faced then would be to build a model from a broad range of variables that should be incorporated into the "optimal" model to predict child stunting in sub-Saharan Africa.

1.4. Aims and Objectives.

The overall objective of this thesis was to develop and validate a child stunting prediction score in the context of sub-Saharan Africa (SSA). This was accomplished through the following objectives:

a) A review of selected studies to detect predictors of stunting in children, aged 0-59 months in SSA.

b) To compare the selected predictor variables of child stunting between six variable selection methods, namely forward selection, backward elimination, and stepwise selection; Least absolute shrinkage and selection operator (LASSO); and random forest

c) To compare the discriminative performance of the selected six sets of variables ((in b) above) in a multivariate logistic regression model for risk prediction score for child stunting.

CHAPTER 2

ANALYSING BINARY OUTCOME DATA

2.1. Binary Outcomes in Health Studies

In health sciences research, several outcomes are measured on different scales. A common measurement is where the interest is in the existence or nonexistence of a disease or condition resulting in binary outcomes. For example, in medical research interest could be in assessing whether a patient is dead or alive, the success of a treatment (cured or not cured) and whether a child has a growth condition (stunted and not stunted). Assume for subject *i*, $i = 1.....,n$ a binary response Y with categories 0 and 1 is observed. In this study Y=1, represents a stunted child and Y=0 represents a child who is not stunted. Several approaches are used to analyze binary outcomes which include probit, logistic regression, naïve Bayes, decision trees, support vector machine, and k-nearest neighbour. The most common approach is the logistic regression model because it does not require greater computational capacity. Therefore, logistic regression is comparatively simpler to implement, interpret, and train when compared to other machine learning models. Logistic and probit models do not have many differences. The differences between probit and logistic regression are just theoretical, the logistic model employs logit transformation, whereas the probit model utilizes the inverse Gaussian link for their respective computations. In this study, logistic regression is employed. The next section provides a brief description of statistical approaches for binary outcomes with much emphasis on the binary logistic model.

2.2. Binary Logistic Regression Model.

Consider the response Y as defined in section 2.0 where Y_i takes values of 0 or 1 for the child i and considers observed data as $O = (Y, X)$, and $X^T = (X_1, X_2 ... X_K)$ is the observed 1 by q vector of covariates representing the characteristics of a child i. If the study assume that $\pi(x_i)$ is a probability that a child i with covariates X_i takes a value $Y_i = 1$, the distribution for this

outcome, $Y_i = 1$ is specified by the Bernoulli distribution as

$$p(Y_i = 1 | X_i = x_i) = p_i(x_i)^{1-y_i}; y_i = 0,1$$

The logistic regression then fits the probability function.

$$\pi(x) = \frac{\exp(\alpha + \beta x)}{1 + \exp(\alpha + \beta x)} \quad (1)$$

The probability that the subject does not have the outcome (stunting) is $1 - \pi(x)$, thus one can have.

$$1 - \pi(x) = 1 - \frac{\exp(\alpha + \beta x)}{1 + \exp(\alpha + \beta x)} \quad (2)$$

$$1 - \pi(x) = \frac{1 + (\exp(\alpha + \beta x)) - (\exp(\alpha + \beta x))}{1 + \exp(\alpha + \beta x)} \quad (3)$$

Equation 3 simplifies equation 4 as follows.

$$1 - \pi(x) = \frac{1}{1 + \exp(\alpha + \beta x)} \quad (4)$$

Therefore, the odds of a child experiencing stunting is expressed as

$$\frac{\pi(x)}{1 - \pi(x)} = \frac{\exp(\alpha + \beta x)[1 + \exp(\alpha + \beta x)]}{1 + \exp(\alpha + \beta x)} \quad (5)$$

Simplifying equation 5 gives equation 6 below.

$$\frac{\pi(x)}{1 - \pi(x)} = \exp(\alpha + \beta x) \quad (6)$$

This is the proportion of the likelihood of being stunted and the likelihood of not being stunted. By taking the log of the odds of being stunted which is expressed as a function of the covariates gives equation 7:

$$\log \frac{\pi(x)}{1 - \pi(x)} = (\alpha + \beta x) \quad (7)$$

A researcher is interested in observing if the probability of being stunted is higher or lower than the odds of not being stunted.

2.3. Parameter Estimation

Logistic regression intends to approximate the unknown parameters, in equation 4. Equations attained with the ultimate likelihood approximation which requires discovering a collection of parameters for which the chance of the recorded data is maximum. The highest likelihood expression is obtained from the probability distribution of the response variable (Czepiel, 2002).

By utilizing this approach, values of β are derived to optimize the likelihood function. As each y_i corresponds to an individual binomial count within the i^{th} population, the contribution of each subject (child) i to the likelihood function is determined for a specific value of the predictor X, and the function can be presented as.

$$P(Y=1|x)^{y} \times P(Y=0|x)^{1-y} \qquad (8)$$

Hence, when Y equals 1, the contribution is represented as $P(Y=1|x)$, and when Y equals 0, the contribution transforms into $P(Y=0|x)$.

Therefore, the joint probability density function of Y is given by multiplying the individual contributions, and it is gotten by:

$$f(y/\beta) = \prod_{i=1}^{N} \binom{n_i}{y_i} \pi_i^{y_i} (1-\pi_i)^{n_i - y_i} \qquad (9)$$

Where $\binom{n_i}{y_i}$ are various permutations for arranging y_i successes (stunted children) from among n_i trials (children). π_i is the chance of a child being stunted for any single of the n_i children, and $1-\pi_i$ is the chance of a child who is not stunted.

The values of β are stated based on predetermined fixed values for y, and this can be presented as.

$$L(\beta/y) = \prod_{i=1}^{N} \binom{n_i}{y_i} \pi_i^{y_i} (1-\pi_i)^{n_i - y_i} \qquad (10)$$

The quantities of β that optimize Equation 10, are referred to as the greatest likelihood estimates. The log-likelihood expression is a more conceivable form of the function above. It is

formed by applying the natural logarithm to equation 10. In general, the log-likelihood is easier to work with, and mathematically it is expressed as.

$$l(\beta / y) = \sum_{i=1}^{n} [y_i \log(\pi_i) + (1 - y_i) \log(1 - \pi_i)] \quad (11).$$

By differentiating this function concerning β and equating the expression to zero, the estimated quantities of the parameters can be obtained. To work out the equations and obtain the resulting values β, an iterative method known as Newton-Raphson is employed. Other methods are used to estimate model parameters. The Markov Chain Monte Carlo (MCMC) estimation algorithm is also used for parameter estimation (Nemeth, 2014). Others have used ridge regression estimation methods to estimate parameters in regression (Dorugade, 2014).

2.4. Variable Selection Methods

Numerous methods for variable selection have been proposed; however, there is no consensus on a single approach that consistently performs well under all circumstances. Therefore, for each dataset, the technique for variable selection should be carefully chosen (Khiabani, Ramezankhani, Azizi, & Hadaegh, 2015). One direct technique for variable selection involves leveraging subject matter expertise acquired through literature review and expert consultations. However, it is worth noting that these options may not always be accessible. Another frequently employed method involves utilizing p-values for examining statistically significant predictors either through univariable analysis or by employing a multivariable forward or backward selection process**.** Various variable selection techniques are formally available in purchased software packages. Commonly used techniques, which are of interest in this thesis are forward selection, backward elimination, stepwise, least angle regression and shrinkage (LARS), Least absolute shrinkage and selection operator **(**LASSO**)** and random forest variable selection. Some of these variable selection methods are discussed below.

2.4.1. Backward Elimination (BE)

It is the most straightforward algorithm for selecting features. It begins with a model that includes all potential features. One by one, the features are removed from the model until only

those that contribute significantly to the outcome remain in the model. The removal process starts with the feature that has minimal impact on the model. The feature that has the smallest p-value below the threshold value or the variable with the utmost p-value above the threshold value is considered to provide a minimal contribution. After removing the least significant feature, the model is modified with that variable excluded, and the p-values are recalculated. This repetitive process continues, removing variables having the minimum p-value or the largest p-value exceeding the designated threshold in each model, which is then refitted accordingly. The approach is replicated until all remaining features are regarded as significant at the specified threshold value. This specified threshold figure is termed as 'p-to-remove' and should not always be put at 0.05 (Chowdhury & Turin, 2020). Pual and others recommended a p-value between 0.15 to 0.2 (Paul , Pennell, & Lemeshow, 2013). This is to make sure that all relevant variables are included in the model.

2.4.2. Forward Selection (FS).

This technique for feature selection is the opposite of the backward elimination algorithm. The method commences with zero variables in the model and afterwards, variables are incrementally added to the model until none of the variables not incorporated in the model can introduce any substantial impact to the model's output. During each repetition, the added variables are assessed for potential addition in the model. The p-value is calculated if an added variable is considered. The variable that yields the highest test statistic exceeding the cutoff value or the smallest p-value below the cutoff value is chosen and incorporated into the model. Essentially, the variable with the highest level of significance is prioritized for addition. Subsequently, the model is readjusted to include this variable, and new p-values are figured for the features that remain in the model. Once more, the variable that has the highest test statistic surpassing the cutoff value or the lowest p-value below the cutoff value is selected from the features that remain and are included in the model. This procedure is continued until no more features are significant at the designated cutoff value when included in the model. A feature that is included in the model will not be removed from the model. (Chowdhury & Turin, 2020)

2.4.3. Stepwise Selection

It involves both forward and backward selection methods, allowing the inclusion and removal of variables in different steps. It can begin with either a backward elimination or a forward selection process. If forward selection is selected, variables are appended to the model one after the other according to their statistical significance. After each addition, the process assesses every variable already incorporated in the model and removes any that are not significant. This process goes on until all variables in the model are significant and all excluded variables are insignificant. This approach is sometimes considered an altered version of forward selection, although variables incorporated into the model may not necessarily stay in it. On the other hand, if backward elimination is the starting point, variables are originally eliminated from the model with all variables based on statistical significance. However, if any of the previously excluded variables later indicate significance, they are added again into the model. This process involves iteratively selecting the feature offering the least contribution to eliminate from the model. After this, all eliminated variables are reassessed for potential reintroduction. Two distinct significance levels (cut-offs) are required in Stepwise selection for removing and adding the variables in the model. The significance value for incorporating features should be more accurate compared to that for removing variables to avoid the process from entering an infinite loop. Backward elimination is often preferred within stepwise selection because it analyzes the model with all features`` and evaluates the impact of all contender variables (Chowdhury & Turin, 2020).

2.4.4. Least Absolute Shrinkage and Selection Operator (LASSO)

A penalty is applied to the totality of squares or log-likelihood, which corresponds to the absolute addition of regression coefficients. LASSO regulates the selection of features by reducing the residual addition of squares while ensuring that the addition of the absolute figures of the coefficients stays below a constant threshold, t . Mathematically, it can be represented as follows.

$$J(\beta)\arg\min_{\beta} = \| y - X\beta \|^2 + \lambda(n)\sum_{j=1}^{p} | \beta | \qquad (12)$$

The use of LASSO as a feature selection technique can be seen from the fact that decreasing the values of λ lead to shrinkage of regression coefficients and some of these even become zero.

2.4.5. Least Angle Regression and Shrinkage (LARS)

Least angle regression (LARS) is a sophisticated approach to model selection that can be considered an advancement of the stagewise algorithm, providing quick calculations (Iturbide et al., 2013). The approach begins by loading all coefficients as zero and captures the covariates that show the highest correlation with the response variable. After that, LARS takes a step of maximum magnitude in the route of this independent variable until another independent variable becomes equally correlated with the remaining residual. At this stage, LARS continues by moving in a route that has equal angles between the two features until the K-th feature is included in the model denoted as β_k. In the case where K is equal to the total number of covariates, a logistic model is obtained. The objective is to select an appropriate value for K that results in a more straightforward and more inclusive model. A cross-validation procedure is employed to choose the optimal number of independent variables to be incorporated into the ultimate model.

2.4.6. Random Forest

Random forest is built upon the bagging technique. This technique involves creating multiple subsets of the original dataset through resampling with replacement. Each subset is then used to train a separate model, and the final prediction is obtained by aggregating the predictions of all the individual models to each sample. The random forest technique is employed to calculate variable importance metrics, allowing for the ranking of variables based on their predictive importance. Permutation importance is employed, which is computed by comparing the prediction performance before and after permuting the variable values, averaged across all trees. The importance calculation in each tree only considers out-of-bag observations (Degenhardt, Seifert, & Szymczake, 2019). The variables that have large importance values are relevant for prediction and those variables with values of importance close to zero, are said to have no association with the outcome of interest.

2.4.6.1. Boruta Method

Boruta was developed as an extension to the Random Forest algorithm, and it is a popular ensemble learning method. Boruta is designed to detect the most relevant features in a dataset by comparing them to randomized versions of the features in the dataset.

The main idea behind Boruta is to determine the importance of features by comparing their performance to that of randomly created "shadow" features. These shadow features are created by permuting the values of the original features while keeping the target variable unchanged. The Boruta algorithm then uses an altered Random Forest model to assess the importance of the original features comparative to the shadow features.

During the procedure, Boruta allocates a measure of importance, called the "Z-score," to each variable. The Z-score indicates the degree of evidence that a variable is truly important compared to the shadow variable. Boruta increasingly eliminates immaterial features by iteratively comparing their Z-scores to a threshold value.

At the end of the procedure, Boruta produces a set of variables that have been selected to be significantly more important than the shadow features. These important features can be used for further analysis or as input to other machine learning models.

2.4.7. Judgement Variable Selection Method

Numerous methods for variable selection have been proposed; however, there is no consensus on a single approach that consistently performs well under all circumstances. Therefore, for each dataset, the technique for variable selection should be carefully chosen (Khiabani, Ramezankhani, Azizi, & Hadaegh, 2015). In statistical analysis, prior knowledge derived from scientific literature is considered the primary basis for determining the inclusion or exclusion of covariates. However, such information may not always be accessible for all research questions (Walter & Tiemeier, 2009). The judgement variable selection method relies on field expertise acquired through reviewing relevant literature and consulting with experts.

2.5. Stopping Rule/Selection Criteria in Variable Selection.

It is important to know when to stop the process of including and excluding variables during the variable selection procedure. A standard significance level for hypothesis testing such as a p-value is often used. Other criteria that are also used as stopping rules are Akaike's information criterion (AIC), Bayesian information criterion (BIC), and Mallows' C_p statistic. These are also employed as model assessment tools, and they are discussed below.

2.6. Model Selection Methods

When selecting a criterion for model selection, it is acknowledged by the researchers that models serve as approximations of reality. When provided with a dataset, the goal is to identify the candidate model that best approximates the data. This entails attempting to minimize the loss or reduction of information. As such, AIC, BIC, and Mallows' C_p statistics are used for model selection.

2.6.1 Akaike's Information Criterion (AIC)

The Akaike's information criterion (AIC) was established by Akaike in 1973. It is a mathematical technique applied to judge the degree of alignment between a model and the data from which it was derived. The AIC (1973) is defined as

$$AIC = 2K - 2\ln L\left(\hat{\beta}\right) \tag{13}$$

Where K is the number of estimated parameters in the candidate model and $L\left(\hat{\beta}\right)$ is the estimate from the log-likelihood function. AIC quantifies the comparative information content of a model by utilizing maximum likelihood estimates and counting the number of parameters involved in the model, as indicated in the above-mentioned formula. It is employed to assess and distinguish various potential models, helping in the identification of the best-fit model that is consistent with the given data. It is also used as a stopping rule in variable selection methods. The model giving the smallest AIC over the set of models considered is selected as the best model.

