2022

Machine learning based prediction of life expectancy.

Lipesa, Brian Aholi
Strathmore Institute of Mathematical Sciences
Strathmore University

Machine Learning Based Prediction of Life Expectancy

Brian Aholi Lipesa

136562

Submitted in partial fulfilment of the requirements for the degree of Master of Science in Statistical Science of Strathmore University

Institute of Mathematical Sciences

Strathmore University

Nairobi, Kenya

October, 2022

Declaration

I declare that this work has not been previously submitted and approved for award of a degree by this or any other University. To the best of my knowledge and belief, the thesis contains no material previously published or written by another person except where due reference is made in the thesis itself.

Name: **Brian Aholi Lipesa**

Signature: .....

Date: August 21, 2022

Approval

The thesis of Brian Aholi Lipesa was reviewed and approved by the following:

Professor Bernard Omolo

Supervisor,

Institute of Mathematical Sciences, Strathmore University.

Dr. Elphas Okango

Supervisor,

Institute of Mathematical Sciences, Strathmore University.

Dr. Godfrey Madigu

Dean,

Institute of Mathematical Sciences, Strathmore University.

Dr. Bernard Shibwabo

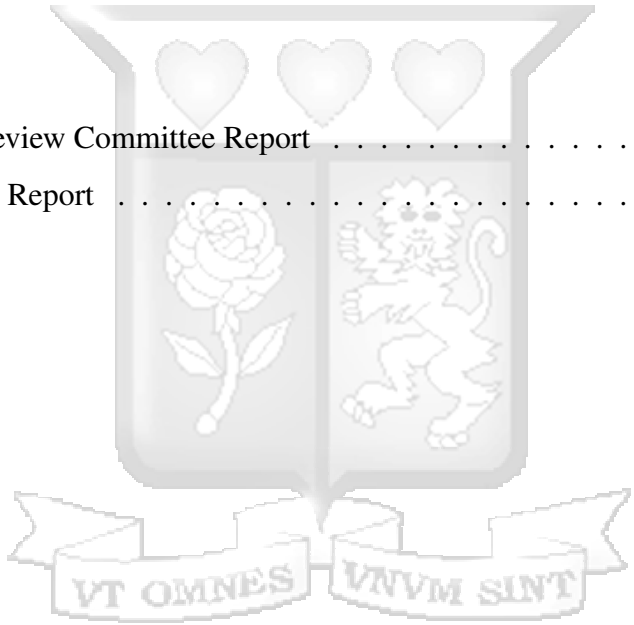
Director,

Office of Graduate Studies, Strathmore University.

Abstract

The social and financial systems of many nations throughout the world are significantly impacted by life expectancy (LE) models. Numerous studies have pointed out the crucial effects that life expectancy projections will have on societal issues and the administration of the global healthcare system. These approaches offer a variety of strategies to enhance society-related advanced care planning and healthcare. Over time, research has proven that the vast majority of the existing factors were insufficient to forecast the lifespan of the general population. An understanding of the chosen sampling population's death rate served as the foundation for earlier models. Researchers have asserted that despite improvements in forecasting approaches and meticulous work in the past, there are still several elements that must be taken into account to determine life expectancy rates in addition to death rates. As a result, life expectancy research now includes a broader focus on issues related to education, health, the economy, and social welfare. In this study, the author developed a model for estimating life expectancy rates taking into consideration health, socioeconomic, and behavioral characteristics by using the eXtreme Gradient Boosting (XGBoost) algorithm to data from 193 UN member states. The effectiveness of the model's prediction was compared to that of the Random Forest (RF) and Artificial Neural Network (ANN) regressors utilized in earlier research. XGBoost attained an MAE and an RMSE of 1.554 and 2.402, respectively. It outperformed the RF and ANN models that achieved MAE and RMSE values of 7.938 and 11.304, and 3.86 and 5.002, respectively. The overall results of this study support XGBoost as a reliable and efficient model for estimating life expectancy.

Table of Contents



List of Figures

List of Tables

List of Abbreviations

AIDS	Acquired Immune Deficiency Syndrome
AKI	Acute Kidney Injury
BMI	Body Mass Index
CHAID	Chi-square Automatic Interaction Detector
COVID-19	Coronavirus Disease
CSM	Conventional Statistical Model
DNN	Deep Neural Network
EU	European Union
GDP	Gross Domestic Product
GHG	Green House Gas
GLM	Generalized Linear Model
HIV	Human Immunodeficiency Virus
KNN	K Nearest Neighbors
LASSO	Least Absolute Shrinkage and Selection Operator
LE	Life Expectancy
MAE	Mean Absolute Error
MAPE	Mean Absolute Percentage Error
MIMIC	Medical Information Mart for Intensive Care
ML	Machine Learning
MSE	Mean Squared Error
RF	Random Forest
RMSE	Root Mean Squared Error
RT-PCR	Real Time Polymerase Chain Reaction
SAPS	Simplified Acute Physiology Score
SDGs	United Nations Sustainable Development Goals
SVM	Support Vector Machine

SVR	Support Vector Regression
UN	United Nations
WHO	World Health Organization
XGBoost	eXtreme Gradient Boosting

Acknowledgement

First and foremost, I thank the Lord God Almighty for his never-ending love, care, grace, wisdom, and good health. Second, I would like to express my gratitude to Professor Bernard Omolo and Dr. Elphas Okango, my supervisors, for their willingness to supervise me, unwavering advice, suggestions, and encouragement throughout the study process. Your unwavering support has allowed me to develop and improve my research abilities. Thirdly, I extend my gratitude to my friends, classmates, academic peers, and the Comprehensive R Archive Network, which includes the R Community, for their continued support and resourcefulness. I have gone this far because of the collective efforts we have put together. Many thanks to the whole Strathmore University administration and faculty for providing me with the chance to be a Stratizen and for creating the most favorable environment amidst the pandemic to reach this milestone. Kanana and my family deserve special thanks for their unconditional support, encouragement, understanding, sacrifice, and willingness to provide a helping hand during this academic journey. You made it possible for me to realize my goal. May God continue to bless you all!

Dedication

This thesis is dedicated to my dad, mom, and siblings: Andrew, Ivonne and Linda.

Chapter 1

Introduction of the Study

1.1 Background of the Study

The Organisation for Economic Cooperation and Development, OECD (2022) defines life expectancy (LE) at birth as how long, on average, a newborn can expect to live, if current death rates remain constant. LE is an estimate of the average number of additional years that a person of a given age can expect to live. Precisely, it tells us the average age of death in a population. Estimates suggest that in the ancient times, LE was around 30 years in all regions of the world (Roser et al., 2013).

Prior to the COVID-19 pandemic, population health was improving globally, increasing the global average LE at birth from 66.8 years in 2000 to 73.3 years in 2019. In 2019, LE for males reached 70.9 years. For females, the equivalent figure was 75.9 years. The African Region still had the lowest LE among WHO regions in 2019, at only 64.5 years, despite experiencing the largest gains over the past two decades. The Region of the Americas had the highest LE (74.1 years) in 2000 but dropped to third place (77.2 years) in 2019 as the European and Western Pacific Regions made accelerated gains, reaching 78.2 and 77.7 years, respectively (World Health Organization, 2021).

At the UN Sustainable Development Summit in September, 2015, world leaders adopted the Sustainable Development Goals (SDGs) as the symbol of the global agenda for development through 2030. These goals are a set of 17 commitments made by the 193 world leaders, for fair and sustainable health at every level. Their aim is to end extreme poverty, inequality and climate change by the year 2030 (Goals, 2022).

Ensuring healthy lives and promotion of the well-being for the population of the world for all at all ages is vital to sustainable development (UN, 2021). Life expectancy is a key measure often used to assess the overall well-being and health of a population. It is an indicator of a country's overall health, reflecting its social and economic conditions, as well as the quality of its public health and healthcare infrastructure, among other factors (Ho and Hendi, 2018).

1.2 Problem Definition

Wang et al. (2016) notes that a fundamental goal of a health system is to prolong life, especially healthy life, into advanced age. A slight growth in LE translates into sizeable increases in the population. Lengthening a population's LE tends to increase the number of individuals living with chronic illnesses, a common feature among the elderly. Thus, the ability to project how populations will age has massive implications on the planning, support and service provision of a Nation.

The computation of life expectancy involves building an ordinary life table (Ayuso et al., 2021). The life table is a population model covering the simple case of a birth cohort, born at the same time period, closed to migration, and followed through successive ages until they die (Guillaume et al., 2002). The life table assumes the homogeneity of cohorts (i.e., subjects have the same distribution of survival times) (Anderson et al., 1980). In life tables, the LE of persons alive at age x is computed as:

$$\mathrm{e}_x = \frac{T_x}{l_x}, \tag{1.1}$$

where T_x is the total number of person-years lived by the cohort from age x until all members of the cohort have died and l_x is the number of persons alive at age x (Guillaume et al., 2002).

Since the use of life tables takes a long time following through successive ages till individuals' departure from life, the resultant estimates from this approach may not be applicable to the current populations due to the current technological and other factors affecting LE. The homogeneity assumption of life tables is not very realistic in practice.

Furthermore, in practice, individuals are prone to censoring either by death or out-migration resulting into biasness of the estimates from life tables. As a result, a robust and more accurate approach is inevitable, hence this study. This study proposed the use of eXtreme Gradient Boosting (XGBoost) algorithm to predict life expectancy, considering health, socioeconomic and behavioral factors.

XGBoost has been widely used to produce state-of-the-art results on many machine learning problems (Chen and Guestrin, 2016). XGBoost is competitive in terms of its ability to: handle missingness, build machine learning models more quickly and add built-in regularization to achieve accuracy gains (Wade, 2020). These features renders XGBoost to be a faster, more accurate, and a more desirable algorithm over other comparable ensemble algorithms, for the dataset at hand.

1.3 Research Objectives

1.3.1 General Objective

The objective of this study was to formulate a supervised machine learning model that will be used to predict life expectancy at birth.

1.3.2 Specific Objectives

The specific study objectives were:

a) To identify and select the predictors considered important in contributing to life expectancy.

b) To assess the predictive performance of the developed model based on life expectancy data.

1.4 Research Questions

The specific research questions depended on whether the study variables were categorical or continuous. Precisely, this study sought to answer the following research questions:

a) Which predictors are considered important in contributing to life expectancy?

b) How does the developed model perform on life expectancy data?

1.5 Scope of the Study

The study focused on the 193 UN member states from the year 2000-2015, sourced from the WHO and UN databases consisting of 2937 observations of 21 variables. The predictor variables for the study were: year, adult mortality, infant deaths, per capita alcohol consumption, Hepatitis B immunization, measles cases, BMI, under-five deaths, polio immunization, government health expenditure, diphtheria immunization, HIV/AIDS deaths, thinness (percent of thinness among children from age 10-19 and age 5-9), country, region, income group, GDP, population and the number of years at school. The response variable was the life expectancy in age. The detailed description of the study variables is as presented in Table 1.1.

Table 1.1: Description of the Response and the Twenty Independent Study Variables.

Variable Name	**Description**
Country	Country Name
Region	Global regional location
IncomeG	Income Group *(Low, Middle, High)*
Year	The calendar year of interest
LifeExp	Life expectancy in age
AdultM	Number of people dying between 15-60 years per 1000 population
InfantD	Number of infant deaths per 1000 population
Alcohol	Recorded per capita alcohol consumption *(in litres)*

Variable Name	Description
HepsB	Percentage Hepatitis B immunization coverage among 1 year olds
Measles	Number of measles reported cases per 1000 population
BMI	Average body mass index of the entire population
Und5Deaths	Number of under five deaths per 1000 population
Polio	Percent of polio immunization coverage among one year olds
GovHealthExp	Government expenditure of health as a percentage of total government expenditure
Diph	Percent of diphtheria immunization coverage among one year olds
HIV	HIV/AIDS deaths per 1000 population
GDP	Gross domestic product per capita in dollars
Pop	The country's population
Thin10-19Yrs	Percent of thinness among children from age 10-19
Thin5-9Yrs	Percent of thinness among children from age 5-9
Schooling	Number of years at school

1.6 Significance of the Study

The research outputs from this study shall inform governments in the development of policies that will be used to improve the longevity of life for their citizens. Governments will have a better outlook on how to address policies relating to lifestyles, elderly care, accessible healthcare for all, prenatal and postnatal care, and transition from secondary to post-secondary education.

Additionally, utilization of these outputs will better the distribution of scarce resources such as societal well-being support and health care financing, to persons and much deserving communities. Besides, governments will be able to develop programmes that promote life expectancy of their citizens and plan on how to implement them, including any eventualities that might shorten the lifespan of the people.

The research findings shall also contribute new knowledge in machine learning and life expectancy research, which scholars with similar interests can use to identify and explore new areas for future research. Similarly, these findings may be used by practitioners to provide better healthcare services and improve the socioeconomic factors that affect life expectancy through targeted medical campaigns.

1.7 Dissemination and Utilisation of Study Results

The results from this study will be publicized through publications in journals, conference presentations, academic blogs, the Strathmore University online repository and the Library catalogue. It is the hope of the researcher that the results from this study shall be applied in policy decision making, development of lifespan promotion programmes, and similar research studies on machine learning based life expectancy prediction.

Chapter 2

Literature Review

2.1 Introduction

The section presents empirical literature on the overview of life expectancy, discusses the concept of machine learning and its application, the use of XGBoost in predictions, and reviews the various machine learning based algorithms employed in life expectancy prediction.

2.2 Global Life Expectancy

WHO (2022) defines life expectancy at birth as the average number of years that a newborn could expect to live, if he or she were to pass through life exposed to the sex and age-specific death rates prevailing at the time of his or her birth, for a specific year, in a given country, territory, or geographic area. WorldBank (2022) on the other hand terms life expectancy at birth as an indication of the number of years a newborn infant would live if prevailing patterns of mortality at the time of its birth were to stay the same throughout its life.