For a small sample size, a modified form called AICc is used instead of the AIC above. The AICc is given by.

$$AIC_c = -2\log L(\hat{\beta}) + 2K + \frac{(2K+1)}{(n-k-1)} \quad (14)$$

Where n is the sample size

2.6.2. Bayesian Information Criterion (BIC)

Another method for scoring and selecting a model is the Bayesian information criterion (BIC). It uses optimum likelihood estimates like AIC. Mathematically it is expressed as

$$BIC = -2L(\hat{\beta}) + K\log(n) \quad (15)$$

Where K is the number of parameters estimated in the candidate model and $L(\hat{\beta})$ is the estimate from the log-likelihood function and n is the size of the sample. Through the blending of a punishment term based on the number of independent parameters, the Bayesian Information Criterion (BIC) tends to prioritize models that display simplicity or parsimony (A. Berchtold, 2010). The BIC imposes a severe penalty on more complex models, making them have larger scores and less likely to be selected (Jason Brownlee, 2019). Like in AIC, the model exhibiting the minimum BIC score is selected as the superior model.

2.6.3. Mallows' Cp Statistic

The Mallows' C_p criterion was put forward by Mallows in 1972. It relies on the calculation of the mean sum of squared errors (MSSE) as the basis. In the context of a model with P-independent features, the MSSE can be stated as.

$$MSSE_P = E(RSS_P) + 2p\sigma^2 - n\sigma^2 \quad (16)$$

The assumption in the C_P criterion is that the model with all the K-independent variables involved is correct (Sembiring and Tarigan, 2018). C_p criterion for a smaller model fitted using any subset with p-independent variables where P<k, is expressed as.

$$C_P = RSS_p + 2p\sigma^2 - n\sigma^2 \quad (17)$$

This σ^2 is an unbiased estimator and is estimated by the following, $\hat{\sigma}^2 = \frac{RSS_P}{n-p}$, where RSS_K is the residual of the sum of squared values within the model with all the K variables. As in AIC

and BIC minimum C_p denote the best model.

2.7. Assessment Methods for Prediction Models

A variety of varied algorithms and measures of performance can be utilized to evaluate the effectiveness of prediction models. The commonly used measures for binary response variables include the following: Sensitivity, Positive Predictive Value (PPV) and Negative Predictive Value (NPV), Specificity and receiver operating characteristics (ROC). The Brier score, concordance (or C) statistic, and the goodness-of-fit statistic have also been used (Steyerberg & Vergouwe, 2014). The area under the receiver operator curve is applied to judge the capability of the overall model to classify the outcomes of a disease condition.

2.7.1. Receiver Operating Characteristics (ROC)

Receiver Operating Characteristics (ROC) curves are often used to access the ability of a risk factor to predict an outcome. Often a risk factor is included in a logistic regression model to forecast the likelihood, for example, of a child being stunted. These predictive probabilities or risks can be examined to see how accurate they are at identifying children who would be stunted or not stunted. Discrimination is commonly measured using ROC curves. The AUC - ROC curve is a way to evaluate how well a model can classify data into different categories at different threshold levels. The ROC curve is a graphical representation of the model's capability to differentiate between the categories, and the AUC (Area Under the Curve) is a numerical representation of this ability. A larger AUC reveals a better ability of the model to distinguish between the categories, like how a model that can better differentiate between stunted and non-stunted children would have a higher AUC.

In this process, the predicted probabilities of stunting are repeatedly dichotomized into above versus below cut-off points. For each cut-off point, one can estimate the sensitivity (probability that the predicted risk is above the cut-off point among stunted children) and specificity (probability that the predicted risk is below the cut-off point among children who are not stunted). The ROC curve is a plot that illustrates the correlation between sensitivity and 1-

specificity across various potential cut-off points. This is plotted by varying the cut-off points to display a spectrum of sensitivity versus specificity. The area under the ROC curve (AUROC) is a useful metric for summarizing the ROC curve. If the ROC curve reaches the top corner of the plot (100% sensitivity and 100% specificity) then the model is said to have perfect discrimination. A diagonal ROC curve indicates random classification. For binary outcomes, the concordance statistic is identical to the AUROC (Steyerberg & Vergouwe, 2014). The AUROC was used to measure discrimination in which models were used to predict acute kidney injury (Davis, Lasko, Chen, Siew, & Matheny, 2017).

2.7.2. **Sensitivity, Specificity, Positive Predictive Value (PPV) and Negative Predictive Value (NPV).**

Sensitivity, negative predictive value (NPV), and positive predictive value (PPV), specificity, are significant statistical measures used in diagnostic testing and screening tools. These measures provide details regarding the effectiveness or accuracy of a test or screening tool.

Sensitivity refers to the capacity of a test to accurately detect individuals who exhibit stunting (true positives). By dividing the count of correctly identified positive cases (TP) by the summation of true positives (TP) and false negatives (FN), sensitivity can be calculated. Sensitivity = TP / (TP + FN). Sensitivity is an indicator of the test's capability to correctly detect stunting when it is present. Higher sensitivity means that the test has a lower rate of false negatives. Sensitivity varies with disease prevalence (Maxim, Niebo, & Utell, 2014)Specificity refers to the test's competence to accurately identify individuals who are not stunted (true negatives). By dividing the count of accurately identified negative cases by the sum of true negatives and false positives, specificity can be figured out. Mathematically it can be indicated as: Specificity = TN / (TN + FP). Specificity is a metric of the test's proficiency to precisely rule out stunting when it is not present. Higher specificity means that the test has a lower rate of false positives. Like sensitivity, specificity is not independent of prevalence (Maxim, Niebo, & Utell, 2014)

Negative Predictive Value (NPV) is the likelihood that a child who is predicted not to be stunted is not stunted. It is determined as the ratio of true negatives to the sum of those accurately

identified negatives and false negatives. NPV can be calculated using the following formula: NPV = TN / (TN + FN). NPV depends on both sensitivity and specificity, as well as prevalence. As prevalence increases, the NPV decreases because there is a higher chance of false negatives, irrespective of whether the test exhibits high sensitivity and specificity. Positive Predictive Value (PPV) is the chance that a child who is predicted to be stunted is stunted. It is determined by dividing the count of true positives by the summation of true positives and false positives. The formula is given by: PPV = TP / (TP + FP). PPV also depends on sensitivity, specificity, and prevalence. The PPV increases as prevalence increases because there is a higher chance of true positives, even if the test has the same sensitivity and specificity.

Sensitivity and specificity play a crucial role in prediction because they directly reflect the performance of a diagnostic or predictive test. They offer information about the precision of the test in correctly identifying individuals with or without a particular condition or outcome. However, the frequency (prevalence) of the disease condition in the population being tested affects Sensitivity, specificity, PPV and NPV.

High sensitivity is most useful in situations where the consequences of a false negative result are significant. For example, in disease screening and infectious disease testing where it is important to identify as many true positive cases as possible to ensure early detection and intervention. On the other hand, high specificity is most useful in situations where the consequences of a false positive result are significant. For example, in confirmatory Tests.

It is important to note that striking the desired equilibrium between sensitivity and specificity depends on the specific context and potential consequences of false positives and false negatives. The appropriate choice of sensitivity or specificity is influenced by the objectives of the test, the prevalence of the condition, the availability of follow-up tests, and the potential risks associated with false results.

2.7.3. Likelihood Ratio

Likelihood ratios (LR) are statistical measures utilized to evaluate the diagnostic or prognostic value of an examination result. They provide information about how much a positive or negative

test result changes the odds of having a disease or experiencing a particular outcome. There are two types of likelihood ratios: the positive likelihood ratio (LR+) and the negative likelihood ratio (LR-). The positive Likelihood Ratio (LR+) refers to the ratio of the chance of acquiring a positive test result in children who are stunted to the likelihood of acquiring a positive test result in children who are not stunted. Mathematically, LR+ is determined as the quotient of sensitivity of the test and one minus the specificity of the test. Thus LR+ = Sensitivity / (1 – Specificity).

The LR+ indicates how much the chances of being stunted are raised given a positive test result is obtained. An LR+ larger than one suggests a correlation between a positive test result and an increased likelihood of being stunted. A stronger association is indicated by higher LR+. Generally, LR+ values above ten are considered strong evidence for ruling in the disease (stunting), while values below one suggest a weak association or a test result that has a higher chance of being a false positive.

The term Negative Likelihood Ratio (LR-) is given to the ratio of the odds of getting a negative test result in children who are stunted to the odds of having a negative test result (not stunted) in children who are not stunted. Mathematically, the following formula calculates LR-. LR- = (1 - Sensitivity) / Specificity. The LR- indicates how significantly the odds of being stunted are scaled down given that a negative test result is observed. An LR- lower than one denotes that a negative test result is linked to a diminished likelihood of having the disease. The lower the LR-, the stronger the association. LR- values closer to zero indicate a strong rule-out potential, while values above one suggest a weak association or a test result that has higher odds of being a false negative.

2.7.4. Brier Score

The Brier score is a quadratic principle that determines the squared differences $(Y-P)^2$ between true binary results (Y) and projections (P). It ranges from 0 to 0.25. Zero indicates a complete model and 0.25 denotes a non-informative model assuming a 50 per cent occurrence of the disease condition. When the occurrence of the disease condition is less frequent, the highest possible mark for a non-informative model is reduced (Steyerberg & Vergouwe, 2014). The

study conducted by Kantidakis, and others used the Brier score to compare Cox models and machine learning techniques (Kantidakis, et al., 2020).

2.7.5. Calibration

Calibration refers to the precision of risk approximates, specifically the concurrence between the predicted and recorded counts of events (Van Calster, McLernon, van Smeden, Wynants, & Steyerberg, 2019). Calibration is typically assessed graphically as the plot of predicted probability versus observed proportion. The x-axis of the graph represents the predictions, while the y-axis represents the outcome. The ideal prediction would align perfectly with the 45-degree line on the graph. In the case of a binary outcome, the y-axis of the plot includes values of 0 and 1. In research implemented by Dhillon et al., in 2016, calibration was employed to project the likelihood of having a live birth for women undergoing in vitro fertilization (IVF) (Dhillon, et al., 2016).

2.*7.6* Model Testing and Evaluation.

After randomly dividing the dataset into two a training set and a test set, typically using an 80/20 split, the optimal model parameters are adjusted using the training set. To prevent overfitting, the model is evaluated on a separate test set that was not exposed to the models during the training process. The efficiency of the model on the test data set is assessed by generating ROC curves and calculating the corresponding AUC. The AUROC serves as an indicator of the model's proficiency to distinguish or classify disease outcomes. When constructing the ROC curve, the true positive rate (TPR) is compared to the false positive rate at different thresholds. The model's performance is assessed by utilizing the AUROC. The AUC ranges from 0.50 to 1. Values close to 1 indicate stronger classifying capability. A model with a value of 1 represents a perfect classifier. An excellent model has values ranging from 0.90 to 0.99, a range of 0.80 to 0.89 is considered a good classifier while 0.70 to 0.79 is a fair model but 0.50 to 0.69 denote a poor predictive ability. When the curve is diagonal (AUC= 0.50), the model is said to be a random classifier meaning that the classification is by chance. To conduct a thorough evaluation of model performance, the sensitivity, PPV, and NPV are all considered. The model that attains the highest average performance metric (AUROC) is deemed the optimal predictive model for

stunting.

2. *9.* Prediction Score

Scoring refers to the process of generating predictions using a predictive model. Scoring necessitates three components: The first requirement is a predictive model, which is a mathematical approach represented by f (x, β). It combines predictor variable values (x) with specific quantities (β), known as model parameters, to generate predicted values for the target or response variable. Secondly, it is necessary to have specific values for the forecaster variables, usually from new data that the model had not seen. Lastly, specific values of the parameters are also needed. In general, the prediction score would be generated by

$$P_{stunting} = 1/ (1 + \exp (-y)) \qquad (18)$$

Where: p is the probability, exp is the natural number, and y is the logistic equation expressed as $\beta_0 + \beta_1 X_1 + \beta_2 X_2 + ...\beta_i X_i$ in which β_0 is the constant, β and X are vectors of the parameter and predictor variables respectively.

A logistic regression model utilizes a logit link function, which is used to transform the linear predictor into a predicted probability for every category or value of the dependent variable, as shown in the equation provided. The predicted response for each observation is determined by selecting the response level with the highest predicted probability. If the probability is below 0.5, the predicted response is assigned as 0 (not stunted). If the probability is 0.5 or higher, the predicted response is assigned as 1 (indicating stunted) Equation 18 can be used to compute a prediction probability of being stunted manually given the attributes of the child by using equation 19 below.

$$P = \frac{1}{1+\exp(-(\beta_0 + \beta_1 X_1 + \beta_2 X + ... + \beta_n X_n))} \qquad (19)$$

2.10. Sample Size for Model Development and Validation

Determining the adequate sample size for model development and validation depends on various factors, including the complexity of the problem, the available data, and the desired level of

statistical significance. While there is no one-size-fits-all answer, literature has proposed at least 10 events per feature (EPP). Others have loosened up the 10 EPP rule to 5 EPP dependent on the type of the model such as logistic or Cox regression (Baeza-Delgado, et al., 2022)

CHAPTER 3

METHODOLOGY

3.1. Application to data for predictors of child stunting

3.1.1. Data Sources

The data utilized in this research is derived from the 2015-16 Malawi Demographic and Health Survey (MDHS), which is a survey conducted at a national level to ensure representativeness. The MDHS collected up-to-date information on mothers' demographics and health information on child nutrition. The DHS, which was conducted by the National Statistics Office collected anthropometric data for the under-five children in selected households. The analysis focused on a dataset consisting of 5149 children who were included in the study due to their stunting outcome. For more information on the sampling procedure of the DHS, one can refer to the 2015-16 MDHS report to obtain specific details (National Statistical Office, 2017).

3.2. Variables

The potential predictors of stunting in this research were chosen according to existing research studies about predictors or determinants of stunting that were conducted in Sub-Saharan Africa and LMICs. The main predictors were categorized into demographic, economic, child-caring practices, obstetric, and other maternal factors. The response variable of this study was stunting and was calculated based on the anthropometric indicator (height-for-age) among under-five children. The growth standards that were released by the World Health Organization in 2006 were employed to compute the height-for-age index of children (WHO, 2006). The height-for-age index acts as an indicator for both stunted linear growth and the aggregate effects of growth shortfalls in children. Stunting is characterized as a state in which children have a height-for-age Z-score that is lower than two standard deviations (-2SD) from the median of the reference population confirmed by the World Health Organization (Akombi, Agho, Astell-Burt, Hall, & Renzaho, 2017). In this context, the z-score is determined by taking the difference between an individual's height at a given age and the median height of the comparative population, and then dividing it by the standard variations of the cited population at that exact age or height (WHO, 2006). The response variable was defined as a binary variable having the following levels, category 1(stunted < -2SD) and category 0 (not stunted > -2SD).

3.3. Selection of Candidate Predictors

The systematic review produced 68 predictor variables of child stunting, of which 67 were available from the 2016 MDHS dataset, and 27 had complete information. In this study, feature selection techniques, including forward selection, backward elimination, and stepwise selection, were employed. Additionally, the Least Absolute Shrinkage and Selection Operator (LASSO) and random forest techniques were utilized to identify significant variables from the list obtained from the MDHS-2015 dataset.

3.3.1. Automated Variable Selection Method (Backward, forward, and stepwise)

The study used AIC as a selection criterion to select relevant variables in these automated variable selection methods. A package called bootStepAIC in R software was also used to avoid overfitting.

3.3.2. LASSO (variable selection)

Furthermore, the LASSO binary logistic regression model was employed for variable selection. Specifically, features with nonzero values of the coefficients were chosen. The LASSO model utilized tenfold cross-validation with the smallest criteria to determine the best parameter (lambda) selection. By drawing an upright line at the value determined through tenfold cross-validation, the best lambda was identified, resulting in 22 variables with nonzero coefficients selected.

3.4. Model Development

The data extracted from the 2015-16 MDHS were used to develop and train different types of predictive models: Random Forest, LASSO regression and Logistic regression using different automated variable selection methods. R (version 17) software was used to conduct analysis and model development. The data was partitioned into a training set (80%) and a testing set (20%). The partitioning was done in such a way that the training dataset (80%) and the testing dataset (20%) had almost the same proportion of stunting. To ensure statistically significant outcomes and representative characteristics of the entire dataset, the research allocated a 20% portion for testing purposes, ensuring an adequate sample size. A training dataset of limited size can enlarge the variance of the model's parameter estimates, while a small testing dataset can lead to increased variance in the performance statistic of the model (Kohavi, 1995). Consequently, the division of data into an 80/20 split aims to minimize both variance values, guaranteeing their reduction to the lowest possible levels. To develop and refine the predictive models the study used the training dataset. The testing dataset was utilized to gauge the model’s accuracy and performance. The variables that were steadily learned to be substantial predictors of stunting in the articles that were reviewed were also selected to form a set of variables. This set of variables was determined by the researcher’s judgement. The selected predictors were then used to

develop different binary logistic regression models to predict stunting. These models were compared with each other for their discriminative ability and predictive performance.

CHAPTER 4

RESULTS

4.1 Results

4.1.1. Dependent variable

A total of 4976 under-five children were included in the study (table 4.1). The prevalence of stunting in the date set was 35%.

Table 2. Dependent variable

Variable	Frequency (n (%))
Stunting	
not stunted	3205(64.41)
Stunted	1771(35.59)

4.1.2. Independent variables

Tables 3, 4 and **5** present the frequency distribution of potential predictors of stunting in this research which were chosen according to existing research studies about predictors or determinants of stunting that were conducted in Sub-Saharan Africa and LMICs. The main predictors were classified into demographic, economic, child-caring practices, obstetric, and other maternal factors.

Table 3. Demographic variable

Predictor	**Frequency (n (%))**
Maternal age	
<20 yrs	340(6.83)
20-34 yrs	3709(74.54)
>=35 yrs	927(18.63)
Residence	
Rural	4172(83.84)
Urban	804(16.16)
Sex of household head	
Female	1301(26.23)
Male	3671(73.77)
age of household head	
<35 years	2740(55.06)
35+ years	2236(44.94)
sex of the child	
Female	2542(51.09)
Male	2434(48.91)

age of child(months)	
0-6 months	488(9.81)
6-18 months	1160(23.31)
Above 18 months	3328(66.88)
Body mass index	
<18.5	248(4.94)
18.5-24	3701(74.38)
>=25	1027(20.64)
Ethnicity	
Chewa	1647(33.10)
Tumbuka	682(13.71)
Lomwe	1205(24.22)
Ngoni	594(11.94)
Yao	693(13.93)
Other	155(3.11)
Maternal education	
No education	602(12.10)
Primary	3263(65.57)
Secondary and above	1111(22.33)
Region	
Northern	883(17.75)
Central	1759(35.35)
Southern	2334(46.91)
Number of under5 children in the household	
<=1 child	2386(47.95)
>=2 children	2590(52.05)
Marital status	
Single	157(3.16)
Ever married	4251(85.43)
Married	568(11.41)
Religion	
Protestant	1224(24.60)
Catholic	778(15.64)
Muslim	700(14.07)
Other religion	2474(45.70)
Family size	
Small	831(16.70)
Medium	2757(55.41)
Large	1388(27.89)

Table 4. Economic factors

wealth index	
Poorest	1086(21.82)
Poorer	1114(22.39)
Middle	978(19.65)
Richer	930(18.69)
Richest	868(17.44)
Occupation of mother	
Not working	1464(29.42)
Agricultural worker	2200(44.21)
Professional/technical/managerial	336(6.75)
Sales and services	262(5.27)
Domestic and unskilled manual	714(14.35)

Table 5. Obstetric, child morbidity and other maternal factors

Predictor	**Frequency (n (%))**
birth weight	
Low weight	870(17.48)
Normal weight	4106(82.52)
birth order number	
First-born	1237(24.86)
2nd -4th	2564(51.53)
5th or more	1175(23.61)
mode of delivery	
Caesarean	308(6.19)
Normal birth	4668(93.81)
Diarrhea episodes	
No	3937(79.12)
Yes	1039(20.88)
Anaemia level	
Not anemic	3424(68.81)
Anaemic	1552(31.19)
preceding birth interval	
no previous birth	2451(52.33)

<24 months	301(6.43)
>24 months	1932(41.25)
Place of delivery	
Home	376(7.56)
Facility	4600(92.44)
Type of birth	
Singleton	4820(96.86)
Multiple	156(3.14)
Delivery assistance	
Not health professional	390(7.84)
Health professional	4586(92.16)
Cough/fever	
No	3471(69.75)
Yes	1505(30.25)
distance to a health facility	
Short distance	2688(54.02)
Long distance	2288(45.98)

4.2. **Variables Selected by Automated Variable Selection Method (Backward, forward, and stepwise)**

The variables selected by backward, stepwise, and forward feature selection methods are presented in the table below.

Table 6. Variables selected by automated methods.

Backward	**Forward**	**Stepwise**
Age of child	Age of child	Age of child
Type of birth	Birth weight	Birth weight
Wealth index	Type of birth	Type of birth
Mother's BMI	Wealth index	Wealth index
Mother's education	Mother's BMI	Mother's BMI
Sex of the child	Ethnicity	Ethnicity
Number of under-five children	Sex of the child	Sex of the child
Diarrhea	maternal occupation	maternal occupation
Distance to a hospital	Distance to a hospital	Distance to a hospital
Household size	Location	Location
Delivery assistance	Diarrhea	Diarrhea
Age of household head	Number of under-5 children	Number of under-5 children

Marital status		

4.3. Variables Selected by Random Forest (Boruta)

Using Boruta, an algorithm designed specifically for random forests, 11 variables were selected from the 27 identified variables. These selected variables were the type of birth, age of the child, birth weight of the child, location, distance to facility, wealth index, birth order of the child, age of household head, body mass index of the mother and household size. The Boruta variable selection path is shown in **Figure 1** below. The confirmed important variables are the ones in green colour and those that are in red are the ones that are confirmed not to be important and in blue are shadow attributes. Shadow attributes in Boruta refer to a set of randomized or shuffled versions of the original attributes. These shadow attributes are created to serve as a benchmark for assessing the true importance of the original attributes.

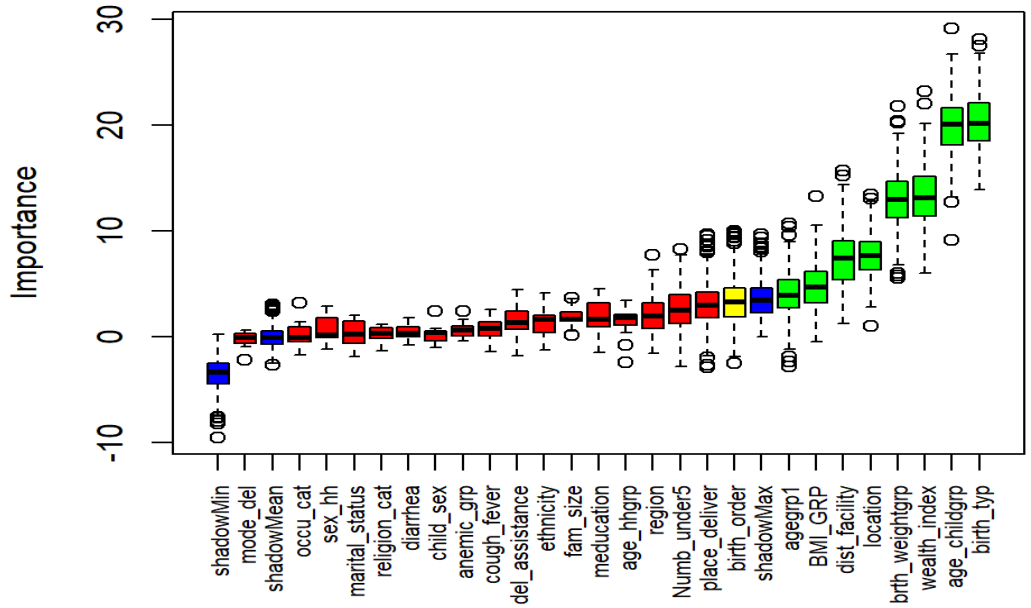

Figure 1: Selected variables: Random Forest, Boruta

(birth_typ =birth type, age_childgrp=age of child, wealth_index = wealth index, brth_weightgrp = child birth weight, location = location, dist_facility = distance to a health facility, birth_order = birth order, age_hhgrp = age of household head, BMI_GRP = body mass index, fam_size = family size, agegrp1 = mother's age, place_deliver =

place of delivery, Numb_under5=number of under five children, meducation = mother's education level, del_assistance = delivery assistance, marital_status = marital status, child_sex = sex of child, sex_hh = sex of household head, religion_cat = religion of mother, anemic_grp = anemia, mode_del = mode of delivery, occu_cat = mother's occupation, cough_fever = cough/fever)

4.4. Variables Selected by LASSO.

Twenty-two variables were selected by LASSO variable selection algorithm. These variables included location, wealth index, maternal age, age of household head, age of the child, household size, body mass index of the mother, distance to a health facility, number of under-5 children, religion, maternal education, type of birth, birth order of the child, region, diarrhea, maternal occupation, anaemia, delivery assistance, sex of the child, and sex of household head.

4.5. Variables Commonly Selected by All Variable Selection Methods.

The research identified factors that were commonly selected by all feature selection algorithms. These included the following: household wealth index, age of the child, household size, type of birth (singleton/multiple births) and birth weight.

4.6. Variables Determined by Judgement.

Using the researcher's judgement, ten variables were identified. The following were the factors that were identified, age of the child, the weight of the child at birth, type of birth, sex of the child, wealth index category of the household, number of under-five children in the household, location, family size, episode of diarrhea and maternal education.

4.7. Development of Prediction Models

All potential variables selected using the different variable selection algorithms were applied to develop different binary logistic regression models for predicting stunting**.** The models that included the above features were fitted and tabulated as shown in **Table 7.** below.

Table 7. Prediction factors for stunting

	Prediction Models											
intercept and variable	**Backward model**		**Forward model**		**Stepwise model**		**Random forest**		**LASSO model**		**Judgement model**	
	coeff	**p-value**	**coeff**	**p-value**	**coeff**	**p-value**	**coeff**	**p-value**	**coeff**	**p-value**	**coeff**	**p-value**
Intercept	-0.895	0.004	-0.802	0.009	-0.802	0.009	-0.698	0.010	-0.693	0.541	-0.734	0.001
Wealth index												
Poorest	ref		ref		ref		ref		ref		ref	
Poorer	-0.081	0.420	-0.084	0.407	-0.084	0.407	-0.117	0.244	-0.075	0.462	-0.098	0.327
Middle	-0.212	0.045	-0.210	0.046	-0.210	0.046	-0.245	0.019	-0.196	0.067	-0.223	0.034
Richer	-0.411	<0.001	-0.404	<0.001	-0.404	<0.001	-0.464	<0.001	-0.382	0.001	-0.419	<0.001
Richest	-0.544	<0.001	-0.508	<0.001	-0.508	<0.001	-0.626	<0.001	-0.458	0.002	-0.552	<0.001
Sex of child												
Male	ref		ref		ref		ref		ref			
Female	-0.175	0.011	-0.178	0.010	-0.178	0.010			-0.169	0.015	-0.178	0.009
Diarrhea												
No	ref		ref		ref		ref		ref			
Yes	0.157	0.087	0.156	0.072	0.156	0.072			0.169	0.053	0.165	0.056
Age of household head												
>35 years	ref		ref		ref		ref		ref			
35+ years							-0.034	0.702	0.019	0.831		

Number of under-five children												
<=1 child	ref		ref		ref		ref		ref			
>=2 children	0.192	0.013	0.187	0.016	0.187	0.016			0.192	0.016	0.162	0.033
Age of child												
0-6 months	ref		ref		ref		ref		ref			
6-18 months	0.486	0.001	0.489	0.001	0.489	0.001	0.532	<0.001	0.488	0.001	0.492	0.001
Above 18 months	1.035	>0.001	1.041	<0.001	1.041	<0.001	1.012	<0.001	1.036	<0.001	1.016	<0.001
Household size												
Small	ref		ref		ref		ref		ref			
Medium	0.044	0.692	0.049	0.656	0.049	0.656	0.139	0.195	0.054	0.639	-0.024	0.813
Large	-0.193	0.143	-0.186	0.159	-0.186	0.159	-0.063	0.631	-0.199	0.170	-0.205	0.083
Body mass index of the mother												
<18.5	ref		ref		ref		ref		ref			
18.5-24	-0.250	0.105	-0.247	0.109	-0.247	0.109	-0.233	0.128	-0.247	0.111		
>=25	-0.499	0.004	-0.489	0.004	-0.489	0.004	-0.496	0.004	-0.498	0.004		
Distance to health a facility												
Long distance	ref		ref		ref		ref		ref			
Short distance	0.156	0.032	0.160	0.028	0.160	0.028	0.137	0.057	0.175	0.017		
Delivery assistance												
Health personnel	ref		ref		ref		ref		ref			
Not health personnel									0.142	0.264		
Type of birth												
Singleton	ref		ref		ref		ref		ref			
Multiple	1.009	<0.001	1.029	<0.001	1.029	<0.001	1.073	<0.001	1.055	<0.001	1.015	<0.001
Maternal education												
No education	ref		ref		ref		ref		ref			
Primary									-0.162	0.139	-0.171	0.099
Secondary and above									-0.237	0.093	-0.279	0.034
Birth weight of the child												
Low birth weight	ref		ref		ref		ref		ref			
Normal birth weight	-0.552	<0.001	-0.557	<0.001	-0.557	<0.001	-0.524	<0.001	-0.546	<0.001	-0.554	<0.001
Location												
Urban	ref		ref		ref		ref		ref			
Rural	0.268	0.026	0.231	0.057	0.231	0.057					0.238	0.042
Mother's occupation												
Not working	ref		ref		ref		ref		ref			
Agricultural worker			0.028	0.740	0.028	0.740			0.013	0.881		
roof/technical/managerial			0.192	0.222	0.192	0.222			-0.187	0.240		
Sales and services			-0.384	0.032	-0.384	0.032			-0.373	0.038		

Domestic and unskilled			0.069	0.533	0.069	0.533			0.054	0.631		
Ethnicity												
Chewa	ref		ref		ref		ref		ref			
Tumbuka	0.043	0.795	0.048	0.769	0.048	0.769			0.034	0.836		
Lomwe	-0.340	0.005	-0.340	0.005	-0.340	0.005			-0.351	0.004		
Ngoni	0.027	0.817	0.035	0.766	0.035	0.766			0.031	0.791		
Yao	-0.179	0.152	-0.177	0.157	-0.177	0.157			-0.168	0.295		
Other	0.188	0.429	0.179	0.454	0.179	0.454			0.144	0.549		
Birth order												
First-born	ref		ref		ref		ref		ref			
2nd-4th	-0.210	0.021	-0.211	0.022	-0.211	0.022	-0.208	0.033	-0.221	0.026		
5th or above	0.053	0.633	0.050	0.660	0.050	0.660	0.009	0.949	-0.030	0.833		
Region												
North	ref		ref		ref		ref		ref			
Central	0.235	0.135	0.209	0.187	0.209	0.187			0.175	0.274		
South	0.320	0.054	0.295	0.075	0.295	0.075			0.265	0.114		
Sex of household head												
Male									ref			
Female									0.032	0.696		
Anaemia												
Not anaemic	ref											
Anaemic									-0.051	0.501		
Religion												
Protestant	ref											
Catholic									0.228	0.045		
Muslim									0.005	0.974		
Other									0.015	0.870		
Maternal age												
>20 years	ref											
20-34 years							-0.013	0.934	-0.021	0.895		
35-39 years							0.038	0.845	0.029	0.884		

Table 8. Prediction factors (factors identified by all variable selection methods) for stunting.

intercept and variable	common variable model	
	coefficient	p-value

Intercept	-0.705	0.004
Wealth index		
Poorest	ref	
Poorer	-0.127	0.201
Middle	-0.261	0.012
Richer	-0.522	<0.001
Richest	-0.808	<0.001
Age of child		
0-6 months	ref	
6-18 months	0.547	<0.001
Above 18 months	1.012	<0.001
Household size		
Small	ref	
Medium	0.061	0.521
Large	-0.071	0.507
Type of birth		
Singleton		
Multiple	1.068	<0.001
Birth weight of the child		
Low birth weight	ref	
Normal birth weight	-0.540	<0.001

4.8. Variable Importance

To help understand the results of the developed models, like in any other machine learning models, variable importance measures were computed. The graphs below present the overall variable importance of the models fitted using variables selected by the different variable selection methods.