Gulland (2016) states that according to data published by WHO, global life expectancy rose by five years since 2000, this being the most rapid increase since the 1960s. The author notes that global life expectancy for children that were birthed in 2015 was 71.4 years, however there existed stark differences between countries. The author further observes that the highest life expectancy in the world was 83.7 years for both genders reported in Japan, followed by 83.4 years in Switzerland, and 83.1 years in Singapore. Sierra Leone had the lowest life expectancy in the world of just 50.1 years, followed by 50.4 years reported in Angola.

United Kingdom's life expectancy was 81.2 years, in comparison to the United States's 79.3 years. Due to the AIDS epidemic and the collapse of the Soviet Union, LE dropped in Africa and Eastern Europe in the 1990s, respectively. The author points out that globally, females lived longer than their male counterparts in every country and WHO region on average. Generally, in the year 2015 life expectancy for women was 73.8 years and for men was 69.1 years (Gulland, 2016).

In their research, Wang et al. (2016) opines that several countries in the sub-Saharan Africa experienced considerable gains in life expectancy in the year 2005 to 2015, recovering from a period of tremendously high loss of lives attributable to HIV/AIDS. Concurrently, many regions experienced declining life expectancy, especially for males and in countries with increasing deaths as a result of war or interpersonal violence. In the period 2005 to 2015, life expectancy for men in Syria declined by 11.3 years (3.7–17.4), to 62.6 years (56.5–70.2).

Studies have found a number of attributes to be crucial in the approximation of life expectancy. Ketenci and Murthy (2018) established the real per capita income and educational attainment levels to be the major determinants affecting the level of life expectancy in the United States. In their study to gauge factors associated with life expectancy in countries with low and medium human development index, Girum et al. (2018) concludes that policy and programs earmarked for improvement of LE ought to account for population dynamics, socioeconomic influence, and health system factors. Their research findings demonstrated that LE is influenced by multiples of socioeconomic factors, health and health care system associated attributes, disease burden, and their complex interactions.

An analysis of the socioeconomic factors by Martín Cervantes et al. (2019) in Spanish regions revealed an association of causality with the life expectancy at birth. Their results showed that, in accordance with the Granger causality test for panel data (Dumitrescu–Hurlin version), hospital beds per 1000 occupants, medical and nursing staff in both specialized and primary care per 1000 occupants, and per capita income caused LE at birth. In a research conducted by Wang and Ren (2019) in China, spatial regression disclosed that income per capita had a positive influence on life expectancy at birth.

According to the study results published by Martin Cervantes et al. (2020), per capita income, the population's educational level, region, and the public expenditure on environmental and social protection, were found to have the highest applicability in explaining life expectancy at birth in Europe over the years 2008–2017. The authors established too that green-house gas (GHG) emissions and public healthcare expenditure had the least relevance to LE at birth.

The findings from a study in Malaysia by Tafran et al. (2020) reveal that poverty and income considerably influence female, male, and total life expectancies. Similarly, unemployment remarkably influences female and total life expectancies, but not male. On the other hand, the study establishes that income inequality and public spending on health (as a percentage of total health expenditure) do not significantly influence life expectancy.

As evidenced in the findings of Nkalu and Edeme (2019) undertaken in Nigeria, carbon dioxide emissions from solid fuel consumption were found to reduce life expectancy by 1 month and 3 weeks. In this study too, population growth and income were found to extend life expectancy by 5 years 5 months and 1 year six months, respectively.

In a study that considered 16 Sub-Saharan African economies conducted by Shahbaz et al. (2019), economic growth, globalization and financial development were revealed to have a positive contribution to life expectancy in these economies except for Gabon and Togo. Analysis by Nketiah-Amponsah (2019) further revealed that, a 1 percent increase in health spending per capita saw a 0.5 percent drop in under-five deaths and a 0.35 percent reduction in maternal mortality, improving life expectancy by 0.06 percent in sub-Saharan Africa.

Research by Freeman et al. (2020) highlights the range and importance of social determinants of health, on life expectancy in Ethiopia, Brazil and USA. The results from this study indicates that in Ethiopia, community based health strategies, improvement in access to safe water, gender empowerment and female education, and the increase of civil society organizations improved life expectancy by 3 years. Similarly, in Brazil, the study revealed that life expectancy improved by 2 years attributable to socio-political and economic enhancements, decreased inequality, female education, health care coverage, civil society, and political participation.

Results from the research conducted in 5 EU accession candidate countries by Miladinov (2020) revealed that, life expectancy at birth was notably influenced by the population health and socioeconomic development of a country. Higher rates of GDP per capita and lower infant mortality levels were found to positively influence life expectancy at birth in Macedonia, Serbia, Bosnia and Herzegovina, Montenegro, and Albania. The study also established causality running one-way from life expectancy at birth to infant mortality rate.

2.3 Machine Learning

Machine learning (ML) has been defined in a number of ways by different authors in literature. El Naqa and Murphy (2015) defined machine learning as "evolving branch of computational algorithms that are designed to emulate human intelligence by learning from the surrounding environment". In his book, Alpaydin (2020) defined machine learning to be "programming computers to optimize a performance criterion using example data or past experience".

Machine learning is comparable to statistics in that, it is employed in data analysis and interpretation. In contrast to statistics however, ML algorithms can apply binary logic (TRUE, /FALSE, AND, OR, NOT), absolute conditional arguments (IF-THEN-ELSE), conditional probability and non-traditional optimization approaches to model relationships present in data. Nevertheless, ML borrows profoundly from probability and statistics, although it is essentially more robust since it permits inferences to be made that could otherwise not be created, utilizing conventional statistical techniques (Duda et al., 2001; Mitchell, 1997).

Machine learning is grouped into four categories: supervised learning, unsupervised learning, semi-supervised learning and reinforcement learning. Supervised learning entails building a mathematical model for prediction of the outcomes for future observations Kassambara (2018), (e.g., classification and regression). In unsupervised learning on the other hand, only input samples are given to the learning algorithm (e.g., clustering and estimation of probability density function) (El Naqa and Murphy, 2015).

Semi-supervised learning combines both supervised and unsupervised learning where part of the data is partially labeled, with the labeled part being used to infer the unlabeled portion (e.g., text/image retrieval systems) (El Naqa and Murphy, 2015). Reinforcement learning involves the finding of suitable actions to take in a given situation in order to maximize a reward. In this type of learning, examples of optimal outputs are not given to the algorithm, but instead, they are discovered through trial and error (Bishop, 2006).

Machine learning algorithms have been applied in a wide range of fields to offer feasible solutions to existing problems. A study undertaken by Vamathevan et al. (2019) found the application of ML approaches to have the potential to promote data-driven decision making, and speed up the process of drug discovery and development, reducing the failure rates. Besides its ability to improve their basic understanding of cancer development and progression, in their review, Cruz and Wishart (2006) established that ML techniques could be employed to substantially $(15-25\%)$ improve the prediction accuracy of cancer susceptibility, recurrence and mortality.

Valletta et al. (2017) presented three case studies to showcase the use of ML in the development of data analytical workflows answering biological questions. To begin with, they use a large number of spectral and morphological attributes that detail the appearance of pheasant, *Phasianus colchicus*, eggs to assign them to putative clutches. Secondly, foraging events are drawn from a continuous data stream of feeder visits out of PIT (passive integrated transponder)-tagged jackdaws, *Corvus monedula*, allowing the building of social networks.

Thirdly, aerial images are used to train a classification algorithm that detects the existence of wildebeest, *Connochaetes taurinus*, to count individuals in a population. The authors conclude that ML would play a crucial role in translation of complex data sets into scientific mastery, becoming a useful inclusion to the animal behaviorist's analytical set of tools.

Clancy et al. (2007) applied ML algorithms to cognitive radio and developed a framework from within which the ML algorithms could be useful. The authors try to formalize a number of architectures on the capacities put forth for cognitive radios and the applications they greatly befit, giving intuition into the dissimilarities amid reasoning and learning. They implement a cognitive radio based on a non-specific cognitive engine.

The radio achieves the optimal configuration of modulation and coding in an environment with time-varying noise power.

Recent studies have found ML techniques to be handy in modelling the novel corona virus (n-cov) disease. Due to their simple and easily understandable nature, Alballa and Al-Turaiki (2021) found that with the use of readily available clinical and laboratory data, supervised ML algorithms have gained popularity in the diagnosis of COVID-19, and the prediction of mortality and severity risks.

The study by Arpaci et al. (2021) retrospected 114 samples from the Taizhou hospital of Zhejiang Province in China. The authors discovered ML predictive models used in their study had the potential of significantly contributing to the early diagnosis of COVID-19, specifically in cases where RT-PCR kits were not sufficient for COVID-19 infection testing.

In their research, Tuli et al. (2020) applied a ML-based improved model in predicting the potential threat posed by COVID-19, among countries globally. The authors present a prediction model deployed via a blockchain-based lightweight framework for edge and fog computing. The proposed Robust Weibull model based on iterative weighting, is reported to output statistically finer predictions over the baseline Gaussian model that exhibits an over-optimistic impression of the COVID-19 scenario.

Lately, novel machine learning approaches have revealed improved predictive performance in comparison to other conventional prediction techniques. The subsequent section highlights a number of studies that have applied the novel Extreme Gradient Boosting (XGBoost) algorithm to various industry problems.

2.4 Application of XGBoost Algorithm in ML Predictions

eXtreme Gradient Boosting (XGBoost) has extensively been recognized in a number of data science challenges for its outstanding predictive performance (Nielsen, 2016). Besides its credit in winning ML competitions, the XGBoost algorithm has gained popularity for various cutting-edge industry usage.

In the gene expression value prediction study by Li et al. (2019), the XGBoost model attained a remarkably lower overall error than the deep learning gene expression (D-GEX), linear regression, and KNN algorithms. The linear regression model achieved an overall error of 0.378 on both the validation set and the test set. On introduction of the L1 and L2 regularization terms, linear regression model recorded overall error rates of 0.377 and 0.378, and 0.378 respectively, on the validation set and the test set, respectively. The KNN model had an error rate of 0.586 on the validation set and 0.587 on the test set.

The D-GEX model reported slightly lower errors than the former three models at 0.312 and 0.320 for the validation and test sets, respectively. The best overall error results on the validation set and test set were achieved by the XGBoost algorithm at 0.280 and 0.282, respectively. In this research Li et al. (2019), the number of landmark genes was found to be large, resulting to high dimensionality of the input features. This characteristic made the models used susceptible to over-fitting.

For the deep network of D-GEX, the researchers acknowledge that not only was the input dimension very high, but also, the output dimension was even higher. It was hence gruelling to train a very accurate model, and the processing of parameter adjustment was tremendously complicated in addition. Besides, poor interpretability was also a stumbling block of the deep network (Li et al., 2019).

In contrast, the control of the model complexity was added for the XGBoost method. This made the resultant model barely likely to overfit, hence boosting the generalization aptness of the model, and eventually significantly lessening the predictive errors. Additionally, XGBoost was more attentive to the interpretability of the model, consequently, landmark genes with considerable effect on the expression value of each target gene, were learnt. Moreover, the competitive feature for parallel computing as highlighted by Wade (2020) rendered the XGBoost model speedy than traditional tree models, with a higher practical value.

Li and Zhang (2020) proposed an orthopedic auxiliary classification prediction method ground on XGBoost algorithm. In the research, the random forest and the associated classification algorithms are compared with XGBoost algorithm.

In comparison of the predictive performance of the three models, the RF model had an accuracy rate of 0.645, a recall rate of 0.572, and a learning time of 36.4 seconds. The associated classification model reported an accuracy rate of 0.737, a recall rate of 0.659, and a corresponding learning time of 39.7 seconds. XGBoost on the other hand attained an accuracy rate, a recall rate and an associated learning time of 0.951, 0.928 and 6.3 seconds, respectively.

The XGBoost model achieved higher accuracy and recall rates in the the classification and prediction of orthopedic diseases. Over and above that, the results show a clear superiority of the XGBoost model over the other two models in terms of the computational speed. It is observed that the XGBoost algorithm could potentially process medical data reliably and rapidly, coinciding with the features of medical data and medical diagnosis.

Wang et al. (2020) proposes a machine learning framework for the prediction of earthquake-induced Newmark sliding displacements of slopes employing the XGBoost model. Three data-driven Newmark displacement models are built via different vector intensity measures (IMs), namely: the XGBoost-peak ground acceleration-peak ground velocity (PGA, PGV), XGBoost-peak ground acceleration-arias intensity (PGA, I_a), and XGBoost-peak ground acceleration-peak ground velocity-arias intensity (PGA, PGV, I_a) models.

The performances of the aforementioned XGBoost models are compared with the Saygili and Rathje (2008) models referred to as SR08 models, and the Updated SR08 models similarly built via the (PGA, PGV), (PGA, I_a) and (PGA, PGV, I_a) intensity measures. The SR08 model attained R^2 values of 0.934, 0.901 and 0.948 for the (PGA, PGV), (PGA, I_a), and (PGA, PGV, I_a) metrics, respectively. The MAE values for the SR08 model were 0.529, 0.651 and 0.447 for the (PGA, PGV), (PGA, I_a), and (PGA, PGV, I_a) metrics, respectively.

The corresponding RMSE of the model was 0.693, 0.849 and 0.606 for the (PGA, PGV), (PGA, I_a), and (PGA, PGV, I_a) metrics, respectively. Similarly, the Updated SR08 model attained R^2 values of 0.936, 0.903 and 0.949 for the (PGA, PGV), (PGA, I_a), and (PGA, PGV, I_a) metrics, respectively. The MAE and RMSE values for the Updated SR08 model were 0.517, 0.647 and 0.441, and 0.675, 0.834 and 0.597 for the (PGA, PGV), (PGA, I_a), and (PGA, PGV, I_a) metrics, respectively.

The XGBoost (PGA, PGV), XGBoost (PGA, I_a), and XGBoost (PGA, PGV, I_a) achieved R^2 values of 0.961, 0.955 and 0.972, respectively. The attained MAE values for the three XGBoost models in the aforementioned order were 0.391, 0.417 and 0.319, respectively. Following the same model order, the reported RMSE values were 0.524, 0.561 and 0.444, respectively.

Additionally, Wang et al. (2020) asserts that the built XGBoost model is extra pliable in capturing high non-linearity embedded in the dataset. Furthermore, it is established that peak ground acceleration (PGA) is the most paramount intensity measure (IM). Moreover, the authors endorse XGBoost (PGA, PGV) and XGBoost (PGA, PGV, I_a) models due to their greater generalization proficiency and robustness.

Hou et al. (2020) on the other hand established that ML based on XGBoost algorithm outperformed conventional logistic regression and SAPS-II score model in the prediction of 30-days mortality for MIMIC-III patients with sepsis-3. The three models manifested good segregation ability with area under curves (AUCs) of 0.797 (95% CI 0.781–0.813), 0.819 (95% CI 0.800–0.838), and 0.857 (95% CI 0.839–0.876), for SAPS-II score model, traditional logistic regression model and XGBoost model, respectively.

The XGBoost model exhibited the largest test AUC but the SAPS-II score model had the smallest test AUC. In accordance with the decision curve analysis (DCA) of the three prognostic models, the net benefit for XGBoost model was greater over the range of traditional logistic model and SAPS-II score model. The insight from these findings rendered the XGBoost model optimal and the SAPS-II score model inferior. These results too are in favour of the XGBoost model.

Abdu-Aljabar et al. (2021) reported XGBoost classification algorithms to have enhanced the prediction of lung cancer diagnosis detection and relapse prediction over SVM and gcForest. The algorithms are applied on six different datasets from two types of gene expressions (microarray and new generation sequence (NGS)). On the first dataset (GSE81089), XGBoost, SVM and gcForest achieved accuracy scores of 0.997, 0.971, and 0.988 respectively. The associated standard errors were 0.006, 0.017 and 0.011 for XGBoost, SVM and gcForest models, respectively.

For the second dataset (GSE30219), XGBoost, SVM and gcForest achieved accuracy scores of 0.995, 0.969, and 0.978 respectively. The associated standard errors were 0.016, 0.021 and 0.0151 for XGBoost, SVM and gcForest models, respectively. Similarly, in the third dataset (GSE19188), XGBoost, SVM and gcForest reported accuracy scores of 0.957, 0.963, and 0.979 respectively.

The corresponding standard errors were 0.014, 0.020 and 0.015 for XGBoost, SVM and gcForest algorithms, respectively. GSE8894 dataset revealed accuracy scores of 0.794, 0.537, and 0.585 for XGBoost, SVM and gcForest algorithms, respectively. The resultant standard errors for XGBoost, SVM and gcForest models were 0.061, 0.061, and 0.064, respectively.

Likewise, for the GSE68219 dataset, XGBoost, SVM and gcForest revealed accuracy scores of 0.632, 0.615, and 0.606 respectively. The corresponding standard errors were 0.026, 0.028 and 0.043 for XGBoost, SVM and gcForest models, respectively. Additionally, this study found XGBoost to have the least learning time for most of the datasets. gcForest on the other hand, had the longest time in comparison to SVM and XGBoost models.

According to the established findings, Abdu-Aljabar et al. (2021) also notes a number of pros and cons of the XGBoost model. XGBoost is reported to: be extremely fast due to its parallel computation ability, exceptionally efficient in both balanced and imbalanced datasets, versatile owing to its use in regression, classification or ranking tasks, and to have abolished the need for feature engineering due to its ability to handle missingness through imputation.

Aside from the aforementioned pros, on the contrary, the authors admit XGBoost to have only worked with numeric features in their study. Moreover, if the algorithm's hyper-parameters are inappropriately tuned, the model could suffer from overfitting.

Similarly, Zhang et al. (2021) found the XGBoost model to have higher accuracy and stronger robustness in comparison to the LASSO-logistic model in the prediction of acute kidney injury (AKI) in patients with hepatobiliary malignancies. Specifically, the study established the Youden's index (the sum of sensitivity and specificity minus one) of the XGBoost model to be outstandingly higher than that of LASSO-logistic model at 47.5% and 59.3%, respectively.

The LASSO-logistic model had a Youden's index of 41.6% and 32.7%, respectively. Both models had $p < 0.001$.

The area under the curve (AUC) for the XGBoost model was 0.822 (95% CI 0.789-0.855) in liver cancer and 0.850 (95% CI 0.775-0.920) in gallbladder cancer. The LASSO-logistic model in contrast reported lower AUC values of 0.793 and 0.740 for liver and gallbladder cancers, respectively. The difference between the two models was found to be statistically significant as verified by DeLong's test ($p = 0.024$ and 0.018). Furthermore, with the buildup of training samples, XGBoost model was found to sustain greater robustness in the learning curve.

2.5 Prediction of Life Expectancy Using Machine Learning Algorithms

A number of studies exist in literature on the estimation of life expectancy with various approaches being employed. Chen and Cheng (2006) considered a linear mean residual life model, developing inference procedures in the presence of potential censoring. The study carried out simulations to assess the finite sample properties of the proposed methods. In this paper, details on the illustration of the efficiency of the proposed approaches are scanty. This calls for more extensive simulation studies to evaluate the efficiency of these methods.

Shang (2012) employed a model averaging approach to predict age-specific life expectancy. The authors argue that by combining and averaging different model results, one could easily acquire better accuracy when the combined outputs employ varied techniques capturing distinct information, specifications or assumptions. They compare fourteen methods for projecting age-specific LE.

The point and interval predictions for LE are evaluated based on ten principal component methods, two random walk approaches and two uni-variate time series techniques using age and sex-specific LE of fourteen developed nations. The findings from the point estimates indicate that the Lee-Miller (LM) method had the least possible bias among all the other

methods, followed closely by the Booth-Maindonald-Smith (BMS) and Random-walk with drift (RWD) methods when applied to the female data.

On the other hand, with the exception of the Lee-Carter (LC) method, all the other methods were found to consistently underestimate the life expectancies when applied to the male data. Contrasted with the LC method, the LM and RWD were too concentrated about zero across all ages. The RWD method exhibited the most precise point forecasts for the one-year-ahead LEs.

For interval forecasts, the prediction intervals are overly slim for all the LC methods (including the BMS), underestimating the LE variability. Conversely, substantially broader prediction intervals consistently arise from the Weighted Hyndman-Ullah (HUw), Random-walk (RW) and RWD methods, giving rise to inferior forecast accuracy. The LM, Hyndman-Ullah (HU), Robust Hyndman-Ullah (HUrob) and the univariate time-series methods all manifest a combination of underestimation and overestimation of the coverage probability for several countries.

For both datasets, the authors assert that HU methods outperform their LC counterparts in terms of the interval forecasts' accuracy. On the flip side, there is scanty to no information about the actual model evaluation metric values used in this study. Moreover, the comparative investigation in this research is limited to techniques mainly ground on principal component approaches and uni-variate methods.

Additionally, a contrast of the applied approaches with other prediction techniques is beyond the scope of this paper. The conclusions from this study are based on age and sex-specific simple and weighted averages across fourteen countries. Some methods perform well for certain countries, but could probably exhibit an overall poor performance. Similarly, a technique could likely perform best overally, but poorly for a specific country.

Raftery et al. (2013) proposed a Bayesian hierarchical model (BHM) for probabilistic projections of life expectancy for all the countries in the globe. The model's resultant performance was compared with the UN methodology. The 10-year out-of-sample predictions for the BHM model attained a mean absolute error (MAE) and a standard absolute prediction

error (SAPE) of 1.07 and 1.04, respectively outmatching the UN approach that had initially reported a MAE of 1.86.

A number of limitations arise from this study. To begin with, the study does not take into account the countries with generalized HIV/AIDS epidemic. Secondly, the scope of this study is solely based on forecasting of life expectancy for males. Additionally, the Bayesian approach has the disadvantage of requiring a prior distribution even when scanty external information is at one's disposal.

A further study undertaken by Raftery et al. (2014) as an extension and remedy to some of the limitations encountered in Raftery et al. (2013), proposed a methodology to obtain joint probabilistic predictions of life expectancy at birth for both genders. The study gauged the calibration, precision and accuracy for model validation.

The performance results calculated for a 15-year out-of-sample cross-validation for four different projection models showed that for the gap between female and male life expectancy at birth, the UN Simulated model attained a MAE of 0.78. The Constant 1995 Gap model achieved a MAE value of 0.73. The BHM obtained a MAE of 0.78. The Male e_0-based Model and the Female e_0-based Model attained MAE values of 0.70 and 0.66, respectively.

For the male life expectancy at birth, the UN Simulated model attained a MAE of 2.32. The Constant 1995 Gap model achieved a MAE value of 1.42. The BHM Projection and the Gap-based Model each reported a MAE of 1.44 and 1.32, respectively. Similarly, for the female life expectancy at birth, the UN Simulated model attained a MAE of 1.94. The Constant 1995 Gap model achieved a MAE value of 1.14. The BHM Projection and the Gap-based Model reported MAE values of 1.16 and 1.17, respectively.

The authors established that the Female e_0-based Model and the Gap-based Model exhibited satisfactory coverage, precision, and gave rise to the most accurate forecasts among the considered models. Conversely, the BHM estimation of male life expectancy attained good coverage with low precision. In addition, it was 11-18% less accurate in comparison to the other two models.

Equivalently, this study too was prone to a number of limitations. Countries with a generalized HIV/AIDS epidemic and those with a population size of 100,000 or less were excluded in the study. The choice of covariates used in the study too is a limitation. The study does not take into account the socioeconomic factors, health behaviors and biological indicators that affect a population's lifespan.

Dias et al. (2017) assessed the impact of sex, death cause, profession and race on life expectancy based on general linear models (GLM) and Kaplan-Meier estimates in Colombo district, Sri Lanka. In this study, the findings indicate that a univariate GLM resulted in a model that could be used in the prediction of the life time for an individual.

The independent samples' t test conducted for the gender variable indicated that at $\alpha = 0.05$, the mean life time of males was found to be insignificantly different from that of the females. A Kruskal–Wallis test on the age distribution across gender suggested that, at $\alpha = 0.05$, the age distribution was not identical for the males and females, implying that an individual's gender had an effect on their lifespan.

Furthermore, the study established that an individual's racial background, cause of death and profession had no effect on their lifespan (Kruskal–Wallis test, $\alpha = 0.05$). Results from the Kaplan-Meier estimates indicated the mean survival times of the males and females to be 76.66 (95% CI: 72.92-80.39, SE $\pm$1.91) and 81.61 (95% CI: 78.84-84.38, SE $\pm$1.41), respectively. The existence of an overlap in the confidence intervals of both genders suggested an unlikely difference in the mean survival times for the two genders.

The corrected generalized linear model attained R^2 and adjusted R^2 values of 0.255 and 0.197, respectively. The associated survival curves for both genders further revealed that, as one ages, their probability of survival decreases regardless of gender. The cumulative hazard plot suggested an increase in the probability of departure from life as individuals aged, for both genders.

The study focused on gender, cause of death, race and profession as the sole predictors. The scope of the study did not consider health, behavioral and other socioeconomic factors that affect a population's health and ultimately, life expectancy. Notably, the considerably lower

values of the R^2 and adjusted R^2 suggest that the fitted model was not a good fit, an indication of underfitting.

A potential opportunity for improvement to this research would call for the enhancement of its scope in terms of the predictors to account for health, behavioral and other socioeconomic features. Secondly, additional of other relevant important life expectancy features and/or use of a more robust model would come in handy to remedy the problem of underfitting.

Karacan et al. (2020) built a decision tree using a Chi-square automatic interaction detector (CHAID) technique, that helped to come up with influential variables on life expectancy at birth. The study established that of the 25 features, percentage of the population using improved sanitation facilities; death rates per 100,000 population for HIV/AIDS, liver disease, stroke and coronary heart disease (CHD); percentage of the urban population using improved drinking-water sources; total fertility rate as births per woman; public health expenditure as a percent of government expenditure; and health expenditure as US$ per capita were the most important variables influencing life expectancy at birth.

The research by Karacan et al. (2020) differs from the present study in that, the authors' main goal was to identify the robust predictors for life expectancy at birth. This research is also limited in the sense that only predictors relating to demography, economy, health and environment, mortality, and child health were considered. The authors acknowledge this as a limitation since other predictors (e.g., employment by level of education and educational level) that were not included in their study, were likely to have an effect on life expectancy at birth. The study also made use of cross-sectional data without regard for sex segregation. This would call for development of independent decision trees for both genders, resulting into distinct classification decision rules. The present study put forth in this work focuses on ML prediction of life expectancy at birth, taking into account health, behavioral and socioeconomic factors.

Meshram (2020) presented a comparative analysis of life expectancy between advanced and developing economies using machine learning. In this research, the forecasting algorithm was trained via the linear regression (LR), decision tree (DT) and the random forest (RF) regressors. For all the three models, the dataset used is partitioned into a training and a testing

set using a split ratio of 80:20% respectively. The adopted model was arrived at through evaluation of the R^2, MSE and the MAE scores. The authors settled on the RF regressor that attained R^2 scores of 0.99 and 0.95 on the train set and test set, respectively.

The MSE and the MAE for the RF regressor were 4.43 and 1.58, respectively. These evaluation scores were reported to be superior in comparison to those registered by the LR and the DT regressor models. However, the details of the model evaluation results for the LR and DT regression models remain undisclosed. The study's findings revealed that adult mortality, infant deaths, measles, under-five deaths, HIV/AIDS, thinness 10-19 and 5-9 years had a negative correlation with Life Expectancy.

Contrastingly, the percentage expenditure on healthcare, BMI, GDP, income composition of resources and schooling had a positive correlation with life expectancy. The feature importance plot associated with the RF regression model suggested that adult mortality, HIV/AIDS deaths, income composition of resources, schooling, percentage health expenditure, prevalence of thinness, and BMI were the key variables having a considerable impact on a country's life expectancy.