4.8.1. Variable Importance for the Backward Model

As indicated in **Figure 2**, the age of a child has the largest contribution to the model seconded by the birth weight of the child. The type of birth, wealth index, ethnicity, body mass index of the mother, sex of the child, number of under-five children, birth order, and location are among the top ten important variables selected via the backward elimination method.

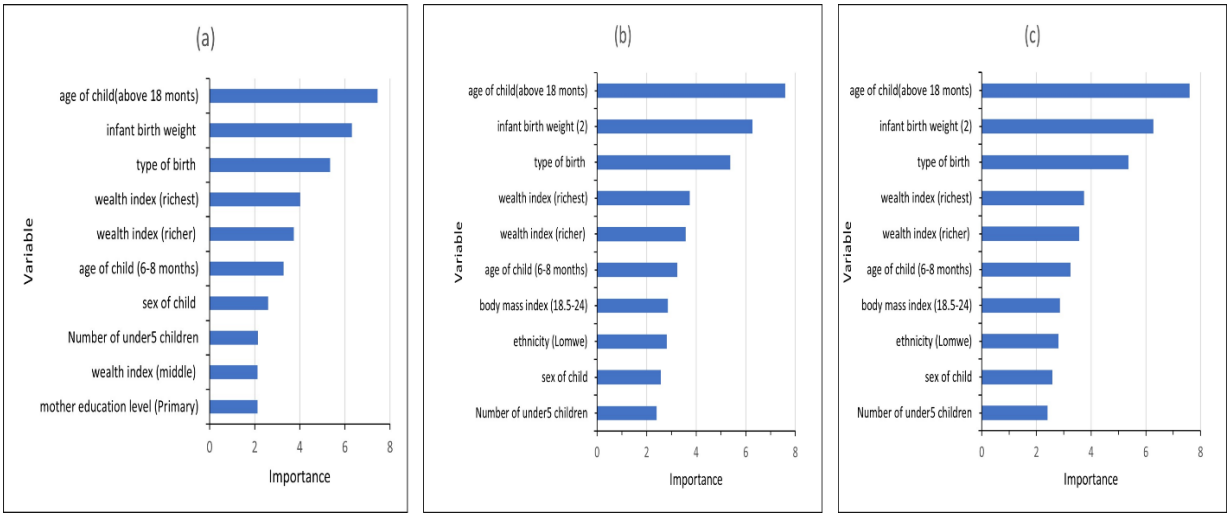

Figure 2 Overall variable importance for top 10 variables: (A) are variables selected by the backward algorithm, (B) variables selected by the forward algorithm and (C) variables selected by the stepwise algorithm.

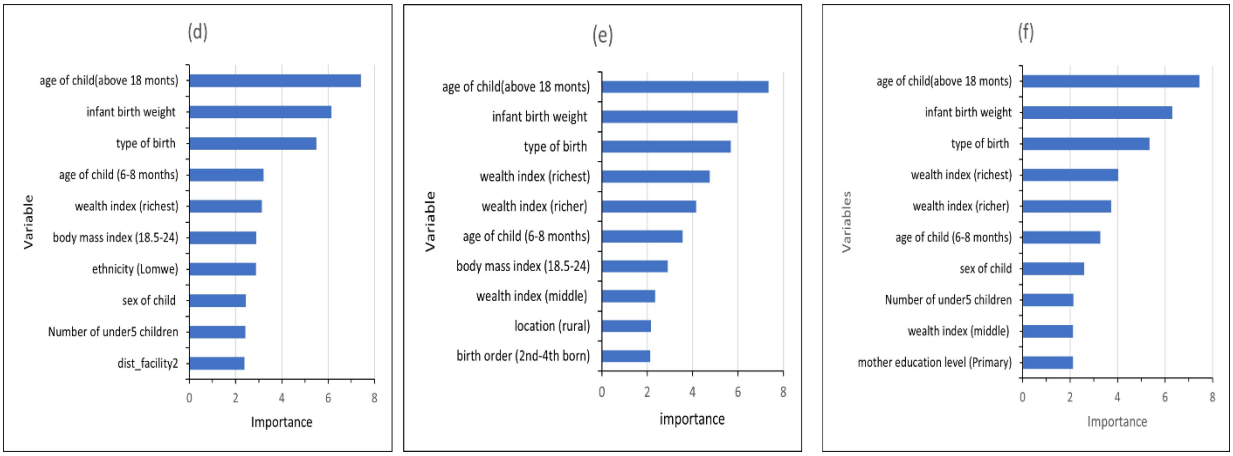

Figure 3: Overall variable importance for top 10 variables: (D) are variables selected by the LASSO algorithm, (E) variables selected by Random Forest algorithm and (F) variables selected by judgement.

4.8.2. Variable Importance for the Forward Model

Figure 2 (B) presents and ranks the variables that were selected using the forward variable selection algorithm. Like in backward variable selection the age of the child, birth weight, type of birth, wealth index, ethnicity, body mass index, sex of the child, number of under-five children, birth order and distance to health facility are the top ten important predictors.

4.8.3 Variable Importance for the Stepwise Model

The variables selected using the stepwise algorithm are identical to those selected via the forward variable selection algorithm. The graph in **Figure 2 (C)** ranks, the age of the child, birth weight, type of birth, wealth index, ethnicity, body mass index, sex of the child, number of under-five children, birth order and distance to a health facility as the top ten important predictors.

4.8.4. Variable Importance for the LASSO

The largest number of predictors were selected using LASSO variable selection algorithm. Twenty predictors were selected and using variable importance measures, the rankings are as presented in the graph, **Figure 3 (E)**. As indicated in this graph the best three variables are the age of the child, birth weight and type of birth just like the other algorithms above.

4.8.5. Variable Importance for the Random Forest

Using the predictors selected by the random forest-based algorithm, **Figure 3 (D)**, the ranking of variable importance measures is not very different from the rankings obtained in the other algorithms. The age of the child, birth weight of the child and type of birth are still ranked as the top three in terms of contribution to the predictive model. The Boruta algorithm has selected the fewest number of predictors, ten compared to the other algorithms.

4.8.6. Variable Importance for the Judgement Model

Another model was fitted using variables selected using the researcher's judgement**.** Ten predictors were incorporated into the model and were chosen based on how many articles among those that were reviewed identified them as statistically significant predictors of stunting. Importance scores of the variables were computed and presented in the graph, **Figure 3 (F)**.

4.9. Model Evaluation and Performance

The models were constructed using the selected predictors of stunting which were selected using backward elimination, forward selection, stepwise selection, random forest-based algorithm (Boruta), LASSO variable selection algorithm and judgment selection method. To further investigate the performance of the models obtained from different sets of variables, the study estimated cut-off points using the **SpEqualSe** method implemented in the **OptimalCutpoins** package in **R.** The cutoff points were estimated using the training data set**.** The analysis used the estimated cutoff points on the fitted logistic regression models to the test set and obtained several performance measures with each estimated cutoff point. Most often a default cut point of 0.5 is used in research studies ((Hasegawa, Ito, & Yamauchi, 2017); (Mukuku, et al., 2019); (van den Brink, et al., 2020)). For comparison purposes, the study also includes results obtained from using the cutoff point of 0.5 from the Bayes rule, **see Table 9**.

Table 9. Summary of probability score from the selected model (Judgement model).

Mean	Median	Range
0.36	0.35	0.74

Table 10. Model performance measures using a cut-off point of 0.5 on test data.

	Cut point	AUC (95% CI)	Sensitivity	Specificity	Misclassification error	Accuracy
Model 1	0.5	0.63(0.59-0.66)	0.34	0.81	0.35	0.65
Model 2	0.5	0.59(0.56 -0.63)	0.27	0.83	0.37	0.64
Model 3	0.5	0.59(0.56 -0.63)	0.27	0.83	0.37	0.64

Model 4	0.5	0.57(0.53 -0.61)	0.25	0.80	0.39	0.61
Model 5	0.5	0.62(0.59-0.66)	0.18	0.91	0.34	0.66
Model 6	0.5	0.64(0.60-0.67)	0.17	0.93	0.33	0.67
Model 7	0.5	0.63(0.59-0.66)	0.15	0.93	0.34	0.66

Model 1 was constructed using the variables selected by backward variable selection algorithms. Model 2 was constructed using the variables selected by forward variable selection algorithms. Model 3 was constructed using the variables selected by stepwise variable selection algorithms. Model 4 was constructed using the variables selected by random forest variable selection algorithms. Model 5 was constructed using the variables selected by LASSO variable selection algorithms. Model 6 was constructed using the variables selected by judgement. Model 7 was constructed using the variables that were common to the 6 models.

In the study, the performance estimates based on the cutoff point of 0.5 do well in terms of specificity, but all have poor sensitivity (**Table 10.**). The estimated cutoff points for each model seek an equilibrium between sensitivity and specificity. This is important because in this study the researchers were more interested in detecting a stunted child than finding a non-stunted child. Hence more attention is paid to sensitivity than specificity.

Training a model is the first step in making good predictions, however identifying how well the predictive power is, is a different question. To conclude if our trained model has good predictive power, the research simply used the trained model and predicted the response for the test data. These predictions were then used to compare with the true response variable. As expected, the models generally performed well when tested against the training dataset, simply because the error is underestimated by using the data that the model has seen as depicted in Appendix 1. The model fitted using the variables selected by the LASSO method has a better performance compared to other models, AUC of 67% (95% CI: 65-69). However, the true performance of the selected models was eventually determined by using the data (test data) that the trained model had not seen. The ROC curves were obtained by sensitivity versus the 1-specificity. The AUC results using the testing data set are tabulated in **Table 11**. Below

Table 11. Model performance measures using estimated cutoff points on test data.

	Cut point	AUC (95% CI)	Sensitivity	Specificity	Misclassification error	Accuracy
Model 1	0.36	0.63(0.59-0.66)	0.68 (0.62 - 0.73)	0.48 (0.44 - 0.52)	0.45	0.55
Model 2	0.36	0.59(0.56 -0.63)	0.66 (0.60 - 0.71)	0.48 (0.44 - 0.51)	0.46	0.54

Model 3	0.36	0.59(0.56 -0.63)	0.66 (0.60 - 0.71)	0.48 (0.44 - 0.51)	0.46	0.54
Model 4	0.37	0.57(0.53 -0.61)	0.68 (0.62 - 0.73)	0.45 (0.41- 0.48)	0.48	0.53
Model 5	0.37	0.62(0.59-0.66)	0.52 (0.47- 0.58)	0.64 (0.60 - 0.68)	0.4	0.6
Model 6	0.37	0.64(0.60-0.67)	0.61 (0.56 - 0.66)	0.60 (0.56 - 0.63)	0.4	0.6
Model 7	0.37	0.63(0.59-0.66)	0.63 (0.58 - 0.68)	0.55- (0.51- 0.59)	0.42	0.58

Model 1 was constructed using the variables selected by backward variable selection algorithms. Model 2 was constructed using the variables selected by forward variable selection algorithms. Model 3 was constructed using the variables selected by stepwise variable selection algorithms. Model 4 was constructed using the variables selected by random forest variable selection algorithms. Model 5 was constructed using the variables selected by LASSO variable selection algorithms. Model 6 was constructed using the variables selected by judgement. Model 7 was constructed using the variables that were common to the 6 models.

From **Table 11.** above, the logistic model from each set of selected variables yields quite similar performance. Nevertheless, the final model is the one with the largest AUC or C-statistic, which is the model fitted using a set of variables determined by the judgement method with an AUC of 64% (95% CI: 60-67%), the accuracy of 60% and sensitivity of 61% and specificity of 60%. The sensitivity and specificity indicate that 61% of the children who had a stunting condition and 60% of the children who did not have a stunting condition were correctly classified by the model. The confusion matrix for the best model is presented in **Table 12** below.

Table 12. Confusion Matrix indicate the performance of the best model at the selected probability cut point.

		Predicted		
		Not stunted	Stunted	All
Actual	Not stunted	385	261	646
	stunted	132	208	340
	All	517	469	986

The Confusion matrix indicates the performance of a classifier on stunted children and those who

are not stunted. The confusion matrix depends on the choice of the probability cutoff point. The research used the best-selected probability cutoff point of 0.37. The model is slightly better at predicting stunted class with a recall value of 0.61.

Table 13. Performance of the selected model after adjusting for sex and residence

	Sex of a child		Residence	
	Female	Male	Urban	Rural
AUC (95% CI)	0.64 (0.59-0.70)	0.63 (0.58 -0.68)	0.67(0.58-0.76)	0.63 (0.59 - 0.67)
Sensitivity (%)	0.86	0.79	0.62	0.7
Specificity (%)	0.26	0.27	0.59	0.45

The model based on risk factors determined by judgement has shown to be a predictive tool that displays a good ability to discriminate between stunted children and non-stunted children, particularly in children dwelling in urban areas (AUC=67% (95% CI: 58-76% in children dwelling in urban versus AUC=63% (95% CI: 59-67 in children dwelling in rural areas). Accordingly, children dwelling in urban areas, with a probability value higher than or equal to 0.37 were identified as stunted. Using this cut-off point, 62% of children residing in urban areas who were stunted and 59% of children residing in urban locations who were not stunted, were correctly classified. There is a small difference in the model's capacity to classify between stunted children and non-stunted children when gender was considered (AUC=64% (95% CI: 59-70%) in female children versus (AUC=63% (95% CI:58-68%) in male children.

The findings of this study show that the six prediction models have a better discrimination ability compared to a random classifier as indicated by the ROC curves in **Figure 4.** below.