The RF regressor algorithm was found to produce the best model for the prediction of life expectancy in comparison to its linear and decision tree counterparts. Nevertheless, a limitation of the RF algorithm is that in most real-life applications, the RF algorithm has been found to be fast to train, but quite slow in predictions after training. The RF regressor's main drawback is that it may suffer ineffectiveness and slowness for real-time predictions for scenarios where large number of trees are involved (Donges, 2021).

In a study on the prediction of life expectancy in Maluku province, Indonesia, Lesnussa et al. (2020) applied the backpropagation artificial neural network approach. The study achieved an average accuracy of 99.65%, with a mean absolute percentage error (MAPE) of 0.0035 and a corresponding learning rate value of 0.1. However, the model of this research was evaluated based on five (5) and two (2) training and testing samples, respectively. This amount of data is considerably small for training neural networks. More data would be desirable for increased model accuracy and reduced overfitting.

Owing to the aforementioned gaps and the inherent limitations of the reviewed literature, a robust and more accurate approach to life expectancy estimation is inevitable. Based on the reviewed empirical studies, the author did not come across any study that applied XGBoost to predict life expectancy at birth considering health, socioeconomic and behavioral factors. Hence, this study.

2.6 Machine Model

The modelling process for this study was as presented in Figure 2.1. The raw data was first pre-processed to transform it into a useful and efficient format, then partitioned into train and test sets in an 80:20% ratio, respectively and subsequently the model was built and optimized. The resultant model was validated and ultimately used to predict life expectancy rates on new data.

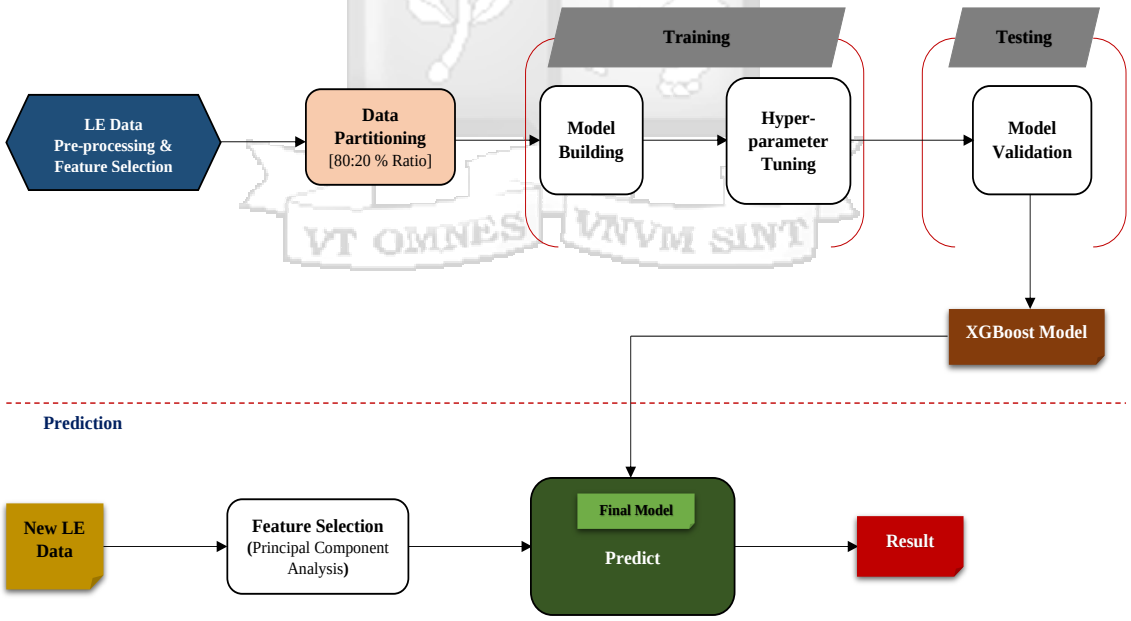


Figure 2.1: A Flowchart Depicting Various Steps of the Modelling Process.

Chapter 3

Research Methodology

3.1 Introduction

In a supervised machine learning challenge, given a set of data points $\{(x_i, y_i)\}_{i=1,\ldots,n}$, where x_i are the predictors and y_i the response, a feasible solution to the function $y = f(x)$ is arrived at through estimating $\hat{f}(x)$. An evaluation of the mapping's quality is made using the loss function $L(y, f)$, and the final loss is then calculated by taking the average of all feasible data points in the dataset.

3.2 eXtreme Gradient Boosting

eXtreme Gradient Boosting, abbreviated as XGBoost is a decision-tree based ensemble machine learning algorithm that uses a gradient boosting framework proposed by Chen and Guestrin (2016) in 2015. It is a novel classification and regression problems implementation algorithm frequently applied due to its rapidness, efficiency and scalability (Wang and Ni, 2019).

3.2.1 Mathematical Setting of XGBoost

In gradient boosting, the function that determines the i^{th} row prediction consists the sum of all previous functions. Suppose there are K boosted trees, mathematically, the XGBoost model will be in the form:

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i),\ f_k \in F, \tag{3.1}$$

where $\hat{y}_i$ is the predicted value by the ML model for the i^{th} row, K is the number of boosted trees, x_i is the i^{th} data point (a vector whose entries are the columns of the i^{th} row), f is a function in the functional space F, and F is the set of all possible Classification And Regression Trees (CARTs).

Similarly, the k^{th} boosted tree prediction will be defined as:

$$\hat{y}_i^{(k)} = \sum_{k=1}^{K} f_k(x_i). \tag{3.2}$$

The general architecture of the XGBoost algorithm is as shown in Figure 3.1 below.

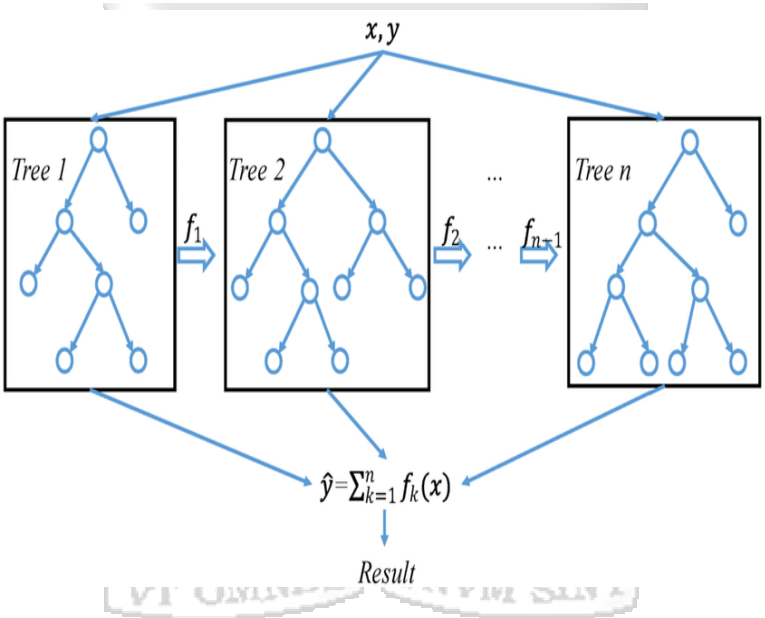


Figure 3.1: A General Architecture of the XGBoost Algorithm.

The trees are trained by defining an objective function and optimizing it. The objective function to be optimized for the k^{th} boosted tree is given by:

$$Obj^{(k)} = \sum_{i=1}^{N} L(y_i, \hat{y}_i^{(k)}) + \sum_{k=1}^{K} \Omega(f_k), \tag{3.3}$$

$L(y_i, \hat{y}_i^{(t)})$ is the training loss function (i.e., MSE) of the k^{th} boosted tree and $\Omega(f_k)$ is the i^{th} regularization term, a penalty term to prevent over-fitting (Wade, 2020). The loss and

regularization functions are derived based on Chen and Guestrin (2016) to come up with the learning objective function as shown in the following subsections.

Loss Function of XGBoost

Since gradient-boosted trees sum the predictions of the previous trees, in addition to the prediction of the new tree, it follows from equation 3.2 that:

$$\hat{y}_i^{(k)} = \hat{y}_i^{(k-1)} + f_k(x_i), \tag{3.4}$$

which is the idea behind additive training. Substituting equation 3.4 into the preceding learning objective in equation 3.3, we obtain:

$$Obj^{(k)} = \sum_{i=1}^{N} L\left(y_i, \hat{y}_i^{(k-1)} + f_k(x_i)\right) + \Omega(f_k). \tag{3.5}$$

The above equation can be re-written as follows:

$$Obj^{(k)} = \sum_{i=1}^{N} \left[y_i - (\hat{y}_i^{(k-1)} + f_k(x_i))\right]^2 + \Omega(f_k). \tag{3.6}$$

Multiplying the polynomial out, the objective function becomes:

$$Obj^{(k)} = \sum_{i=1}^{N} \left[2\left(y_i - \hat{y}_i^{(k-1)}\right) f_k(x_i) + f_k(x_i)^2\right] + \Omega(f_k) + C, \tag{3.7}$$

where C is a constant term independent of k. The goal was to find the optimal value of $Obj^{(k)}$, the optimal function mapping the samples (roots) to the predictions (leaves).

XGBoost uses the Newton Rhapson's Method with a second-order Taylor expansion to get the following:

$$Obj^{(k)} = \sum_{i=1}^{N} \left[g_i f_k(x_i) + \frac{1}{2} h_i f_k(x_i)^2\right] + \Omega(f_k), \tag{3.8}$$

where g_i and h_i are the first and second partial derivatives of the loss function, respectively defined as:

$$\begin{aligned} g_i &= \frac{\partial l(y_i, \hat{y}_i)}{\partial \hat{y}_i} \\ &= \frac{y_i}{\hat{y}_i} - \frac{1-y_i}{1-\hat{y}_i} \\ &= \frac{y_i(1-\hat{y}_i) - \hat{y}_i(1-y_i)}{\hat{y}_i(1-\hat{y}_i)} \\ &= \frac{y_i - y_i\hat{y}_i - \hat{y}_i + y_i\hat{y}_i}{\hat{y}_i(1-\hat{y}_i)} \\ &= \frac{y_i - \hat{y}_i}{\hat{y}_i(1-\hat{y}_i)}, \end{aligned} \tag{3.9}$$

$$\begin{aligned} h_i &= \frac{\partial^2 l(y_i, \hat{y}_i)}{\partial \hat{y}_i^2} = \frac{\partial}{\partial \hat{y}_i} g_i \\ &= \frac{\partial}{\partial \hat{y}_i} \left[\frac{y_i - \hat{y}_i}{\hat{y}_i(1-\hat{y}_i)} \right] \\ &= \frac{y_i - 1}{(\hat{y}_i - 1)^2} - \frac{y_i}{\hat{y}_i^2}. \end{aligned} \tag{3.10}$$

Regularization Function of XGBoost

Having introduced the training step, the complexity of the tree $\Omega(f_k)$, is defined. Let w be the vector space of leaves. Then f, the function mapping the root of the tree to the leaves can be given a different form in terms of w as follows:

$$f_k(x) = w_{q(x)},\ w \in R^T,\ q : R^d \to \{1, 2, \ldots, T\}, \tag{3.11}$$

where w is the vector of scores on leaves, q is the function assigning each data point to the corresponding leaf, and T is the number of leaves.

In XGBoost, the regularization term is defined as highlighted in Equation 3.12, where γ and λ are penalty constants to reduce overfitting (Wade, 2020):

$$\Omega(f_k) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2. \tag{3.12}$$

Combining the loss function with the regularization term yields the learning objective function as:

$$Obj^{(k)} = \sum_{i=1}^{N} \left[g_i w_{q(x_i)} + \frac{1}{2} h_i w_{q(x_i)}^2 \right] + \gamma T + \frac{1}{2} \lambda \sum_{j=1}^{T} w_j^2 , \tag{3.13}$$

which is the result XGBoost uses to determine how well the model fits the data.

3.2.2 Design Features

XGBoost is a cutting-edge implementation of the gradient boosting algorithm. The key features discussed below were derived from (Chen and Guestrin, 2016; Wade, 2020). They distinguish XGBoost from gradient boosting and other tree ensemble methods.

a) Handling Missingness: it has the ability to learn missing values during training using its in-built hyperparameter called *missing*.

b) Speed Gains: XGBoost was designed for speed. This feature allows the algorithm to build models more promptly. The XGBoost's competitive edge in speed is as a result of:

   i) Approximate split-finding algorithm: a greedy algorithm selects the best split at each step.

   ii) Sparsity-aware split finding: when searching for splits, XGBoost is faster since its matrices are scattered.

   iii) Parallel computing: when multiple computational units work together concurrently on a similar problem, XGBoost sorts and compresses the data into blocks. It then provides parallel computing to hasten the model-building process.

   iv) Cache-aware access: XGBoost uses cache-aware prefetching for gradient statistics. It allocates an internal buffer, fetches the gradient statistics, performing accumulation with mini batches.

   v) Block compression and sharding: the algorithm achieves extra gains in speed via block compression and block sharding. Block compression assists with

computationally expensive disk reading by compacting columns. Block sharding reduces read times by sharding the data into multiple disks that alternate when reading the data.

c) Accuracy Gains: XGBoost incorporates regularization as part of its learning objective, which penalizes the model complexity smoothing out the final weights to avert overfitting.

3.2.3 XGBoost Hyperparameters

In machine learning, a hyperparameter is a parameter whose value controls the learning process and determines the values of model parameters that a learning algorithm ends up learning (Nyuytiymbiy, 2020). Table 3.1 gives a comprehensive list of the XGBoost hyperparameters.

Table 3.1: XGBoost Hyperparameters with their Corresponding Usage and Effect.

Hyperparameter	**Default**	**Range**	**Usage/Note**	**Effect**
n_estimators	100	[1,inf)	Provides the number of trees to grow.	Increasing may improve scores with large data.
eta	0.3	[0, inf)	Shrinks the weights of trees for each round of boosting to prevent overfitting.	Decreasing avoids overfitting.
max_depth	6	[0, inf)	Determines the maximum length of a tree. Equivalent to the number of rounds of splitting.	Limiting prevents overfitting.