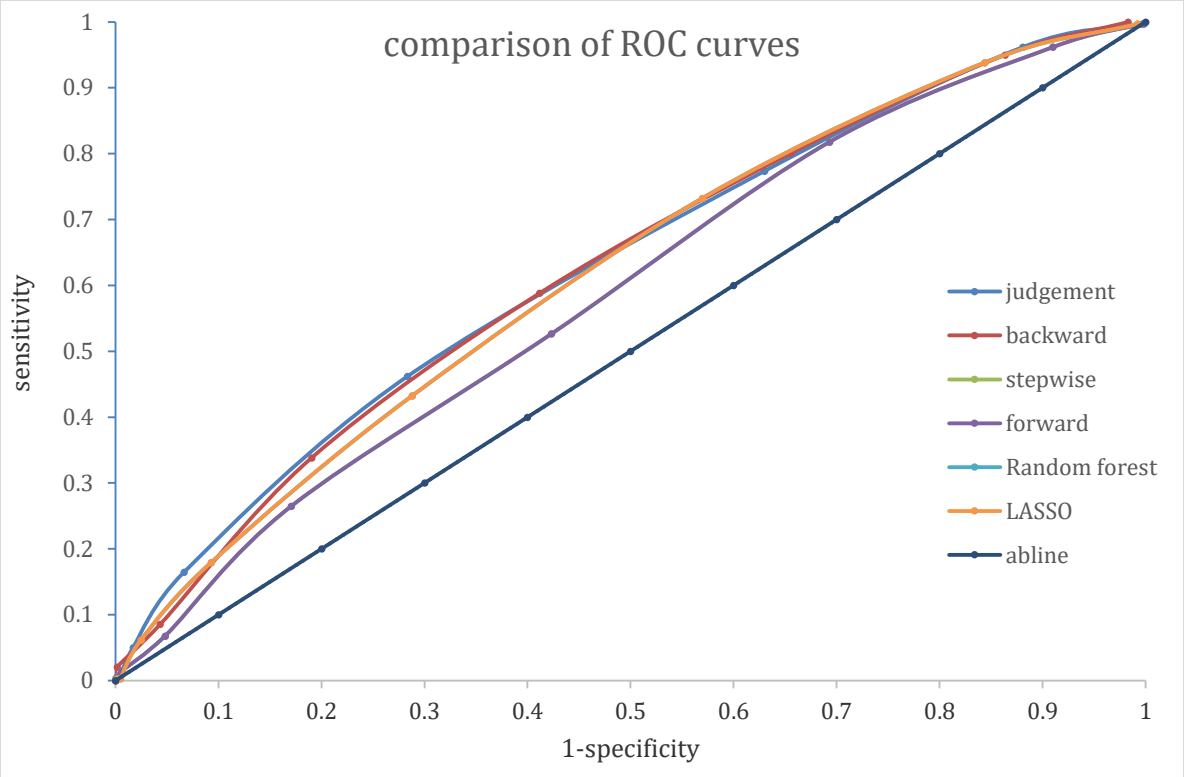

Figure 4. Comparing discrimination of the models fitted using variables selected by the different methods.

CHAPTER 5

DISCUSSION, CONCLUSION AND RECOMMENDATIONS

5.1. Discussion

The primary objective of the study was to create and validate a child stunting prediction score

based on the best predictive model in Malawi. A child stunting prediction score may help decision-makers implement tailor-made interventions to help in achieving a reduction of stunting in Malawi. It was based on predictor variables obtained from fitting a multivariate logistic regression model to child stunting applying data obtained from the 2015-16 MDHS. Using six variable selection methods, namely backward, forward, stepwise, random forest, LASSO, and judgment, the study identified nine easily measured key predictors of child stunting. These included the age of the child, the weight of the child at birth, type of birth, sex of the child, wealth index category of the household, number of under-five children in the household, location, and maternal education. The best predictive model was based on risk factors determined by judgment methods, which had AUROC of 65% (95% CI: 64%-67%) and 64% (95% CI: 60%-67%) in the training data set and the testing data set, respectively. Based on the common risk factors, identified by all the feature selection algorithms, the predictive ability was 62% (95% CI: 59.0%-66.0%). For children residing in urban areas, the AUROC was 67% (95% CI: 58-76%) while for the children living in rural areas, AUROC was 63% (95% CI: 59-67).

Although many studies have identified determinants of stunting, a sparse number of studies have focused on the explicit creation and evaluation of risk prediction models designed to detect children at a high risk of stunting. One such study was carried out by Hasegawa et Al., 2017, in Zambia. The predictive tool that the study developed was aimed at predicting malnourishment in young children. It used maternal age, weight-for-age z-score status, birth weight, feeding status, history of sibling death, multiple births, and maternal education level as important predictors. However, their study differs from this study in that the tool was developed using data collected from one health facility in rural location which limits its generalizability. The Lives Saved Tool has been in use to approximate the influence of specific modifications in important interventions on the decrease of stunting in children less than five years old. Many interventions that influence stunting, both directly and indirectly, have been identified by this tool. Zinc supplementation, education on suitable complementary feeding, and giving supplementary food are some of the interventions that are included in this tool (Hanieh, et al., 2019). Nevertheless, this study's predictive model distinguishes itself from the Lives Saved Tool by its inability to predict which children are likely to be stunted. In the investigation implemented by Hanieh et al. (2019), they constructed and outwardly verified a predictive model to forecast the hazard of stunting when preschool children reach 3 years of age. Their final model contained very important predictors,

which consisted of the height of both the father and mother, the weekly weight gain of the mother during pregnancy, the sex of the infant, the gestational age at birth, as well as the weight and length of the infant at 6 months of age. This study differed in the sense that the model was specifically developed and validated within a rural context, excluding urban areas and other regions. In addition, the model had limitations in predicting all children who were vulnerable to stunting, thereby constraining its utility among children aged over 3 years.

The method of choosing the probability cut points influences the predictive ability of the model. The cost of wrongly classifying those that have the disease (false negative) and those that do not have the disease (false positive) has informed the choice of the appropriate method of choosing the cut points for the classifier's scores (Ferri, Orallo, & Flach, 2019). Sometimes it is important to choose a method that gives high sensitivity and specificity (Bewick, Cheek, & Ball, 2004). During the selection of probability thresholds, a trade-off is made between the false positive rate (FPR) and the false negative rate (FNR). This is perceived as the objective function of the model, wherein the target is to reduce the number of errors, or the cost incurred. In general, there is a tradeoff between specificity and sensitivity, and a decision must be made based on their relative importance (Bewick, Cheek, & Ball, 2004). The method of selecting the cut points must take this into account. It is important to evaluate the efficiency of a model at different cut points but assessing the model at its optimal cut point is also desirable. The ROC curves presented in Figure 4.4 demonstrate the tradeoff between the two measures at various cut points in this study. The metrics were calculated for assessing the performance of the predictive model using cut-points derived based on the **SpEqualSe** method implemented in the **OptimalCutpoins** package in **R.** This method hinges on the principles of balancing sensitivity and specificity with the assumption that the expenses associated with false positives and false negatives are of equal value. This is one of the data-driven methods of choosing optimal cut points and their use in studies with small sample sizes may identify accurate optimal cut points and overstate accuracy estimates (Bhandari, et al., 2021). However, this study used a big sample size which might have avoided what Bhandari observed. There are other data-driven methods of selecting cut points that can also be used in choosing the optimal thresholds such as the Youden's Index (J) (Lai, Tian, & Schisterman). Xu et al.,2019 used Youden's Index to decide the thresholds for predicting AKI in their model. The Youden's Index (J) is defined as the sum of Sensitivity and Specificity minus one ($J_c = SE_c + SP_c - 1$). The method (**SpEqualSe)** used in this study does not differ from

Youden's Index since both use criteria built on sensitivity and specificity measures (López-Ratón, Rodríguez-Álvarez, Cadarso-Suárez, & Gude-Sampedro, 2014). Incorporating predetermined cut points, when accessible, would enhance the credibility of a classification model (Ewald, 2006). These pre-specified cut points are the ones that are predetermined by using previous studies.

There were certain limitations in the study that are worth mentioning and discussing. To begin with, the study encountered missing values for certain potential predictors, including nutritional variables, which could introduce selection bias. In addition, the study did not consider the clustering and weighting of the DHS data, which may have affected the estimated probabilities of being stunted by not being representative. The data used for score development and validation were exclusively obtained from Malawi, which may potentially restrict the generalizability of the risk score to other regions within Sub-Saharan Africa (SSA). This study did not consider LARS due to the unavailability of the software package in R to implement it as a variable selection method for nonlinear models. Another limitation is that the study assumed that a logistic link function will provide a better predictive model. Alternatively, a probit link function or complementary log-log link function could have been employed. The method of selecting the probability cut points that were used in the study is also another limitation. There are other robust approaches for selecting decision boundaries that could have been used.

The effectiveness of this study's approach is rooted in the fact that using nationally representative data (MDHS), the study investigated an extensive array of potential predictors and successfully identified a concise collection of major variables. These variables are commonly assessed in primary healthcare settings in numerous countries or can be readily obtained. Even though factors affecting stunting that have been reported in the literature vary by many attributes such as type of study, region and sample size, and the ones mentioned above, considerable key findings have surfaced that offer support for the predictive variables that the model has identified. However, this might have affected the derived scores since some variables were not captured in MDHS and some had missing values and as such, they were not used in developing the predictive model.

The discriminative ability of this study's model seems to be different depending on other

characteristics such as the residence type of the child. The change in the performance of the model between urban and rural populations is largely due to differences in sample sizes of these two categories in the data set.

The study proposes a similar research area but uses simulated data with sensitivity analysis to improve the predictive ability. Future investigations should also strive to replicate the findings of this study by employing alternative machine learning algorithms for binary classification. There is also a need to explore other variable selection methods such as LARS.

5.2. Conclusions

Though the various selected predictors and models had an unsatisfactory performance at distinguishing between stunted and non-stunted children, this work has shown the potential of using a tool that combines purported child stunting predictors. The study's approach offers a direct estimate of child stunting using a single summary measure, rather than working with multiple predictors of child stunting.

The findings of the systematic review have shown that determinants of stunting are multifaceted and interdependent. The study has identified many predictors of stunting, but the dominant ones are the weight of the child at birth, type of birth, sex of the child, wealth index category of the household, number of under-five children in the household, location, maternal education, family size, diarrhea, birth order, distance to facility and body mass index.

The prediction model shows that the predictors of stunting for children are the weight of the child at birth, type of birth, sex of the child, wealth index category of the household, number of under-five children in the household, location, and maternal education. The precision of the scoring system in predicting the likelihood of children under the age of five experiencing stunting in Malawi was 60% with a sensitivity of 61%, specificity of 60% and AUC of 64% (95% CI: 60-67%). Stunting cases occur usually because this disease is not recognized by the public at an early stage. The research has developed a prediction model that has the possibility of helping in understanding what may influence child stunting and may help policymakers to focus on evidence-based interventions that target specific predictors in low-resource countries. The researchers believe that the predictive model will empower public health practitioners at the

community level or in hospitals to quickly measure the projected subsequent likelihood of stunting in under-five children. By employing early protective measures during the critical developmental stage of the first five years of life, there is a chance to intervene and alter the growth path before it becomes irreversible. This approach permits timely action within the optimal window of opportunity, where the most significant impact is expected to be attained.

5.3 Recommendations

The risk predictive model for child stunting is recommended for children aged 0-59 years in Malawi and similar settings in sub-Saharan Africa. It is necessary to embrace a comprehensive community-oriented strategy that addresses the instantaneous and fundamental factors contributing to child malnutrition. This strategy should encompass counselling phases for mothers to enhance infant feeding behaviours and maternal dietary intake as well as health promotion initiatives to raise awareness about the significance of appropriate public health measures for cleanliness and hygiene.

REFERENCES

Aguayo, V. M., Nair, R., Badgaiyan, N., & Krishna, V. (2016). Determinants of stunting and poor linear growth in children under 2 years of age in India: an in-depth analysis of Maharashtra's comprehensive nutrition survey. *Maternal & child nutrition*, 121–140. doi:https://doi.org/10.1111/mcn.12259

Bhandari, P. M., Lewis, B., Neupane, D., Patten, S. B., Shrier, I., & Thombs, B. D. (2021). Data-driven methods distort optimal cutoffs and accurate estimates of depression screening tools: a simulation study using individual participant data. . *Journal of clinical epidemiology, 137*, 137–147. doi:. https://doi.org/10.1016/j.jclinepi.2021.03.031

Dake, S. K., Solomon, F. B., Bobe, T. M., Tekle, H. A., & Tufa, E. G. (2019). Predictors of stunting among children 6–59 months of age in Sodo Zuria District, South Ethiopia: a community-based cross-sectional study. . *BMC Nutr, 5*(23). doi:https://doi.org/10.1186/s40795-019-0287-6

Khan, J. R., Tomal, J. H., & Raheem, E. (2021). Model and variable selection using machine learning methods with applications to childhood stunting in Bangladesh. *Informatics for Health and Social Care, 46*(4), 425-442.

Steyerberg, E. W., & Vergouwe, Y. (2014). Towards better clinical prediction models: seven steps for development and an ABCD for validation. *European heart journal, 35*(29), 1925–1931. doi:https://doi.org/10.1093/eurheartj/ehu207

Afolabi, F., & Palamuleni, M. E. (2019). Determinants of Stunting among Under-five Children in Malawi. African Population Conference. doi:http://uaps2019.popconf.org

Ahmed, S. M., Brintz , B. J., Pavlinac, P. B., Shahrin, L., Huq , S., Levine , A. C., . . . Leung, D. T. (2023). Derivation and external validation of clinical prediction rules identifying children at risk of linear growth faltering . *Elife, 12*.