Hyperparameter	Default	Range	Usage/Note	Effect
gamma	0	[0, inf)	Minimum loss reduction required to make a further partition on a leaf node of the tree.	Increasing prevents overfitting.
min_child_weight	1	[0, inf)	The minimum sum of weights required for a node to split into a child.	Increasing avoids overfitting.
subsample	1	(0, 1]	Subsample ratio of the training instances. Setting it to 0.5 implies that XGBoost would randomly sample half of the training data prior to growing trees.	Decreasing avoids overfitting.
colsample_bytree	1	(0, 1]	Subsample ratio of columns when constructing each tree. Subsampling occurs once for every tree constructed.	Lowering makes the model to generalise better.

3.2.4 Tree Splitting Mechanism

This study used decision trees as the base learners for the model. XGBoost builds each new tree from the previous one. The algorithm then proceeds to fit each new tree entirely based on the errors of the previous tree's predictions. Each tree is split using a greedy algorithm that selects the best split at each step, without backtracking to look at previous branches. Besides, XGBoost employs an approximate split-finding algorithm to split each

tree to produce optimal predictions. The latter algorithm uses quantiles to propose candidate splits (Wade, 2020).

3.3 Data Splitting

Data splitting is a routinely employed technique for model validation where a given dataset is partitioned into two disjoint sets: training and testing. The proposed model in this study was applied to a real-life dataset. The dataset was randomly split into a training and testing set following the commonly applied ratio of 80% and 20% respectively (Joseph, 2022; Wang et al., 2020). The training set was used for the model development and the hold-out (testing) set employed for model validation.

Automatic random selection was used to avoid the introduction of potential figure selection bias via use of the caret (Kuhn, 2021) package in R. The study used R version 4.2.0 program from the R Core Team (2022) as the primary software for all the statistical computations and analyses.

3.4 Model Validation Metrics

As Bakas and Kontoleon (2021) recommends, after the model was developed, trained and found to be functional, its performance was evaluated before proposing its use for industry or further scientific research. Due to their commonness in measuring accuracy for continuous variables and their nature of being quick and easy to calculate, the study employed the Root Mean Squared Error (RMSE) and the Mean Absolute Error (MAE) to examine the accuracy of the resultant prediction model.

RMSE is naturally influenced by outliers since squaring the errors in the MSE gives extreme values (usually having higher errors than other samples) more attention and dominance in the final error, thereby impacting the model parameters (Minaee, 2019). MAE has been known

to be more robust to outliers (Kassambara, 2018). To address the limitation of RMSE, the study used the MAE as a complementary measure for model validation.

3.4.1 Root Mean Squared Error (RMSE)

RMSE essentially finds the square root of the average squared error between the actual value and the predicted value by the model. It is a standard way to measure the error of a model in numerical predictions. It is defined as:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{3.14}$$

where, y_i is the i^{th} observed value of the response for the data, $\hat{y}_i$ is the corresponding i^{th} predicted value using the fitted model and the predictors from the data, and n is the number of observations. Lower RMSE values indicate a good fit (Kassambara, 2018).

3.4.2 Mean Absolute Error (MAE)

MAE (or the mean absolute deviation) finds the average absolute distance between the target value and the predicted value. It is defined as:

$$\text{MAE} = \frac{1}{n}\left[\sum_{i=1}^{n}|y_i - \hat{y}_i|\right] \tag{3.15}$$

where, y_i, $\hat{y}_i$, and n are elaborated in sub-section 3.4.1. $|y_i - \hat{y}_i|$ is the absolute error. A lower MAE value is an indication of a good fit.

3.5 Study Design and Population

This study followed a cross-sectional design where information on 21 study variables described in Table 1.1 for the 193 UN member states from the year 2000 to 2015 were obtained.

The study population were the UN member states that had data for all the years from 2000 through 2015. Countries that completely missed data on any particular year were excluded from the study.

As of the year 2022, the Worldatlas states that there were 195 countries in the world. The study population was based on the readily available data for the 193 countries that represent nearly the entire globe. The data partitioning was based on the split ratio discussed in section 3.3.

3.6 Statistical Analysis

In order to convert the raw data into a meaningful and efficient format, data pre-processing was undertaken in the preliminary phase of the analysis. Data was inspected to check for missingness and structure of the variables, and appropriately cleaned. Summary statistics were employed to establish the measures of central tendency. EDA was applied to understand the underlying structure of the dataset and uncover useful insights that were relevant for this study.

Spearman's rank correlation coefficient was used to measure the associations between study variables Sedgwick (2014), and the resultant correlation matrix visualized through a heatmap via the corrplot package in R. A multivariate normality check was carried out using the MVN package in Korkmaz et al. (2014) to ascertain whether the data was drawn from a normally distributed population.

The multivariate normality assumption was violated. Thus, an assessment of whether or not the Regions and Income Groups had different mean vectors across the different study variables was conducted using the Extended Multivariate Kruskal-Wallis (E-MKW) test with missing data put forth by Fanyin et al. (2017) in their paper.

Data for the year 2015 for all the countries was reserved for making predictions. The remaining data was split into two disjoint partitions for the train and test sets by use of the caret package (Kuhn, 2021). The baseline model was initialized using the default XGBoost

hyperparameters. Subsequently, the hyperparameters were fine-tuned via a grid search cross-validation to arrive at an optimal set of hyperparameters used in developing the final XGBoost model. The final model was evaluated on both the training set and the testing set using the metrics in section 3.4. The resultant findings are as documented in chapter 4.

3.6.1 Feature Scaling

Feature scaling (also known as data normalization) is a process to standardize the variables present in a dataset to a constant scale (Mamidanna et al., 2022). Prior to the selection of variables, feature scaling was undertaken on the numeric variables using the Z-Score technique. The formula used for computing the Z-Score of a data point, x, was as follows:

$$x' = \frac{(x - \mu)}{\sigma}, \tag{3.16}$$

where μ and σ are the mean and standard deviation of the feature vector, respectively. The predictors were selected through principal component analysis (PCA) where, the variables with significant contribution to the first principal component were chosen. Since the position of the split point is unaffected by feature scaling in XGBoost, there was no scaling applied to the chosen predictors for the XGBoost model (Sun, 2021).

3.6.2 Missingness

As a preliminary step prior to performing PCA, the missing values in the life expectancy dataset were imputed with the Principal Components Analysis model via the *missMDA* package (Josse and Husson, 2016). *missMDA* imputes the incomplete data set in such a way that the imputed values will not have any weight on the results of PCA. The missing values were predicted using the iterative PCA algorithm with two dimensions being used to predict the missing entries.

3.7 Preliminary Data Analysis and Results

The LASSO, KNN, RF, SVR, XGBoost, CatBoost and DNN regressors were run on a real estate dataset (with 267 observations of 9 variables), a built-in housing dataset in R for 506 census tracts of Boston from the 1970 census (with 506 observations of 19 variables) and the dataset for this study (with 1,649 observations of 22 variables) after removal of the observations with missing values.

3.7.1 Model Comparison Results

An evaluation of the performances for the 7 models on the three datasets was undertaken using the MAE, RMSE and R-Squared metrics, and their respective training runtimes computed. A comparison of the preliminary results from the evaluation of the 7 models is as outlined in Tables 3.2, 3.3 and 3.4.

Table 3.2: Real Estate Price Prediction Validation Metrics.

SN	Model	MAE	RMSE	RSquared	Runtime (in minutes)
1.	KNN	39.961	50.864	0.771	1.337
2.	RF	33.525	42.05	0.859	0.499
3.	SVR	28.003	34.122	0.871	0.466
4.	XGBoost	24.473	31.569	0.896	0.203
5.	CatBoost	25.925	31.154	0.907	30.768
6.	LASSO	22.709	27.835	0.919	0.075
7.	DNN	22.417	27.71	0.894	34.34

Table 3.3: Homes' Median Value Prediction Validation Metrics.

SN	Model	MAE	RMSE	RSquared	Runtime (in minutes)
1.	KNN	3.406	5.427	0.719	0.833
2.	LASSO	3.477	5.073	0.735	0.062
3.	SVR	2.646	4.546	0.807	0.256
4.	CatBoost	2.809	4.216	0.858	28.069
5.	RF	2.555	3.659	0.885	0.955
6.	XGBoost	2.366	3.264	0.891	0.127
7.	DNN	2.311	2.975	0.886	24.862

Table 3.4: LE Data Set (with NAs dropped) Prediction Validation Metrics.

SN	Model	MAE	RMSE	RSquared	Runtime (in minutes)
1.	DNN	6.488	7.262	0.845	45.999
2.	LASSO	2.795	3.621	0.824	0.104
3.	KNN	2.019	2.978	0.881	2.254
4.	SVR	1.811	2.628	0.907	1.066
5.	CatBoost	1.561	2.349	0.932	1.404
6.	XGBoost	1.395	2.042	0.943	0.442
7.	RF	1.237	1.896	0.953	12.575

Based on the results in Tables 3.2, 3.3 and 3.4, and further backed by the reviewed literature on supervised machine learning techniques that found the eXtreme Gradient Boosting algorithm to offer the best predictions, the study proceeded to use the XGBoost algorithm for modelling.

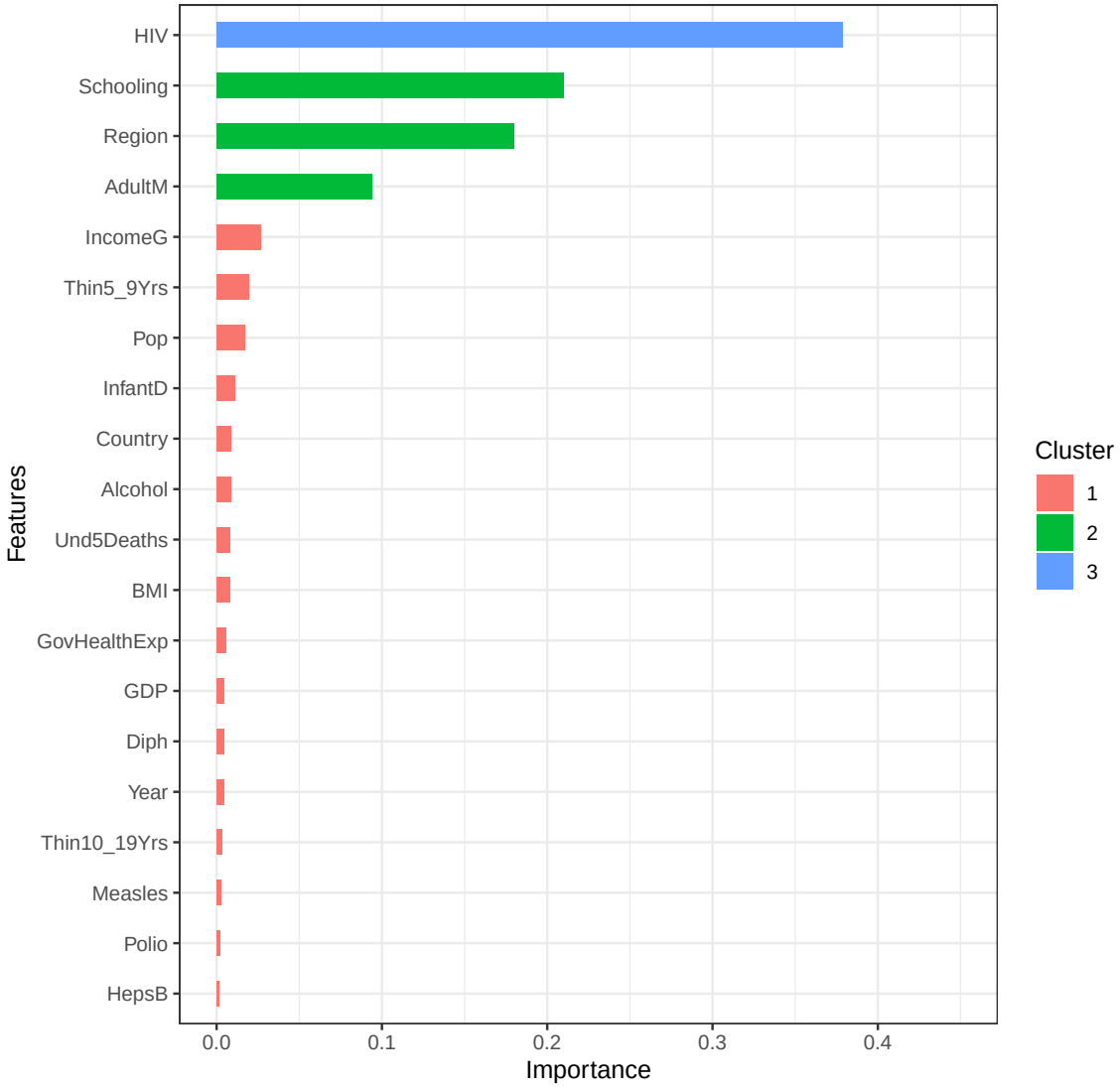


Figure 3.2: Importance Contribution of the Predictor Variables.

The insights from Figure 3.2 above suggests that HIV/AIDS deaths per 1000 population, the number of years at school, regional location and the number of people dying between 15-60 years per 1000 population are the major determinants of the life expectancy at birth rates for a nation.

Chapter 4

Presentation of Research Findings

4.1 Introduction

The data for this study were sourced from the databases of the World Health Organization (WHO) and the United Nations (UN). The data was on 193 UN member states from the year 2000-2015, with the LE health related factors drawn from the Global Health Observatory data repository. The UN data repository provided the corresponding socioeconomic related factors for the 193 Countries. The individual data files were merged into a single data set with 2937 observations of 22 variables.

During the data pre-processing phase, the data values for the *"population"* variable were found to be incorrect. For this reason, the *"population"* values in the initial dataset were replaced with new values from the World Bank Development Indicators' dataset. A further scrutiny into the variable on *"percentage expenditure on health as per GDP"* established that the values for this variable were erroneous, hence, this variable was excluded from the study.

Similarly, the data for Niue, San Marino, Cook Islands, Marshall Islands, Monaco, Palau, Tuvalu and Dominica were missing for considerable number of years. Consequently, these countries were excluded from this analysis. The resultant dataset used in this study had 2832 observations of 21 variables. The study variables are as depicted in Figure 4.1.

4.2 Exploratory Data Analysis

Prior to making inferences from the analysis, the author inspected the dataset to understand its various aspects via exploratory data analysis (EDA).

As noted by Sonal (2021), EDA is the task of utilizing summary statistics and graphical visualizations to conduct preliminary examinations on data in order to reveal patterns, detect anomalies, test hypotheses, and verify assumptions. EDA was conducted to maximize the author's insight into the LE dataset and its underlying structure, while providing the precise elements to be extract from the dataset. The exploration was aided by Cui (2020).

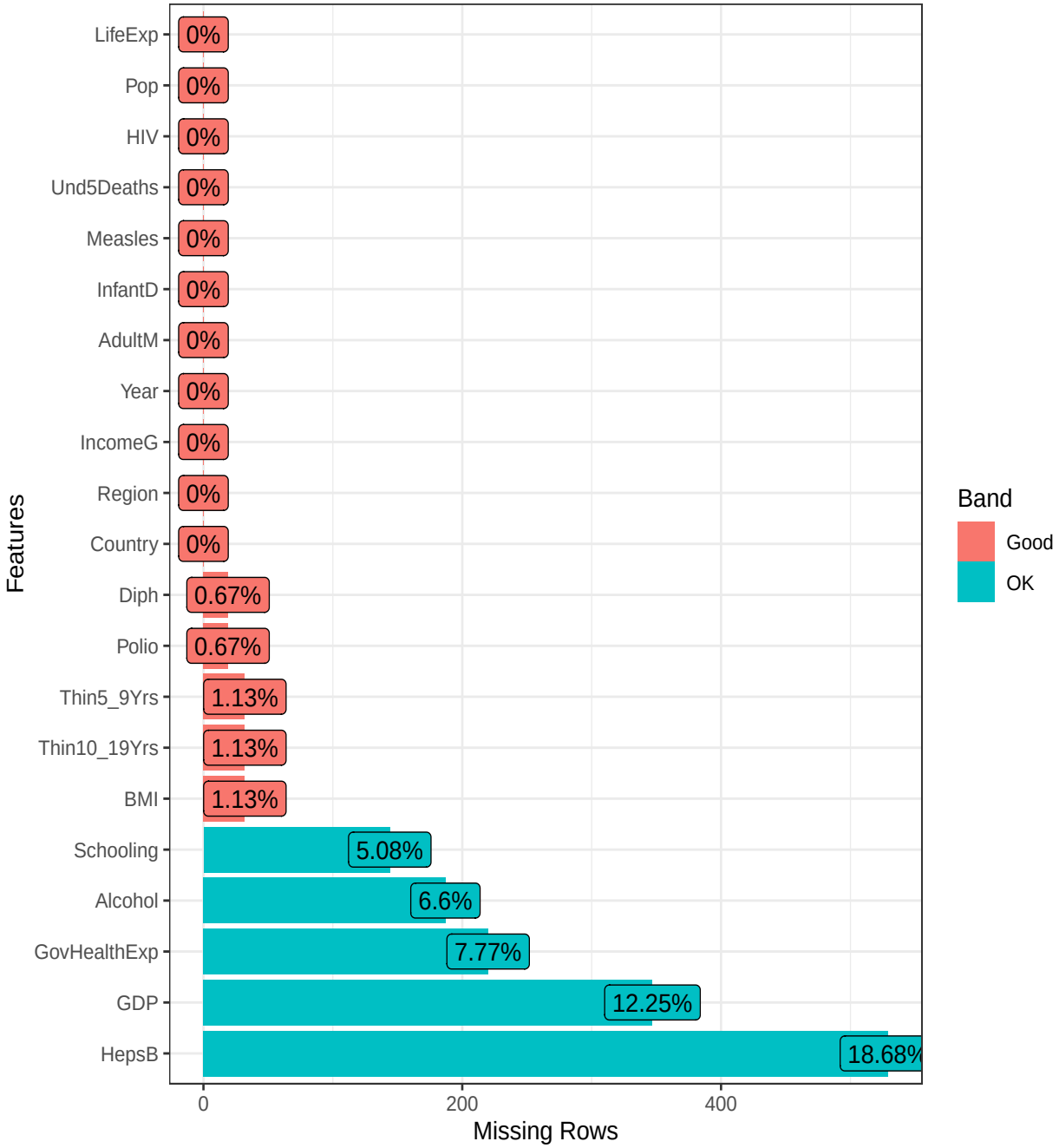


Figure 4.1: Missingness Profile of the Twenty-One Study Variables.

As is the case with any real world data set, the LE data set used had missing values. It can be observed from Figure 4.1 that 18.9%, 12.3%, 7.8%, 6.6% and 5.1% of the countries had missing data values for percentage of Hepatitis B immunization coverage among 1 year olds, gross domestic product per capita, government expenditure of health as a percentage of total government expenditure, per capita alcohol consumption and the number of years at school, respectively.

The amount of null values in BMI, percent of thinness among children from age 10-19 and age 5-9, and the percent of polio and diphtheria immunization coverage among one year olds were considerably small. Since XGBoost is robust to missingness and can automatically detect and deal with missing values, manual transformation of the null values was redundant. As a preliminary step to feature selection however, missing values were imputed using the PCA model via the *imputePCA* function (Josse and Husson, 2016).

An assessment of the multivariate normality for the dataset was conducted using the MVN package (Korkmaz et al., 2014). Evidently from the results given in Table 4.1, both univariate and multivariate tests indicate a deviation from normality. In accordance with the Royston's test, the LE data set does not follow a multivariate normal distribution (H = 2642.338, $p = 0$). This finding is further affirmed by Figure 4.2 and Figure 4.3.

Table 4.1: Univariate and Multivariate Normality Assessment Test Results.

SN	Test	Variable	Statistic	P-Value	Normality Met?
1.	Shapiro-Wilk	AdultM	0.90	<0.001	NO
2.	Shapiro-Wilk	InfantD	0.24	<0.001	NO
3.	Shapiro-Wilk	Alcohol	0.90	<0.001	NO
4.	Shapiro-Wilk	HepsB	0.70	<0.001	NO
5.	Shapiro-Wilk	Measles	0.21	<0.001	NO
6.	Shapiro-Wilk	BMI	0.93	<0.001	NO
7.	Shapiro-Wilk	Und5Deaths	0.24	<0.001	NO
8.	Shapiro-Wilk	Polio	0.64	<0.001	NO
9.	Shapiro-Wilk	GovHealthExp	0.99	<0.001	NO

SN	Test	Variable	Statistic	P-Value	Normality Met?
10.	Shapiro-Wilk	Diph	0.63	<0.001	NO
11.	Shapiro-Wilk	HIV	0.33	<0.001	NO
12.	Shapiro-Wilk	GDP	0.54	<0.001	NO
13.	Shapiro-Wilk	Pop	0.22	<0.001	NO
14.	Shapiro-Wilk	Thin10_19Yrs	0.83	<0.001	NO
15.	Shapiro-Wilk	Thin5_9Yrs	0.82	<0.001	NO
16.	Shapiro-Wilk	Schooling	0.99	<0.001	NO
17.	Shapiro-Wilk	LifeExp	0.95	<0.001	NO
18.	**Royston**	**MVN Test**	**2642.34**	**0**	**NO**

The histograms in Figure 4.2 and Figure 4.3 suggest that the underlying distributions of the continuous predictors are not approximately normal. The variables histograms reveal skewed distributions across all the variables. XGBoost algorithm is a non-parametric method. For this reason, the reliability of our study results is not bound by the multivariate normality assumptions.

Insights on health-related factors are drawn from Figure 4.2. It can be observed that, the number of individuals dying between 15-60 years per 1000 population is skewed to the right. In most of the countries, 0 to around 250 people die in this age bracket. Similarly, 0 to 5 individuals die as a result of HIV/AIDS per 1000 population in most countries.

The percent of diphtheria, Hepatitis B and polio immunization coverage among one year olds are skewed to the left. In most countries, the immunization coverage ranges between 80% to 100%. The number of infants dying, under five deaths and measles reported cases per 1000 population are right skewed. Few instances of deaths are reported among infants and children below the age of five in nearly all the countries. Likewise, fewer cases of measles are reported per 1000 population in nearly all the countries.

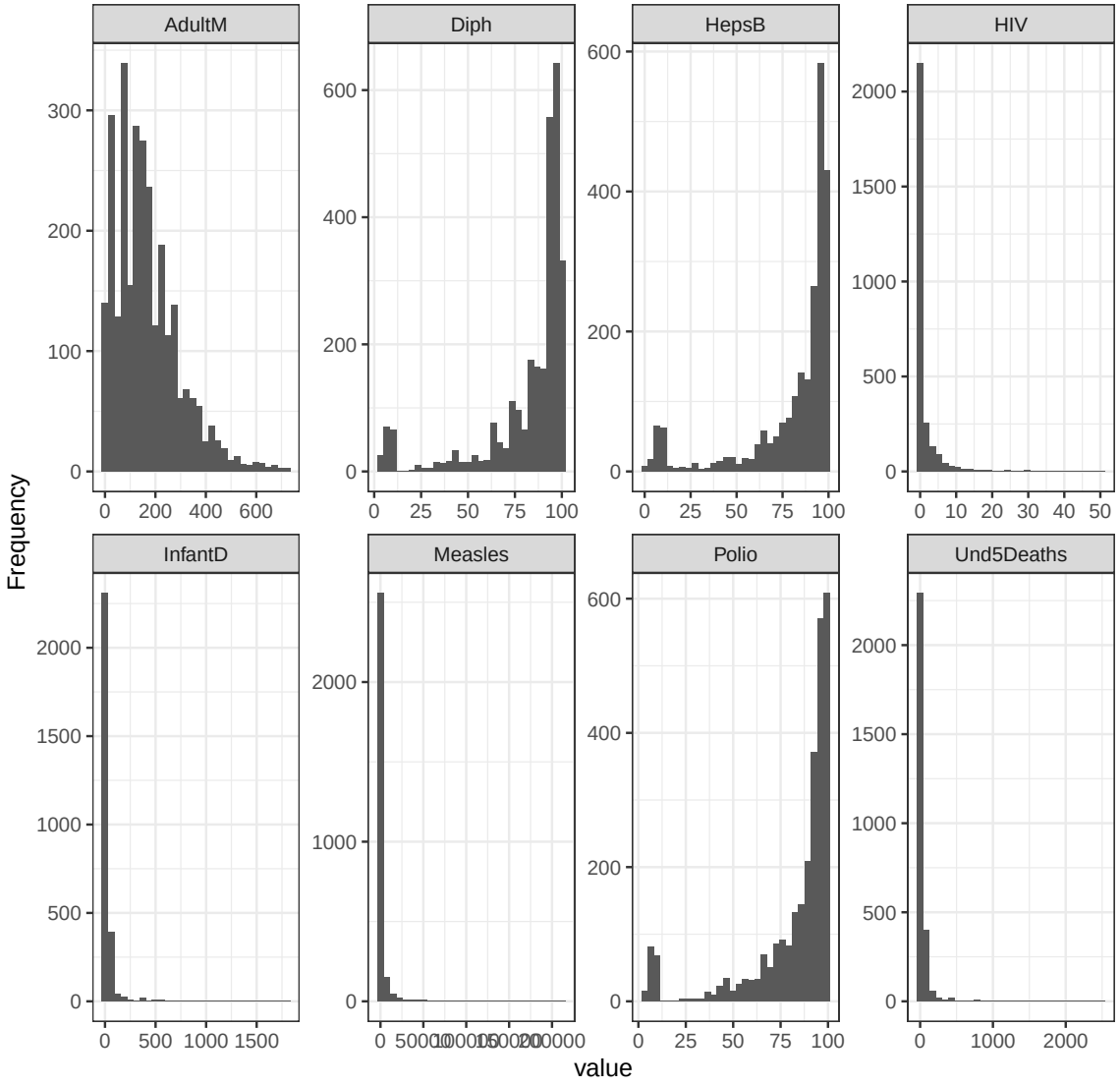


Figure 4.2: Histograms of Health-Related Factors.

Insights from Figure 4.3 reveal skewed, multimodal and near normal distributions for the socioeconomic factors. In most countries, the recorded per capita alcohol consumption is between 0 and 5 litres. The multimodal distribution for the average BMI of the populations suggests an existence of sub-populations with different underlying traits. Majority of these sub-populations seem to have BMI ranging between 15 to 30 and 45 to 60 across the countries.

GDP per capita in dollars, population and thinness appear to be skewed to the right. In most countries, the GDP per capita was between 0 and 10,000 dollars across the study period.

The percentage expenditure on health against the total government expenditure ranged between 2.5% to 10% in most countries. All the countries had populations below 5 billion people with fewer countries crossing the 1 billion mark. The number of years spent in school by individuals for the majority of countries was between 10 and 17.5 across the 15-year-study period.