Baeza-Delgado, C., Alberich, L. C., Carot-Sierra, J. M., Veiga-Canuto, D., de las Heras, B. M., Raza, B., & Martí-Bonmatí, L. (2022). A practical solution to estimate the sample size required for clinical prediction models generated from observational research on data. 6,. *Eur Radiol Exp, 6*. doi:https://doi.org/10.1186/s41747-022-00276-y

Berhe, K., Seid, O., Gebremariam, Y., Berhe, A., & Etsay, N. (2019). Risk factors of stunting (chronic undernutrition) of children aged 6 to 24 months in Mekelle City, Tigray Region, North Ethiopia: An unmatched case-control study. . *PLoS ONE, 14*(6). doi:https://doi.org/10.1371/journal.pone.0217736

Bewick, V., Cheek, L., & Ball, J. (2004). Statistics review 13: receiver operating characteristic curves. *Critical care, 8*(6), 508–512. doi:https://doi.org/10.1186/cc3000

Bukusuba, J., Kaaya, A. N., & Atukwase, A. (2017). Predictors of Stunting in Children Aged 6 to 59 Months: A Case-Control Study in Southwest Uganda. *Food and nutrition bulletin, 38*(4), 542–553. doi:https://doi.org/10.1177/0379572117731666

Chirande, L., Charwe, D., Mbwana, H., Victor, R., Kimboka, S., Issaka, A. I., . . . Dibley, M. J. (2015). Determinants of stunting and severe stunting among under-fives in Tanzania: evidence from the 2010 cross-sectional household survey. *BMC Pediatr, 15*, 165. doi:https://doi.org/10.1186/s12887-015-0482-9

Fikadu , T., Assegid, S., & Dube, L. (2014). Factors associated with stunting among children of age 24 to 59 months in Meskan district, Gurage Zone, South Ethiopia: a case-control study. . *BMC Public Health, 14*. doi:https://doi.org/10.1186/1471-2458-14-800

Gebru , K. F., Haileselassie, W. M., Temesgen, A. H., Seid, A. O., & Mulugeta, A. (2019). Determinants of stunting among under-five children in Ethiopia: a multilevel mixed-effects analysis of 2016 Ethiopian demographic and health survey data. *BMC Pediatr, 19*. doi:https://doi.org/10.1186/s12887-019-1545-0

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*, 115–130. doi:https://doi.org/10.2147/PHM

Haile , D., Azage , M., Mola , T., & Rainey, R. (2016). Exploring spatial variations and factors associated with childhood stunting in Ethiopia: a spatial and multilevel analysis . *BMC pediatrics, 16*. doi:https://doi.org/10.1186/s12887-016-0587-9

Hanieh, S., Braat, S., Simpson, J. A., Ha, T. T., Tran, T. D., Tuan, T., . . . Biggs, B.-A. (2019). The Stunting Tool for Early Prevention: development and external validation of a novel tool to predict the risk of stunting in children at 3 years of age. *BMJ Global Health, 4*. doi:doi:10.1136/ bmjgh-2019-001801

Kantidakis, G., Putter, H., Lancia, C., de Boer, J., Braat, A. E., & Fiocco, M. (2020). Survival prediction models since liver transplantation - comparisons between Cox models and machine learning techniques. 20, 277. *BMC Med Res Methodol, 20*(277). doi:https://doi.org/10.1186/s12874-020-01153-1

Khiabani, F. B., Ramezankhani, A., Azizi, F., & Hadaegh, F. (2015). A tutorial on variable selection for clinical prediction models: Feature selection methods in data-mining could improve the results. *Journal of Clinical Epidemiology, 71*. doi:10.1016/j.jclinepi.2015.10.002

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*. doi:https://doi.org/10.1186/s12889-017-4621-0

Krasevec, J., An, X., Kumapley, R., Bégin, F., & Frongillo, E. A. (2017). Diet quality and risk of stunting among infants and young children in low- and middle-income countries. *Maternal & child nutrition, 13*. doi:https://doi.org/10.1111/mcn.12430

Liao, H., & Lynn, H. S. (2010). A Survey of Variable Selection Methods in Two Chinese Epidemiology Journals. . . *BMC medical research methodology*. doi:10.1186/1471-2288-10-87

López-Ratón, M., Rodríguez-Álvarez, M., Cadarso-Suárez, C., & Gude-Sampedro, F. (2014). OptimalCutpoints: An R Package for Selecting Optimal Cutpoints in Diagnostic Tests. *Journal of Statistical Software, 61*(8), 1–36. doi:https://doi.org/10.18637/jss.v061.i08

Maxim, L. D., Niebo, R., & Utell, M. J. (2014, Sep 29). Screening tests: a review with examples. *Inhal Toxicol, 26*(13), 811-28. doi: 10.3109/08958378.2014.955932

McDonald, C. M., Kupka, R., Manji, K. P., Okuma, J., Bosch, R. J., Aboud, S., . . . Duggan, C. P. (2012). Predictors of stunting, wasting and underweight among Tanzanian children born to HIV-infected women. . *Eur J Clin Nutr, 66*, 1265–1276. doi: https://doi.org/10.1038/ejcn.2012.136

Mehta, R., Suchdev, P., Rhodes, E., & Williams, A. (2018). Determinants of Stunting among Preschool Children in the 2015–2016 Malawi Micronutrient Survey. doi:10.1093/cdn/nzy039.

Mtambo , O., Masangwi, S., & Kazembe , L. (2014). Analysis of Childhood Stunting in Malawi Using Bayesian Structured Additive Quantile Regression Model. *International Journal of Statistics and Applications, 2014*. doi:10.5923/j.statistics.20140403.04

Mukuku, O., Mutombo, A. M., Kamona, L. K., Lubala, T. K., Mawaw, P. M., Aloni, M. N., . . . Luboya, O. N. (2019). Predictive Model for the Risk of Severe Acute Malnutrition in Children. *Journal of Nutrition and Metabolism, 2019*, 1-7. doi:10.1155/2019/4740825

Ndagijimana, S., Kabano, I. H., Masabo, E., & Ntaganda, J. M. (2023). Prediction of Stunting Among Under-5 Children in Rwanda Using Machine Learning Techniques. *J Prev Med Public Health, 56*(1), 41-49. doi:doi: 10.3961/jpmph.22.388

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

Nkurunziza , S., Meessen, B., Van geertruyden , J.-P., & Korachais, C. (2017). Determinants of stunting and severe stunting among Burundian children aged 6-23 months: evidence from a national cross-sectional household survey. *BMC pediatrics, 17*(1). doi:https://doi.org/10.1186/s12887-017-0929-2

Nsereko, E., Mukabutera, A., Iyakaremye, D., Umwungerimwiza, Y. D., Mbarushimana, V., & Nzayirambaho, M. (2018). Early feeding practices and stunting in Rwandan children: a cross-sectional study from the 2010 Rwanda demographic and health survey. *The Pan African medical journal, 29*. doi:https://doi.org/10.11604/pamj.2018.29.157.10151

Nshimyiryo , A., Hedt-Gauthier , B., Mutaganzwa , C., Kirk , C. M., Beck , K., Ndayisaba , A., . . . El-Khatib , Z. (2019). Risk factors for stunting among children under five years: a cross-sectional population-based study in Rwanda using the 2015 Demographic and Health Survey. . *BMC Public Health, 19*. doi:https://doi.org/10.1186/s12889-019-6504-z

Sema , B., Azage , M., & Tirfie, M. (2021). Childhood stunting and associated factors among irrigation and non-irrigation user northwest, Ethiopia: a comparative cross-sectional study. . *Ital J Pediatr, 47*. doi:https://doi.org/10.1186/s13052-021-01048-x

Shinsugi , C., Matsumura , M., Karama , M., Tanaka , J., Changoma , M., & Kaneko , S. (2015). Factors associated with stunting among children according to the level of food insecurity in the household: a cross-sectional study in a rural community of Southeastern Kenya. *BMC public health, 15*, 441.

Tariku , A., Biks , G. A., Derso , T., Wassie, M. M., & Abebe, S. M. (2017). Stunting and its determinant factors among children aged 6-59 months in Ethiopia. *Italian journal of pediatrics, 43*(1). doi:https://doi.org/10.1186/s13052-017-0433-1

Uwiringiyimana , V., Veldkamp, A., & Ocke, M. C. (2019). Predictors of stunting with particular focus on complementary feeding practices: A cross-sectional study in the northern province of Rwanda. *Nutrition, 60*, 11-18. doi:https://doi.org/10.1016/j.nut.2018.07.016

Van Calster, B., McLernon, D. J., van Smeden, M., Wynants, L., & Steyerberg, E. W. (2019). Calibration: the Achilles heel of predictive analytics. . *BMC Med, 17*(230).

Akombi, J. B., Agho, E. K., Astell-Burt, T., Hall, J. J., & Renzaho, A. M. (2017, Jan 13). Stunting and severe stunting among children under-5 years in Nigeria: A multilevel analysis. *BMC Pediatrics*. doi:10.1186/s12887-016-0770-z

Ayelign, A., & Zerf, T. (2021). Household, Dietary and Healthcare Factors Predicting Childhood Stunting in Ethiopia. *Heliyon*. doi:10.1016/j.heliyon

Chowdhury, M. Z., & Turin, T. C. (2020). Variable selection strategies and its importance in clinical prediction modelling. *Family medicine and community health, 8*(1). doi: https://doi.org/10.1136/fmch-2019-000262

Cruz, L. M., Azpaitia, G. G., Suarez, R. D., Rodriguez, A. S., & Ferrer, J. F. (2017). Factors Associated with Stunting among Children Aged 0 to 59 Months from the Central Region of Mozambique. *Nutrients, 9*(5). doi:https://doi.org/10.3390/nu9050491

Czepiel, S. A. (2002). Maximum Likelihood Estimation of Logistic Regression Models. doi:http://czep.net

Davis, S. E., Lasko, T. A., Chen, G., Siew, E. D., & Matheny, M. E. (2017). Calibration drift in regression and machine learning models for acute kidney injury. *Journal of the American Medical Informatics Association, 24*, 1052–1061.

Degenhardt, F., Seifert, S., & Szymczake, S. (2019, March). Evaluation of variable selection methods for random forests and omics data sets. *Briefings in Bioinformatics, 20*(2), 492–503. doi: https://doi.org/10.1093/bib/bbx124

Dhillon, R. K., Mclernon, D. J., Smith, P. P., Fishel, S., Dowell, K., Deeks, J. J., . . . Coomarasamy, A. (2016). Predicting the chance of live birth for women undergoing IVF: a novel pretreatment counselling tool 31(1), . *Human reproduction, 31*(1), 84–92. doi:https://doi.org/10.1093/humrep/dev268

Dorugade, A. V. (2014). New ridge parameters for ridge regression. . *Journal of the Association of Arab Universities for Basic and Applied Sciences, 15*(1), 94-99. doi:10.1016/j.jaubas.2013.03.005

Ewald, B. (2006, Aug). Post hoc choice of cut points introduced bias to diagnostic research. *J Clin Epidemiol, 59*(8), 798-801. doi: doi: 10.1016/j.jclinepi.2005.11.025

Ferri, C., Orallo, J. H., & Flach, P. (2019). Setting decision thresholds when operating conditions are uncertain. *Data Min Knowl Disc, 33*, 805–847. doi:https://doi.org/10.1007/s10618-019-00613-7

Hasegawa, J., Ito, Y., & Yamauchi, T. (2017). Development of a screening tool to predict malnutrition among children under two years old in Zambia. *Global health action, 10*(1). doi:https://doi.org/10.1080/16549716.2017.1339981

Heinze, G., Wallisch, C., & Dunkler, D. (2018). Variable selection - A review and recommendations for the practising statistician. Biometrical journal. Biometrische Zeitschrift. *60*(3), 431–449. doi:https://doi.org/10.1002/bimj.201700067

Keino, S., Ettyang, G. P., & Borne, B. V. (2014). Determinants of stunting and overweight among young children and adolescents in sub-Saharan Africa. *Food and nutrition bulletin, 35*(2), 167–178. doi:https://doi.org/10.1177/156482651403500203

Kofi, J. O. (2018). Predictors of childhood stunting in Ghana: A cross-sectional survey of the association between stunting among children under age five and maternal bio-demographic and socioeconomic characteristics in Ghana 2014 (Dissertation). doi: http://urn.k

Kohavi, R. (1995). Study of cross-validation and bootstrap for accuracy estimation and model selection. *In: Ijcai*.

Lai, C.-T., Tian, L., & Schisterman, E. (n.d.). Exact confidence interval estimation for the Youden index and its corresponding optimal cut-point. *Computational statistics & data analysis, 56*(5), 1103–1114. doi:https://doi.org/10.1016/j.csda.2010.11.023

Mtenda, P. A. (2019). Association of low birth weight with undernutrition in preschool-aged children in Malawi. *Nutrition, 18*.

National Statistical Office. (2017). *2015-16 Malawi Demographic & Health Survey.* Maryland: Rockville , USA.

Nemeth, C. J. (2014). Parameter estimation for state space models using sequential Monte Carlo algorithms.

Paul , P., Pennell, M., & Lemeshow, S. (2013). Standardizing the power of the Hosmer-Lemeshow goodness of fit test in large data sets. *Statistics in medicine, 32*. doi:10.1002/sim.5525

Predictors of stunting with particular focus on complementary feeding practices: A cross-sectional study in the northern province of Rwanda. (n.d.). *Nutrition (Burbank, Los Angeles County, Calif.), 60*, 11–18. doi: https://doi.org/10.1016/j.nut.2018.07.016

Ratner, B. (2010). Variable selection methods in regression: Ignorable problem, outing notable solution. *J Target Meas Anal Mark, 18*, 65–75.

Semali, A. (2015). Prevalence and determinants of stunting in under-five children in central Tanzania: remaining threats to achieving Millennium Development Goal 4. *BMC Public Health, 15*. doi:https://doi.org/10.1186/s12889-015-2507-6.

Stewart, C. P., Iannotti, L., Dewey, K. G., Michaelsen, F. K., & Onyango, A. W. (2013). Contextualizing complementary feeding in a broader framework for stunting prevention. . *Maternal & child nutrition, 9*, 27–45. doi:https://doi.org/10.1111/mcn.12088

van den Brink, D. A., de Meij, T., Brals, D., Bandsma, R. H., Thitiri, J., Ngari, M., . . . Voskuijl, W. P. (2020). Prediction of mortality in severe acute malnutrition in hospitalized children by faecal volatile organic compound analysis: proof of concept. *Sci Rep, 10*. doi: https://doi.org/10.1038/s41598-020-75515-6

Vu, N. U. (2022). *Childhood Stunting Prediction in Bangladesh A Machine Learning Approach (Doctoral dissertation, Tilburg University).*

Walter, S., & Tiemeier, H. (2009). Variable selection: current practice in epidemiological studies. . *European journal of epidemiology, 24*(12), 733–736.

WHO. (2006). *WHO Child Growth Standards based on length/height, weight and age.* WHO MULTICENTRE GROWTH REFERENCE STUDY GROUP.

Woldeamanuel, B. T., & Tesfaye, T. T. (2019). Risk Factors Associated with Under-Five Stunting, Wasting, and Underweight Based on Ethiopian Demographic Health Survey Datasets in Tigray Region, Ethiopia . *Journal of Nutrition and Metabolism, vol. 2019*, 11. doi:https://doi.org/10.1155/2019/6967170

APPENDICES

Appendix 1: Model performance measures using estimated cutoff points on the training set.

	Cut point	AUC (95% CI)	Sensitivity	Specificity	Misclassification error	Predictive value
Model 1	0.36	0.66(0.64-0.68)	0.62	0.62	0.39	0.61
Model 2	0.36	0.66(0.65-0.68)	0.62	0.62	0.38	0.62
Model 3	0.36	0.66(0.65-0.68)	0.62	0.62	0.38	0.62
Model 4	0.37	0.65(0.64-0.67)	0.61	0.61	0.39	0.61
Model 5	0.37	0.67(0.65-0.69)	0.63	0.63	0.37	0.63
Model 6	0.37	0.65(0.64-0.67)	0.61	0.60	0.39	0.61
Model 7	0.37	0.64(0.62-0.66)	0.6	0.61	0.39	0.61

Model 1 was constructed using the variables selected by backward variable selection algorithms. Model 2 was constructed using the variables selected by forward variable selection algorithms. Model 3 was constructed using the variables selected by stepwise variable selection algorithms. Model 4 was constructed using the variables selected by random forest variable selection algorithms. Model 5 was constructed using the variables selected by LASSO variable selection algorithms. Model 6 was constructed using the variables selected by judgement. Model 7 was constructed using the variables that were common to the 6 models.