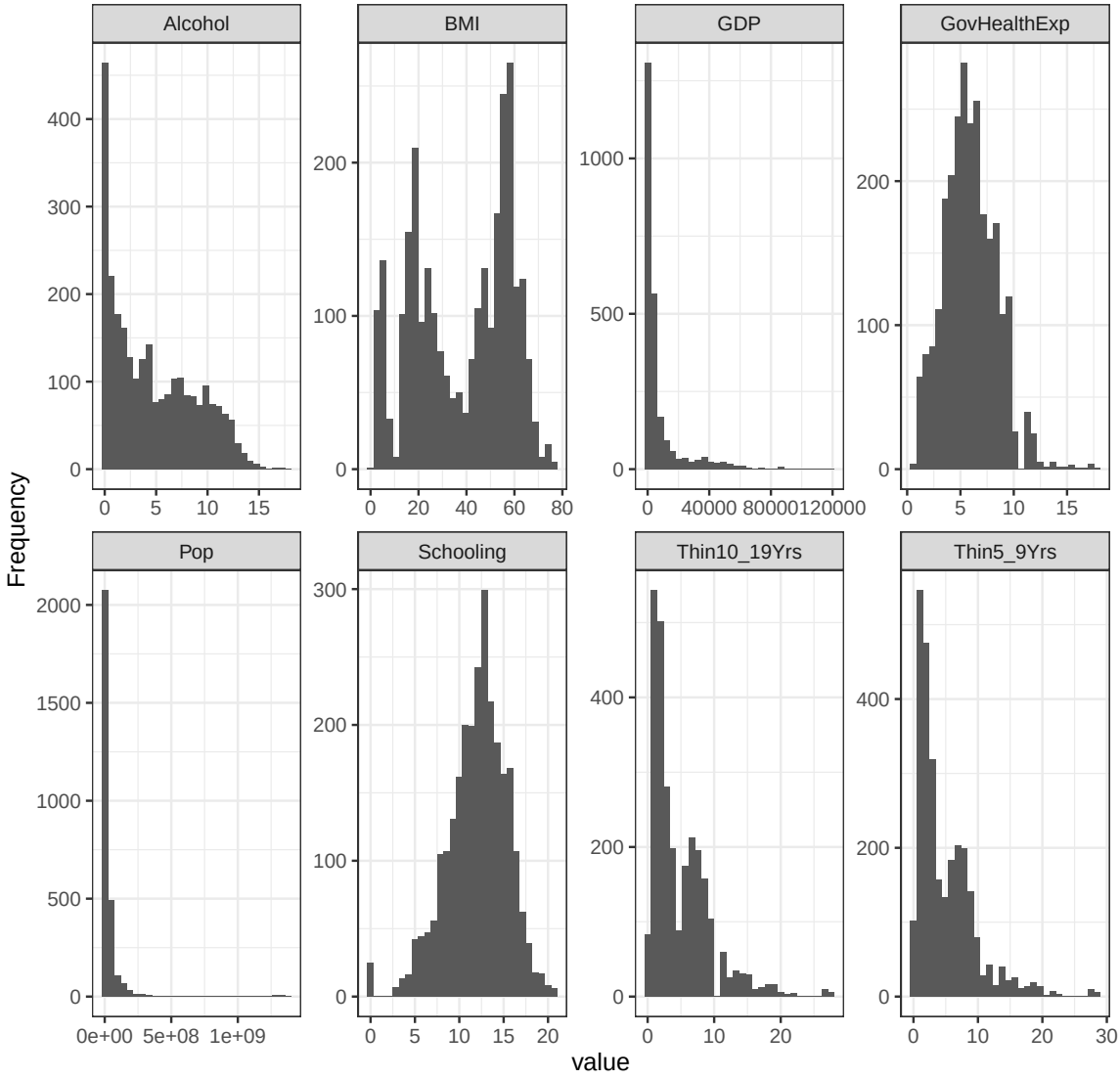


Figure 4.3: Histograms of Socioeconomic Factors.

Thinness among children aged 10-19 and 5-9 ranged between 0-10% in most countries. A handful of the countries experienced thinness levels that were way above 25% for these ages.

The relationship between variables was examined to determine their association. The heatmap in Figure 4.4 depicts presence of multicollinearity among the predictor variables. Adult mortality is highly influenced by the HIV/AIDS deaths. Infant deaths (under 5 deaths inclusive) are influenced by measles cases and the population density. Per capita alcohol consumption is influenced by the number of schooling years. There exists a high positive association between Hepatitis B and Diphtheria immunizations.

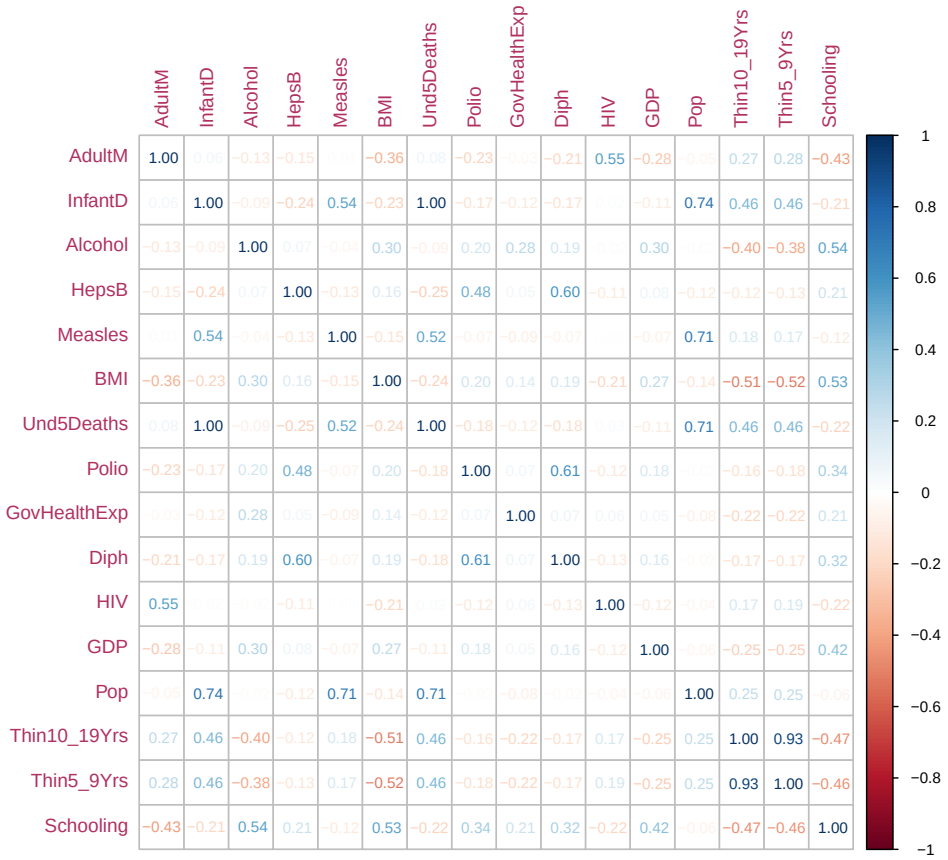


Figure 4.4: A Visualization of the Correlation Matrix Heatmap.

The average body mass index of the entire population is negatively associated with thinness and positively influenced by the number years at school. Polio and Diphtheria immunizations are also positively related. Similarly, a country's population is highly influenced by the number of infant deaths (under 5 deaths inclusive) per 1000 population and the number of measles reported cases per 1000 population. PCA was employed to address multicollinearity.

4.3 Results from the Analysis

4.3.1 Overall Findings

Results from the Extended Multivariate Kruskal-Wallis (E-MKW) rank sum test revealed significant differences in the mean vectors of the numeric variables across the regions (EM Kruskal-Wallis chi-squared = 2615.833, df = 220.4847, p-value = 0). Similarly, there were significant differences among the variables across the three income groups (EM Kruskal-Wallis chi-squared = 1128.493, df = 73.49491, p-value = 1.448515e-188).

The populations of the richest countries in the world have life expectancies of over 60 years with the North American countries taking the lead with an average life expectancy of 79.88 years. The countries in the Sub-Saharan Africa region have the lowest life expectancy rates with a mean of 57.12 years. The details of the respective regional mean life expectancy rates are as highlighted in Table 4.2.

Table 4.2: Average Life Expectancy Rates in Years per Region.

SN	Region	Mean Life Expectancy (in years)
1.	North America	79.88
2.	Europe & Central Asia	75.99
3.	Middle East & North Africa	73.16
4.	Latin America & Caribbean	73.07
5.	East Asia & Pacific	71.47
6.	South Asia	67.37
7.	Sub-Saharan Africa	57.12

A notable finding to note from Figure 4.5 is that generally, the mean life expectancy rates at birth have been increasing progressively over the years from 2000 to 2015 in all the seven regions. The North America region seems to have experienced some slight decrease in the mean life expectancy rate at birth in the year 2009. Similarly, South Asia also experienced a slight decrease in the average life expectancy rate at birth, in the year 2015.

It is also crystal clear from Figure 4.5 that the North American region consists of only high income countries. Generally, Sub-Saharan Africa and South Asian countries have low incomes, with Sub-Saharan African countries having the least incomes, over the years. In the

Sub-Saharan Africa region, high income countries seem to have higher life expectancy rates in comparison to their low and middle income counterparts. South Asia has no country in the high income cohort.

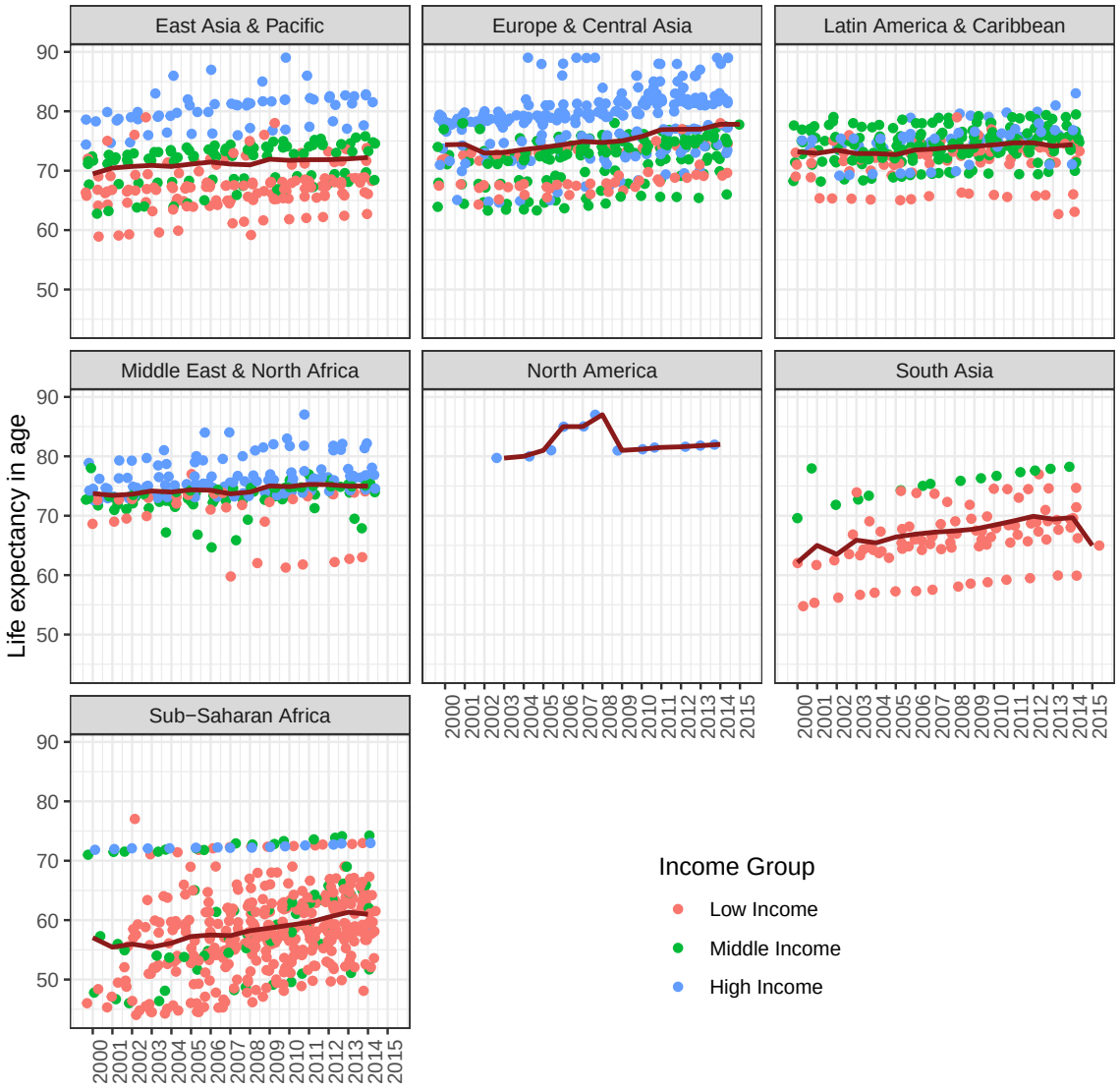


Figure 4.5: Regional Life Expectancy Scatter Plots by Countries' Income Groups.

The Box and Whisker Plots in Figure 4.6 illustrates that in general, the number of people dying between 15-60 years per 1000 population is high in the low income countries for East Asia and Pacific, Europe and Central Asia, Latin America and Caribbean, Middle East and North Africa, and South Asia regions.

Likewise, this number is high for the middle income countries in the Sub-Saharan Africa region. Fewer number of individuals die between 15-60 years per 1000 population in the high income countries for all the regions, except for South Asia where no country is in this group.

In East Asia and Pacific, the number of majority of the people dying between 15-60 years per 1000 population is between 150-225, 80-200, and 75-90 for low, middle and high income countries, respectively. In Europe and Central Asia, this number is between 120-190, 75-175 and 65-100 for low, middle and high income countries, respectively.

Latin America and Caribbean has the number of the majority of individuals dying between 15-60 years per 1000 population between 175-200, 100-175, and 100-130 for low, middle and high income countries, respectively. Middle East and North Africa region has this number between 130-190, 15-130, and 10-100 for low, middle and high income countries, respectively.

This number is between 80-95 in North America for the high income countries, this being the only income group in the region. Similarly from Figure 4.6, South Asia has the number of most individuals who die between 15-60 years per 1000 population ranging between 130-240 and 65-115 for low and middle income countries, respectively. This region has no country in the high income group.

On the other hand, the Sub-Saharan Africa region has this number between 215-385, 225-425, and 125-190 for for low, middle and high income countries, respectively. This region is the leading in the number of individuals dying between 15-60 years per 1000 population for all the income groups.

Comparably, the median number of individuals dying between 15-60 years per 1000 population in the North American countries is 70. The Sub-Saharan Africa region has the median at 285, 324 and 186 for low, middle and high income countries, respectively. For the South Asian region, the median is 178 and 78 for low and middle income countries, respectively.