Appendix 2: Analysis: Stata commands (Data cleaning)

```
clear all
cd c:\thesis
set more off
cap log close
log using thesis_res2.log,append
use "C:\Users\Jonathan Mkungudza\Documents\MASTERS_BIOSTATISTICS_updated\MDHS_DATA\MWKR7ADT\MWK
> R7AFL.DTA"
***stunting***
recode hw70(min/-200=1 "stunted") (-200/9990=0 "not stunetd") (else=.), gen(stunting)
drop if stunting==.
***mother's age ***
rename v012 age_women
recode  age_women  (0/19=1  ">20  yrs")  (20/34=2  "20-34  yrs")  (35/49  =3  "35-39  yrs"),
generate(agegrp1)
tab agegrp1,m
***sex of the child***
rename b4 child_sex
tab child_sex
***sex of household head ***
rename v151 sex_hh
tab sex_hh

***region***
tab v024
rename v024 region

***location***
tab v025
rename v025 location

***age of household head (v152)***
gen age_HH2=v152
replace age_HH2=999 if v152==98
replace age_HH2=. if age_HH2==999
recode age_HH2 (0/34=1 ">35 yrs") (35/max=2 "35+ yrs"), generate(age_hhgrp)
tab age_hhgrp,m

***Number of underfive children
rename v137 childnumb_und5
recode childnumb_und5 (0/1=1 "<=1 children") (2/max=2 ">=2 children"), generate(Numb_under5)
tab Numb_under5,m
```

```
*** mode of delivery***
rename m17 mode_del
tab mode_del,m

***wealth index***
rename v190 wealth_index
tab1 wealth_index,m

***Ethnicity***
gen ethnicity=v131
replace ethnicity=1 if ethnicity==1 |ethnicity==10
replace ethnicity=2 if ethnicity==2 |ethnicity==4 |ethnicity==7
replace ethnicity=3 if ethnicity==3 |ethnicity==9 |ethnicity==6
replace ethnicity=4 if ethnicity==8
replace ethnicity=6 if ethnicity==96
label define ethnicity 1"Chewa" 2"Tumbuka" 3"Lomwe" 4"Ngoni" 5"Yao" 6"other"
label values ethnicity ethnicity
tab ethnicity,m

***Diarrhea
rename h11 diarrhea
replace diarrhea=. if diarrhea==8
replace diarrhea=1 if diarrhea==2
tab diarrhea,m

***Number of ANC Visits***
gen ANCvisit= m14
replace ANCvisit=. if m14==98
recode ANCvisit (0/3=1 "<= 3")(4/max=2 "4 above"),gen(ANC_VISGRP)
tab ANC_VISGRP,m ///Jonathan: thinking of excluding this from analysis since 1045 participants
have missing records

***Child age in months***
rename b19 curr_agemonth
recode curr_agemonth (0/5=1 "0-6 months")(6/18=2 "6-18 months")(19/max=4 "above 18 months"),gen
(age_childgrp)
tab age_childgrp,m

***Family size***
rename v136 hh_meb_nu
recode hh_meb_nu (0/3=1 "small")(4/6=2 "medium")(7/max=3 "large"),gen(fam_size)
tab fam_size,m

***Body mass index***
gen BM_index=v445/100
recode BM_index (0/18.49=1 "<18.5")(18.5/24.99=2 "18.5-24")(25/max=3 ">=25"),gen(BMI_GRP)
```

```
tab BMI_GRP,m

***marital status***
gen mar_status=v501
recode mar_status (0=0 "single")(1/2=1 "married") (3 4 5 =2 "ever married"), gen(marital_status)
tab marital_status,m

***place of delivery***
rename m15 del_place
recode  del_place(11/12  96  =1  "home")(21  22  23  26  31  32  33  34=0  "health  facility"),  gen
(place_deliver)
label var place_deliver "place of delivery"
tab place_deliver,m

***distance to health facility***
recode v467d (1=1 "long distance")(2=2 "short distance"), gen(dist_facility)
label var dist_facility "distance to a health facility"
tab dist_facility,m

***delivery assistance***
gen del_assistance=.
replace del_assistance=1 if m3a==1 |m3b==1|m3h ==1
replace del_assistance=2 if m3g ==1|m3i ==1|m3k ==1|m3n ==1
label define del_assistance 1"health personnel" 2"not health personnel"
label values del_assistance del_assistance
tab del_assistance,m

***birth number***
recode b0 (0=0 "singleton" )(1 2 3=1 "multiple"),gen(birth_typ)
tab birth_typ,m

***Birth order***
recode bord (1/1=1 "first_born")(2/4=2 "2nd-4th")(5/max=3 "5th or above"),gen(birth_order)
tab birth_order

***education of the mother***
recode v106 (0=0 "No education")(1=1 "Primary")(2 3=2 "Secondary and above"),gen(meducation)
tab meducation,m

***Mother's religion***
 gen religion_cat=v130
 replace religion_cat=1 if v130==2 | v130==3 |v130==4
 replace religion_cat=2 if v130==1
 replace religion_cat=3 if v130==6
 replace religion_cat=4 if v130==5
 replace religion_cat=. if v130==7 |v130==96
 label define religion_cat  1"Protestant" 2"catholic" 3"muslim" 4"other"
```

```
 label values religion_cat religion_cat
 tab religion_cat,m


***/birthweight***
gen brth_weight=m19
replace brth_weight=. if m19==9996 | m19==9998
recode brth_weight (0/2499=1 "Low birth")(2500/max=2 "Normal birth"),gen(brth_weightgrp)
replace brth_weightgrp=2 if m18==1 | m18==2 |m18==3 & brth_weight==.
replace brth_weightgrp=1 if m18==4 | m18==5 & brth_weight==.
tab brth_weightgrp


gen cough_fever=.
replace cough_fever=0 if h31==0 |h22==0
replace cough_fever=1 if h31==1 |h22==1
label define cough_fever 0"No" 1"Yes"
label values cough_fever cough_fever
tab cough_fever,m

*** Mother's occupation***
recode    v717    (0=0    "Not    working")(4    5=1    "Agricultural    worker")    (1    8=2
"Proff/technical/managerial")(3  7=3  "Sales  and  services")(6  9=4  "Domestic  and  unskilled"),
gen(occu_cat)
 tab occu_cat,m

 ***child's anemia level***
 recode v457 (4=0 "Not anemic")(1 2 3=1 "Anemic"), gen (anemic_grp)
 tab anemic_grp,m

drop if anemic_grp==.
drop if brth_weightgrp==.
drop if cough_fever==.
drop if religion_cat==.
drop if BMI_GRP==.
drop if diarrhea==.
drop if mode_del==.
drop if age_hhgrp==.

tab anemic_grp,m
tab brth_weightgrp,m
tab cough_fever,m
tab religion_cat,m
tab BMI_GRP,m
tab diarrhea,m
tab mode_del,m
tab age_hhgrp,m
```

```
keep stunting agegrp1 region ANC_VISGRP      sex_hh  age_hhgrp   child_sex   age_childgrp   BMI_GRP
ethnicity meducation location Numb_under5    marital_status fam_size    occu_cat    brth_weightgrp
mode_del diarrhea      anemic_grp     place_deliver  birth_typ      cough_fever      dist_facility
wealth_index del_assistance religion_cat  birth_order

tab1 stunting agegrp1 region ANC_VISGRP      sex_hh  age_hhgrp   child_sex   age_childgrp   BMI_GRP
ethnicity meducation location Numb_under5    marital_status fam_size    occu_cat    brth_weightgrp
mode_del diarrhea      anemic_grp     place_deliver  birth_typ      cough_fever      dist_facility
wealth_index del_assistance religion_cat  birth_order,m

save "C:\Users\jmkungudza\Documents\MASTERS_BIOSTATISTICS_updated\DATA\prediction\data.dta"


y
```

Appendix 2: R Script (Model training and Evaluation)

```
setwd ("~/MASTERS_BIOSTATISTICS_updated/DATA/prediction")
data <- read.csv("~/MASTERS_BIOSTATISTICS_updated/DATA/prediction/Mydata.csv")

data$stunting<-factor(data$stunting,level=c(0,1),labels=c("not stunted", "stunted"))


data$wealth_index<-as.factor(data$wealth_index)
data$child_sex<-as.factor(data$child_sex)
data$diarrhea<-as.factor(data$diarrhea)
data$Numb_under5<-as.factor(data$Numb_under5)
data$age_childgrp<-as.factor(data$age_childgrp)
data$fam_size<-as.factor(data$fam_size)
data$BMI_GRP<-as.factor(data$BMI_GRP)
data$marital_status<-as.factor(data$marital_status)
data$dist_facility<-as.factor(data$dist_facility)
data$del_assistance<-as.factor(data$del_assistance)
data$birth_typ<-as.factor(data$birth_typ)
data$meducation<-as.factor(data$meducation)
data$location<-as.factor(data$location)
data$ethnicity<-as.factor(data$ethnicity)

data$agegrp1<-as.factor(data$agegrp1)
data$mode_del<-as.factor(data$mode_del)

data$age_hhgrp<-as.factor(data$age_hhgrp)
data$place_deliver<-as.factor(data$place_deliver)
data$birth_order<-as.factor(data$birth_order)

data$religion_cat<-as.factor(data$religion_cat)
data$brth_weightgrp<-as.factor(data$brth_weightgrp)
data$cough_fever<-as.factor(data$cough_fever)
data$occu_cat<-as.factor(data$occu_cat)
data$anemic_grp<-as.factor(data$anemic_grp)
data$region<-as.factor(data$region)



set.seed (123)
ind<-sample(2,nrow(data),replace=T,prob=c(0.8,0.2))
traindata<-data [ind==1, ]
```

```
testdata<-data [ind==2, ]


str(data)
#Required packages
install.packages("InformationValue")
install.packages("pROC")
install.packages("ggpubr")
install.packages("OptimalCutpoints")
install.packages("dplyr")
library("dplyr")
library("OptimalCutpoints")
library(MASS)
library(pROC)
library(ROCR)
library(caret)
library(InformationValue)
library(bootStepAIC)
library(Boruta)
library(randomForest)
library(glmnet)
library(ggplot2)
library(ggpubr)


table(traindata$stunting)
table(testdata$stunting)
table(data$stunting)

prop.test(x=1771,n=4976)
prop.test(x=1431, n=3990)
prop.test(x=340, n=986)


#Backward variable selection method
model_All<-glm(stunting ~.,data=traindata, family="binomial")
mod_step<-stepAIC(model_All,direction="backward",trace=FALSE)
mod_step

model_boot<-boot.stepAIC(model_All,traindata,B=50)
model_boot

#Forward variable selection method
fitAll<-glm(stunting~.,data=traindata,family="binomial")
model_intecept<-glm(stunting~1,data=traindata,family="binomial")
summary(model_intecept)
mod_stepFo<-stepAIC(model_intecept,direction="forward",scope=formula(fitAll))
```

```
summary(mod_stepFo)


#stepwise variable selection method
mod_stepwise<-stepAIC(model_intecept,direction="both",scope=formula(fitAll))
summary(mod_stepwise)

#Boruta:Random forest variable selection
boruta <-Boruta(stunting~.,data=traindata,doTrace=2,maxRuns=500)
print(boruta)
plot(boruta,las=2,cex.axis=0.7)

#getting selected variables
getNonRejectedFormula(boruta)


# 5) Training LASSO model
train_x<-model.matrix(stunting~ .,data=traindata)[, -8]
train_y<-traindata[,"stunting"]

test_x<-model.matrix(stunting~ .,data=testdata)[, -8]
test_y<-testdata[,"stunting"]

#adjust x,y size of plot
options(repr.plot.width=10,repr.plot.height=8)
mod_lasso<-glmnet(
  x=train_x,
  y=train_y,
  family="binomial",
  alpha=1
)



#CROSS VALIDATION
set.seed(2345)
mod_lasso_cv<-cv.glmnet(
  x=train_x,
  y=train_y,
  type.measure="class",
  family="binomial",
  alpha=1
)

#plot results of cv
par(mfrow=c(1,2))
plot(mod_lasso_cv, main="Misclsification error curve")
```

```
plot(mod_lasso, xvar="lambda", main="LASSO coefficient profile")

(best.lambda<-mod_lasso_cv$lambda.min)

#final variable selected with the best lambda
lasso_model<-glmnet(x=train_x,y=train_y,family="binomial",alpha=1,lambda=best.lambda)
lasso_model$beta


#Training logistic models

# 1)fit the selected model (AIC_backward)
model_backward<-glm(formula=stunting ~ region + location + wealth_index + diarrhea +
                      Numb_under5 + ethnicity + age_childgrp + fam_size + BMI_GRP +
                      dist_facility + birth_typ + birth_order + brth_weightgrp
                    ,data=traindata,family="binomial")
summary(model_backward)

# Estimating cutpoints on training data
p<-predict(model_backward,newdata=traindata, type="response")


traindata1 <- cbind(traindata, p)
traindata1$stunting<- ifelse(traindata1$stunting=="stunted",1,0)
cutpoint1<-optimal.cutpoints(X="p",status="stunting",
tag.healthy=0,method=c("MaxSe"),data=traindata1,
categorical.cov=NULL,pop.prev=NULL,control=control.cutpoints(),ci.fit=TRUE)


summary(cutpoint1)


# performance measure(ROC) on test data
predicted<-predict(model_backward, newdata=testdata, type="response")

pt<-predict(model_backward,newdata=testdata, type="response")
pb <- prediction(pt, testdata$stunting)
prfb <- performance(pb, measure = "tpr", x.measure = "fpr")

#Cut off points at 0.5
testdata$stunting<- ifelse(testdata$stunting=="stunted",1,0)
confusionMatrix(testdata$stunting,predicted)
misClassError(testdata$stunting,predicted)
sensitivity(testdata$stunting,predicted)
specificity(testdata$stunting,predicted)

#Cut off points at 0.36
```

```
confusionMatrix(testdata$stunting,predicted,threshold=0.36)
misClassError(testdata$stunting,predicted,threshold=0.36)
sensitivity(testdata$stunting,predicted,threshold=0.36)
specificity(testdata$stunting,predicted,threshold=0.36)

auc <- performance(pb, measure = "auc")
auc <- auc1@y.values[[1]]
auc
ci.auc(testdata$stunting, pt)

V = caret::varImp(model_backward)

ggplot2::ggplot(V, aes(x=reorder(rownames(V),Overall), y=Overall)) +
  geom_point( color="blue", size=4, alpha=0.6)+
  geom_segment( aes(x=rownames(V), xend=rownames(V), y=0, yend=Overall),
                color='skyblue') +
  xlab('Variable')+
  ylab('Overall Importance (backward model)')+
  theme_light() +
  coord_flip()

# 2)fit the selected model (AIC_forward)
model_forward<-glm(formula = stunting ~ age_childgrp + wealth_index + brth_weightgrp + birth_typ
+
                     BMI_GRP + child_sex + occu_cat + birth_order + Numb_under5 +
                     fam_size + dist_facility + location + diarrhea + ethnicity +
                     region, family = "binomial", data = traindata)
summary(model_forward)

# Estimating cutpoints on training data
pf<-predict(model_forward,newdata=traindata, type="response")

traindataf1 <- cbind(traindata, pf)
traindataf1$stunting<- ifelse(traindataf1$stunting=="stunted",1,0)
cutpointf1<-optimal.cutpoints(X="pf",status="stunting",
tag.healthy=0,method=c("SpEqualSe"),data=traindataf1,
categorical.cov=NULL,pop.prev=NULL,control=control.cutpoints(),ci.fit=TRUE)
table(traindataf1$stunting)


summary(cutpointf1)

# ## performance measure(ROC) on test data
predicted1<-predict(model_forward, newdata=testdata, type="response")

pfo<-predict(model_forward,newdata=testdata, type="response")
pfo1 <- prediction(pfo,testdata$stunting)
```

```
prfo1 <- performance(pfo1, measure = "tpr", x.measure = "fpr")

#Cut off points at 0.5
confusionMatrix(testdata$stunting,predicted1)
misClassError(testdata$stunting,predicted1)
sensitivity(testdata$stunting,predicted1)
specificity(testdata$stunting,predicted1)

#Cut off points at 0.36
confusionMatrix(testdata$stunting,predicted1,threshold=0.36)
misClassError(testdata$stunting,predicted1,threshold=0.36)
sensitivity(testdata$stunting,predicted1,threshold=0.36)
specificity(testdata$stunting,predicted1,threshold=0.36)

auc1 <- performance(pfo1, measure = "auc")
auc1 <- auc1@y.values[[1]]
auc1
ci.auc(testdata$stunting, pfo)

B = caret::varImp(model_forward)

ggplot2::ggplot(B, aes(x=reorder(rownames(B),Overall), y=Overall)) +
  geom_point( color="blue", size=4, alpha=0.6)+
  geom_segment( aes(x=rownames(B), xend=rownames(B), y=0, yend=Overall),
                color='skyblue') +
  xlab('Variable')+
  ylab('Overall Importance (forward model)')+
  theme_light() +
  coord_flip()


# 3)fit the selected model (AIC_BOTH)
model_stepwise<-glm(formula = stunting ~ age_childgrp + wealth_index + brth_weightgrp +
                      birth_typ + BMI_GRP + child_sex + occu_cat + birth_order +
                      Numb_under5 + fam_size + dist_facility + location + diarrhea +
                      ethnicity + region, family = "binomial",
                    data = traindata)

summary(model_stepwise)


# Estimating cutpoints on training data
ps<-predict(model_stepwise,newdata=traindata, type="response")

traindatas1 <- cbind(traindata, ps)
traindatas1$stunting<- ifelse(traindatas1$stunting=="stunted",1,0)
cutpoints1<-optimal.cutpoints(X="ps",status="stunting",
```

```
tag.healthy=0,method=c("SpEqualSe"),data=traindatas1,
categorical.cov=NULL,pop.prev=NULL,control=control.cutpoints(),ci.fit=TRUE)
table(traindatas1$stunting)


summary(cutpoints1)

## performance measure(ROC) on test data
predicted2<-predict(model_stepwise, newdata=testdata, type="response")

ps1<-predict(model_stepwise,newdata=testdata, type="response")
ps2 <- prediction(ps1, testdata$stunting)
psf2 <- performance(ps2, measure = "tpr", x.measure = "fpr")

#Cut off points at 0.5
confusionMatrix(testdata$stunting,predicted2)
misClassError(testdata$stunting,predicted2)
sensitivity(testdata$stunting,predicted2)
specificity(testdata$stunting,predicted2)

#Cut off points at 0.36
confusionMatrix(testdata$stunting,predicted2,threshold=0.36)
misClassError(testdata$stunting,predicted2,threshold=0.36)
sensitivity(testdata$stunting,predicted2,threshold=0.36)
specificity(testdata$stunting,predicted2,threshold=0.36)

auc2 <- performance(ps2, measure = "auc")
auc2 <- auc2@y.values[[1]]
auc2
ci.auc(testdata$stunting, ps2)


# 4)Fitting selected model (RF_Logistic)

rf_sel<-glm(formula=stunting ~ location + wealth_index + agegrp1 + age_hhgrp + age_childgrp +
              fam_size + BMI_GRP + dist_facility + birth_typ + birth_order +
              brth_weightgrp,family="binomial",data=traindata)
summary(rf_sel)

# Estimating cutpoints on training data
prf<-predict(rf_sel,newdata=traindata, type="response")

traindatar1 <- cbind(traindata, prf)
traindatar1$stunting<- ifelse(traindatar1$stunting=="stunted",1,0)
cutpointr1<-optimal.cutpoints(X="prf",status="stunting",
tag.healthy=0,method=c("SpEqualSe"),data=traindatar1,
categorical.cov=NULL,pop.prev=NULL,control=control.cutpoints(),ci.fit=TRUE)
```

```
table(traindatar1$stunting)


summary(cutpointr1)

## performance measure(ROC) on test data
predicted23<-predict(rf_sel, newdata=testdata, type="response")

p23<-predict(rf_sel,newdata=testdata, type="response")
pr23 <- prediction(p23, testdata$stunting)
prf23 <- performance(pr23, measure = "tpr", x.measure = "fpr")

#Cut off points at 0.5
confusionMatrix(testdata$stunting,predicted23)
misClassError(testdata$stunting,predicted23)
sensitivity(testdata$stunting,predicted23)
specificity(testdata$stunting,predicted23

#Cut off points at 0.37
confusionMatrix(testdata$stunting,predicted23,threshold=0.37)
misClassError(testdata$stunting,predicted23,threshold=0.37)
sensitivity(testdata$stunting,predicted23,threshold=0.37)
specificity(testdata$stunting,predicted23,threshold=0.37)

auc23 <- performance(pr23, measure = "auc")
auc23 <- auc23@y.values[[1]]
auc23
ci.auc(testdata$stunting, p23)

#4)Fitting selected model (LASSO_Logistic)

LASSO_sel<-glm(formula=stunting ~ location + wealth_index + agegrp1 + age_hhgrp + age_childgrp +
                fam_size + BMI_GRP + dist_facility + Numb_under5 + religion_cat + meducation +
                birth_typ  +  birth_order  +  region  +  diarrhea  +  occu_cat  +  anemic_grp  +
del_assistance +
                ethnicity + child_sex + sex_hh +
                brth_weightgrp,family="binomial",data=traindata)
summary(LASSO_sel)

# Estimating cutpoints on training data
pl<-predict(LASSO_sel,newdata=traindata, type="response")

traindatal1 <- cbind(traindata, pl)
traindatal1$stunting<- ifelse(traindatal1$stunting=="stunted",1,0)
cutpointl1<-optimal.cutpoints(X="pl",status="stunting",
tag.healthy=0,method=c("SpEqualSe"),data=traindatal1,
categorical.cov=NULL,pop.prev=NULL,control=control.cutpoints(),ci.fit=TRUE)
```

```
table(traindatal1$stunting)


summary(cutpointl1)

## performance measure(ROC) on test data
predicted24<-predict(LASSO_sel, newdata=testdata, type="response")

p24<-predict(LASSO_sel,newdata=testdata, type="response")
pr24 <- prediction(p24, testdata$stunting)
prf24 <- performance(pr24, measure = "tpr", x.measure = "fpr")

#Cut off points at 0.5
confusionMatrix(testdata$stunting,predicted24)
misClassError(testdata$stunting,predicted24)
sensitivity(testdata$stunting,predicted24)
specificity(testdata$stunting,predicted24)

#Cut off points at 0.37
confusionMatrix(testdata$stunting,predicted24,threshold=0.37)
misClassError(testdata$stunting,predicted24,threshold=0.37)
sensitivity(testdata$stunting,predicted24,threshold=0.37)
specificity(testdata$stunting,predicted24,threshold=0.37)

auc24 <- performance(pr24, measure = "auc")
auc24 <- auc24@y.values[[1]]
auc24
ci.auc(testdata$stunting, p24)

#Model judgement

model_judge<-glm(formula=stunting ~ location + wealth_index + child_sex +  age_childgrp +
                   fam_size + Numb_under5 + meducation +
                   birth_typ +  diarrhea  +
             brth_weightgrp,family="binomial",data=traindata)

summary(model_judge)
ODDS<-exp(cbind("odds ratio"=coef(model_judge),confint.default(model_judge,level = 0.95)))
print(ODDS)

# Estimating cutpoints on training data
pj<-predict(model_judge,newdata=traindata, type="response")

traindataj1 <- cbind(traindata, pj)
traindataj1$stunting<- ifelse(traindataj1$stunting=="stunted",1,0)
cutpointj1<-optimal.cutpoints(X="pj",status="stunting",
tag.healthy=0,method=c("SpEqualSe"),data=traindataj1,
```

```
categorical.cov=NULL,pop.prev=NULL,control=control.cutpoints(),ci.fit=TRUE)
table(traindataj1$stunting)

summary(cutpointj1)


# performance measure(ROC) on test data
predicted7<-predict(model_judge, newdata=testdata, type="response")

p7<-predict(model_judge,newdata=testdata, type="response",threshold=0.37)
pr7 <- prediction(p7, testdata$stunting)
prf7 <- performance(pr7, measure = "tpr", x.measure = "fpr")

#Cut off points at 0.5
confusionMatrix(testdata$stunting,predicted7)
misClassError(testdata$stunting,predicted7)
sensitivity(testdata$stunting,predicted7)
specificity(testdata$stunting,predicted7)

#Cut off points at 0.37
confusionMatrix(testdata$stunting,predicted7,threshold=0.37)
misClassError(testdata$stunting,predicted7,threshold=0.37)
sensitivity(testdata$stunting,predicted7,threshold=0.37)
specificity(testdata$stunting,predicted7,threshold=0.37)

auc7 <- performance(pr7, measure = "auc",threshold=0.37)
auc7 <- auc7@y.values[[1]]
auc7
ci.auc(testdata$stunting, p7)


#analysis by gender (male vs female)

model_judge_gender<-glm(formula=stunting ~ location + wealth_index +  age_childgrp +
                          fam_size + Numb_under5 + meducation +
                          birth_typ +  diarrhea  +
                          brth_weightgrp,family="binomial",data=traindata)

pg<-predict(model_judge_gender,newdata=testdata, type="response")

testdatag1 <- cbind(testdata, pg)
testdatag1$stunting<- ifelse(testdatag1$stunting=="stunted",1,0)
cutpointg1<-optimal.cutpoints(X="pg",status="stunting",
tag.healthy=0,method=c("SpEqualSe"),data=testdatag1,
categorical.cov="child_sex",pop.prev=NULL,control=control.cutpoints(),ci.fit=TRUE)

summary(cutpointg1)
```

```
#analysis by location (rural vs urban)

model_judge_loc<-glm(formula=stunting ~ child_sex + wealth_index +  age_childgrp +
                       fam_size + Numb_under5 + meducation +
                       birth_typ +  diarrhea  +
                       brth_weightgrp,family="binomial",data=traindata)

pl<-predict(model_judge_loc,newdata=testdata, type="response")

testdatal1 <- cbind(testdata, pl)
testdatal1$stunting<- ifelse(testdatal1$stunting=="stunted",1,0)
cutpointl1<-optimal.cutpoints(X="pl",status="stunting",
tag.healthy=0,method=c("SpEqualSe"),data=testdatal1,
categorical.cov="location",pop.prev=NULL,control=control.cutpoints(),ci.fit=TRUE)

summary(cutpointl1)



#model fit_,common variable
model_com<-glm(formula=stunting ~ wealth_index +  age_childgrp +
                 fam_size + birth_typ + brth_weightgrp,family="binomial",data=traindata)

summary(model_com)
# Estimating cutpoints on training data
pc<-predict(model_com,newdata=traindata, type="response")

traindatac1 <- cbind(traindata, pc)
traindatac1$stunting<- ifelse(traindatac1$stunting=="stunted",1,0)
cutpointc1<-optimal.cutpoints(X="pc",status="stunting",
tag.healthy=0,method=c("SpEqualSe"),data=traindatac1,
categorical.cov=NULL,pop.prev=NULL,control=control.cutpoints(),ci.fit=TRUE)
table(traindatac1$stunting)


summary(cutpointc1)

# performance measure(ROC) on test data
predicted8<-predict(model_com, newdata=testdata, type="response")

p8<-predict(model_com,newdata=testdata, type="response")
pr8 <- prediction(p8, testdata$stunting)
prf8 <- performance(pr8, measure = "tpr", x.measure = "fpr")

#Cut off points at 0.5
confusionMatrix(testdata$stunting,predicted8)
```

```
misClassError(testdata$stunting,predicted8)
sensitivity(testdata$stunting,predicted8)
specificity(testdata$stunting,predicted8)

#Cut off points at 0.37
confusionMatrix(testdata$stunting,predicted8,threshold=0.37)
misClassError(testdata$stunting,predicted8,threshold=0.37)
sensitivity(testdata$stunting,predicted8,threshold=0.37)
specificity(testdata$stunting,predicted8,threshold=0.37)

auc8 <- performance(pr8, measure = "auc",threshold=0.37)
auc8 <- auc8@y.values[[1]]
auc8
ci.auc(testdata$stunting, p8,threshold=0.37)


V = caret::varImp(model_backward)
B = caret::varImp(model_forward)
S = caret::varImp(model_stepwise)
L = caret::varImp(LASSO_sel)
F = caret::varImp(rf_sel)
Q = caret::varImp(model_judge)
C = caret::varImp(model_com)

bacwd<-ggplot2::ggplot(V, aes(x=reorder(rownames(V),Overall), y=Overall)) +
  geom_point( color="blue", size=2, alpha=0.6)+
  geom_segment( aes(x=rownames(V), xend=rownames(V), y=0, yend=Overall),
                color='skyblue') +
  xlab('Variable')+
  ylab('importance')+
  theme_light() +
  coord_flip()


forwd<-ggplot2::ggplot(B, aes(x=reorder(rownames(B),Overall), y=Overall)) +
  geom_point( color="blue", size=2, alpha=0.6)+
  geom_segment( aes(x=rownames(B), xend=rownames(B), y=0, yend=Overall),
                color='skyblue') +
  xlab('Variable')+
  ylab('importance')+
  theme_light() +
  coord_flip()


step<-ggplot2::ggplot(S, aes(x=reorder(rownames(S),Overall), y=Overall)) +
  geom_point( color="blue", size=2, alpha=0.6)+
  geom_segment( aes(x=rownames(S), xend=rownames(S), y=0, yend=Overall),
```

```
                color='skyblue') +
  xlab('Variable')+
  ylab('importance')+
  theme_light() +
  coord_flip()

rf<-ggplot2::ggplot(F, aes(x=reorder(rownames(F),Overall), y=Overall)) +
  geom_point( color="blue", size=2, alpha=0.6)+
  geom_segment( aes(x=rownames(F), xend=rownames(F), y=0, yend=Overall),
                color='skyblue') +
  xlab('Variable')+
  ylab('importance')+
  theme_light() +
  coord_flip()

laso<-ggplot2::ggplot(L, aes(x=reorder(rownames(L),Overall), y=Overall)) +
  geom_point( color="blue", size=2, alpha=0.6)+
  geom_segment( aes(x=rownames(L), xend=rownames(L), y=0, yend=Overall),
                color='skyblue') +
  xlab('Variable')+
  ylab('importance')+
  theme_light() +
  coord_flip()



judge<-ggplot2::ggplot(Q, aes(x=reorder(rownames(Q),Overall), y=Overall)) +
  geom_point( color="blue", size=2, alpha=0.6)+
  geom_segment( aes(x=rownames(Q), xend=rownames(Q), y=0, yend=Overall),
                color='skyblue') +
  xlab('Variable')+
  ylab('importance')+
  theme_light() +
  coord_flip()

com_var<-ggplot2::ggplot(C, aes(x=reorder(rownames(C),Overall), y=Overall)) +
  geom_point( color="blue", size=2, alpha=0.6)+
  geom_segment( aes(x=rownames(C), xend=rownames(C), y=0, yend=Overall),
                color='skyblue') +
  xlab('Variable')+
  ylab('importance')+
  theme_light() +
  coord_flip()

ggarrange(bacwd,forwd,step +rremove("x.text"),
          labels=c("A","B","C"),
          ncol=3,nrow=1)
```

```
ggarrange(laso,rf, judge,com_var +rremove("x.text"),
          labels=c("D","E","F"),
          ncol=3,nrow=1)


#COMPARING ROC PLOT of 7 Models

plot(prfb,col="red",lwd=2)
plot(prfo1,add=TRUE,col="green",lwd=2)
plot(psf2,add=TRUE,col="blue",lwd=2)
plot(prf23,add=TRUE,col="black",lwd=2)
plot(prf24,add=TRUE,col="yellow",lwd=2)
plot(prf6,add=TRUE,col="pink",lwd=2)
plot(prf8,add=TRUE,col="orange",lwd=2)
title(main="Comparison of ROC Curves", font.main=4)
plot_range<-range(0,0.5,0.5,0.5,0.5)
legend(0.5,                                plot_range[2],c("backward","forward","stepwise","random
forest","LASSO","judgement","common_var"), cex=0.2,
       col=c("red","green","blue","black","yellow","pink","orange"), pch=21:22, lty=1:2)
abline(a=0,b=1,lwd=2,lty=2,col="gray")
```