In Middle East and North Africa region, the median number of individuals dying between 15-60 years per 1000 population is 172, 118 and 75.5 for low, middle and high income

countries, respectively. Latin America and Caribbean region has a median of 188, 142 and 128 for low, middle and high income countries, respectively. Europe and Central Asia region has a median of 182, 134 and 77 for low, middle and high income countries, respectively. Similarly, East Asia and Pacific region has a median of 179, 144 and 69 for the low, middle and high income countries, respectively.

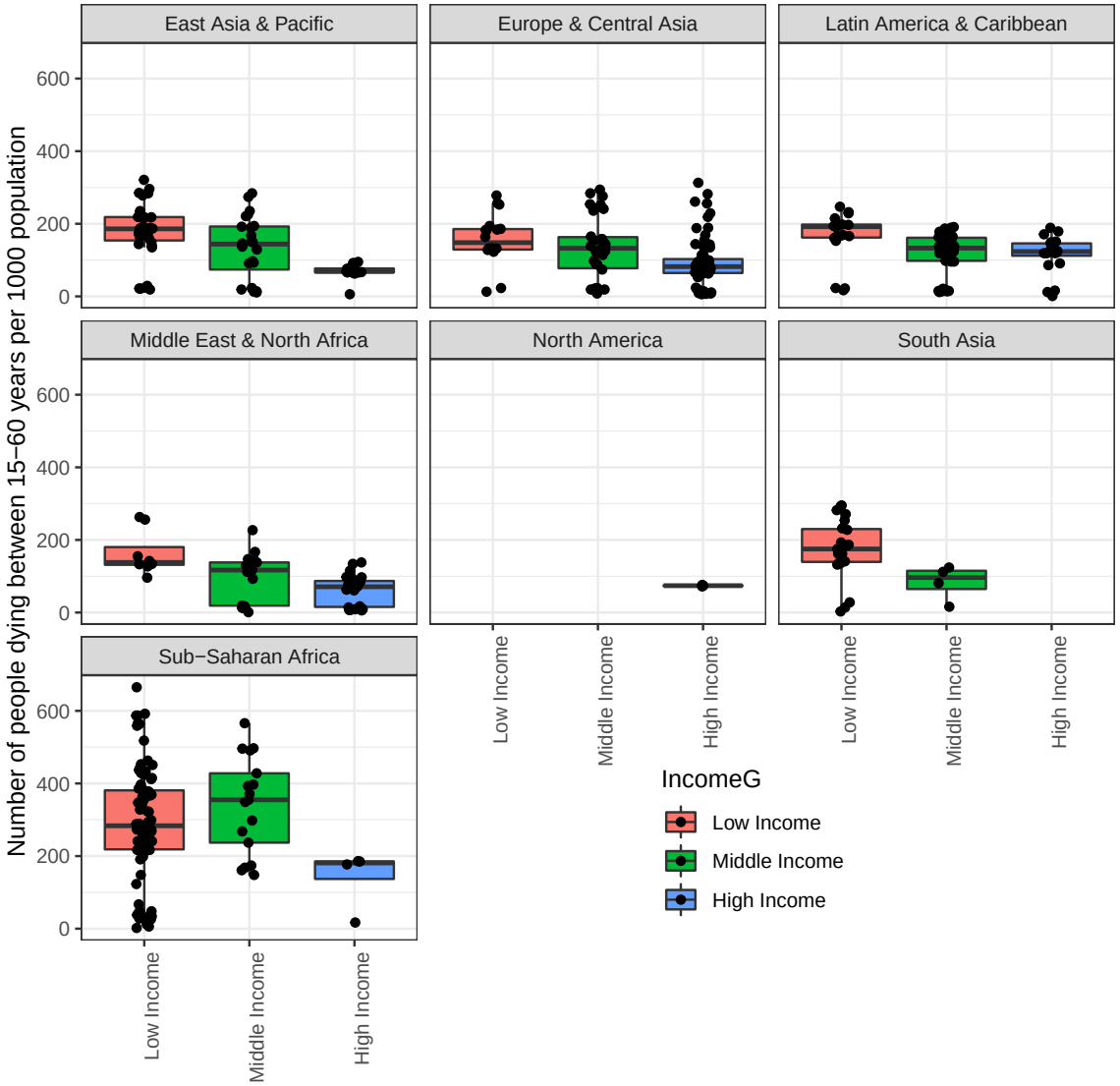


Figure 4.6: Regional Adult Mortality by Countries' Income Groups Box Plots.

Table 4.3 highlights the average number of people living in the seven regions. South Asia is the most populous region globally with a mean population density of 197.01 million people.

Table 4.3: Average Population Densities in Millions per Region.

SN	Region	Mean Population Density (in millions)
1.	South Asia	197.01
2.	North America	167.63
3.	East Asia & Pacific	85.53
4.	Latin America & Caribbean	19.95
5.	Europe & Central Asia	18.64
6.	Middle East & North Africa	18.21
7.	Sub-Saharan Africa	17.34

Sub-Saharan Africa on the flip side is the least populated region with an average population density of 17.34 million people.

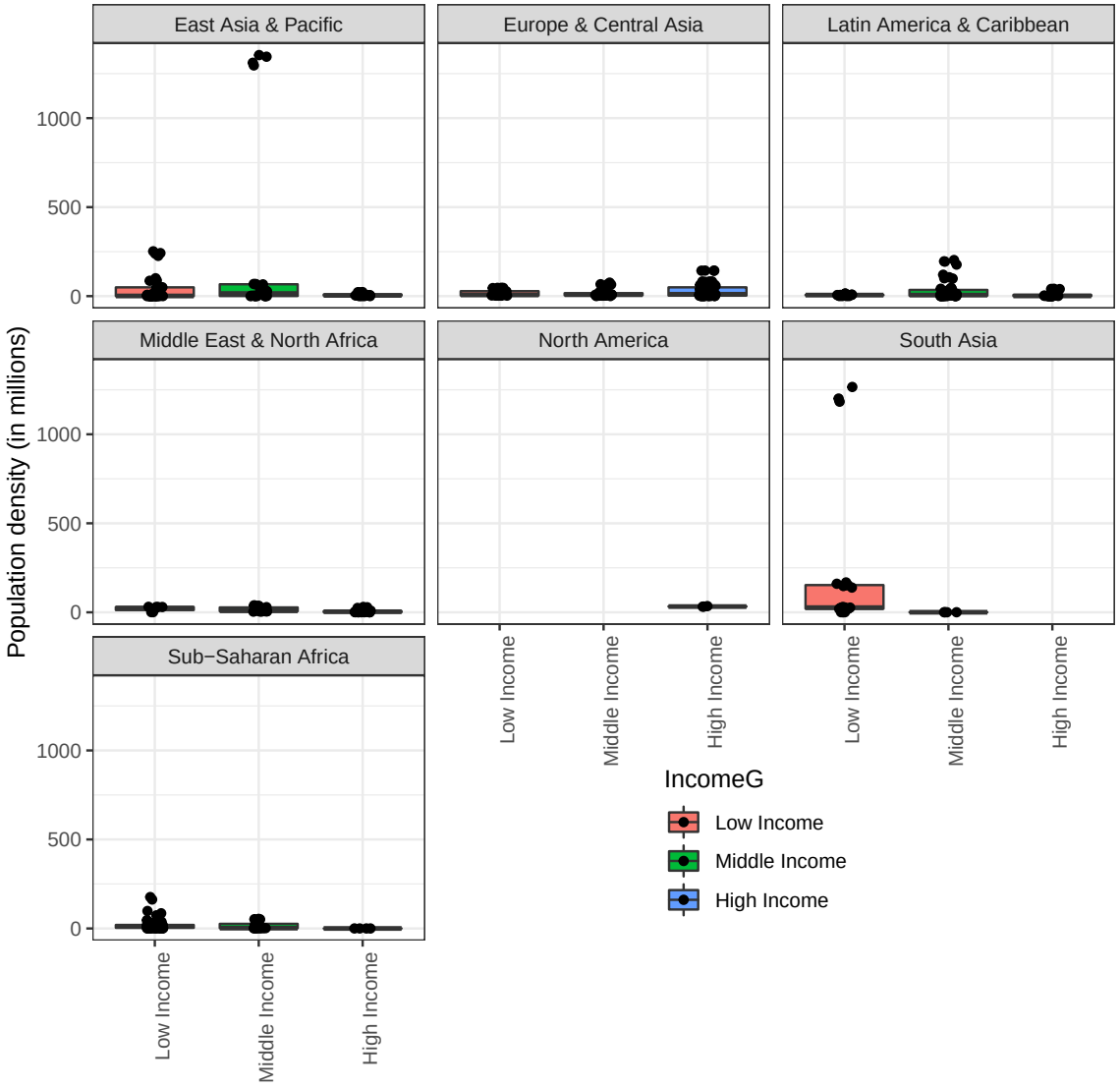


Figure 4.7: Regional Population Densities by Countries' Income Groups Box Plots.

Notably from Figure 4.7, it is observed that the low income countries in East Asia and Pacific, and the South Asian regions are populous. In East Asia and Pacific region, it is evident that some countries under the middle income cohort have the highest population densities. Similarly, there exists some low income countries with the highest population densities in the South Asian region. The high income countries in Europe and Central Asia seem to be populous too.

Figure 4.8 visualizes the respective health expenditure rates by different regional governments.

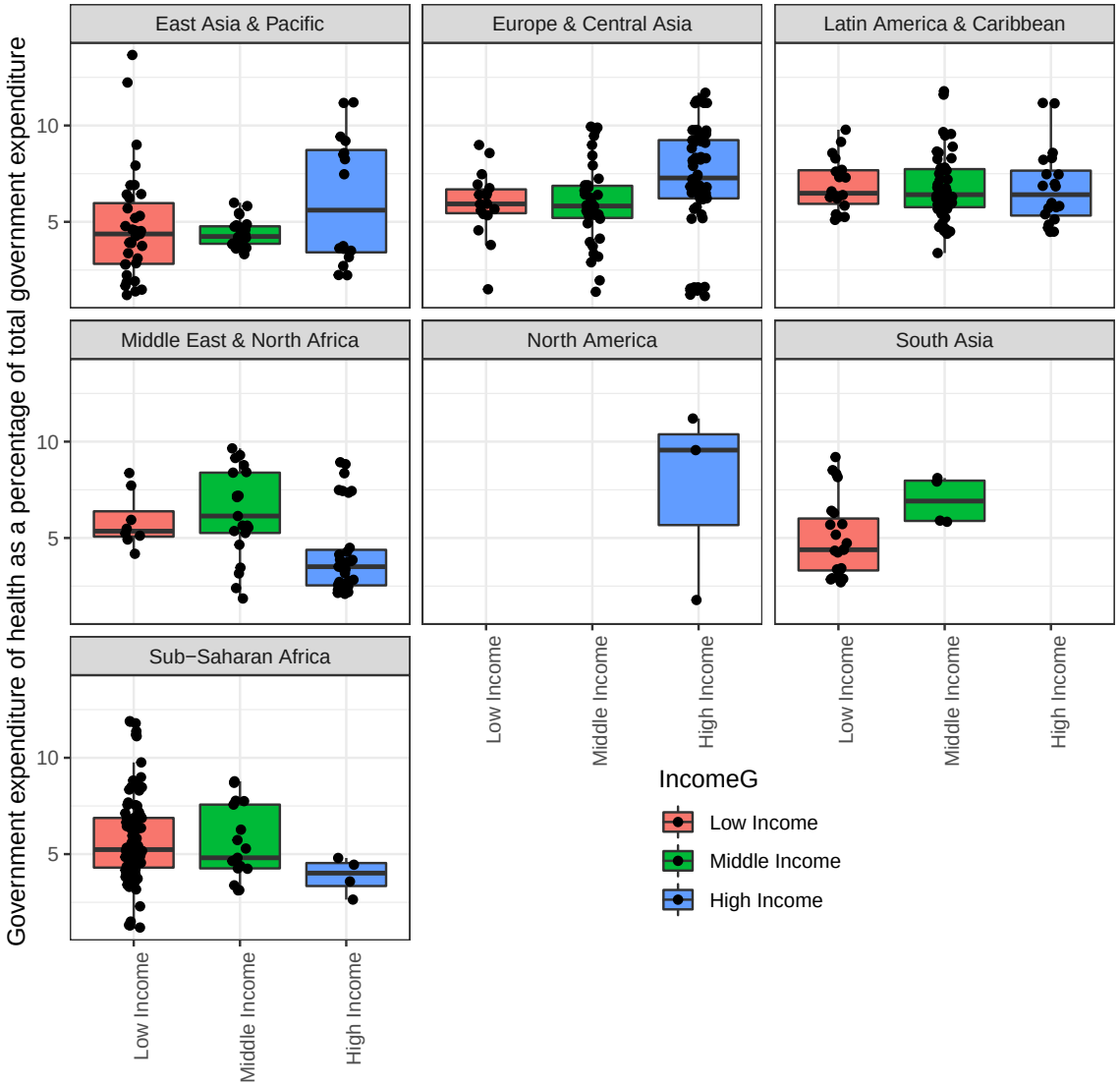


Figure 4.8: Regional Government Health Expenditures by Income Groups Box Plots.

It is evident from Figure 4.8 that the high income countries in East Asia and Pacific, and the North America regions have higher health expenditure rates as a percentage of the total government expenditure in comparison to the other regions. North American countries seem to have the highest health expenditures as a percentage of the total government spending.

In Middle East and North Africa, and the Sub-Saharan Africa regions, the expenditure rates on health are higher in the middle income countries. Notably, some low income countries in East Asia and Pacific region have high health expenditure rates. Generally, the low income countries in nearly all the regions seem to have the least expenditures on health.

4.3.2 Key Determinants of Life Expectancy

Figure 4.9 visualizes the contribution of study variables to the first principal component. The anticipated average contribution of the variables is shown by the red dashed line (6.25%).

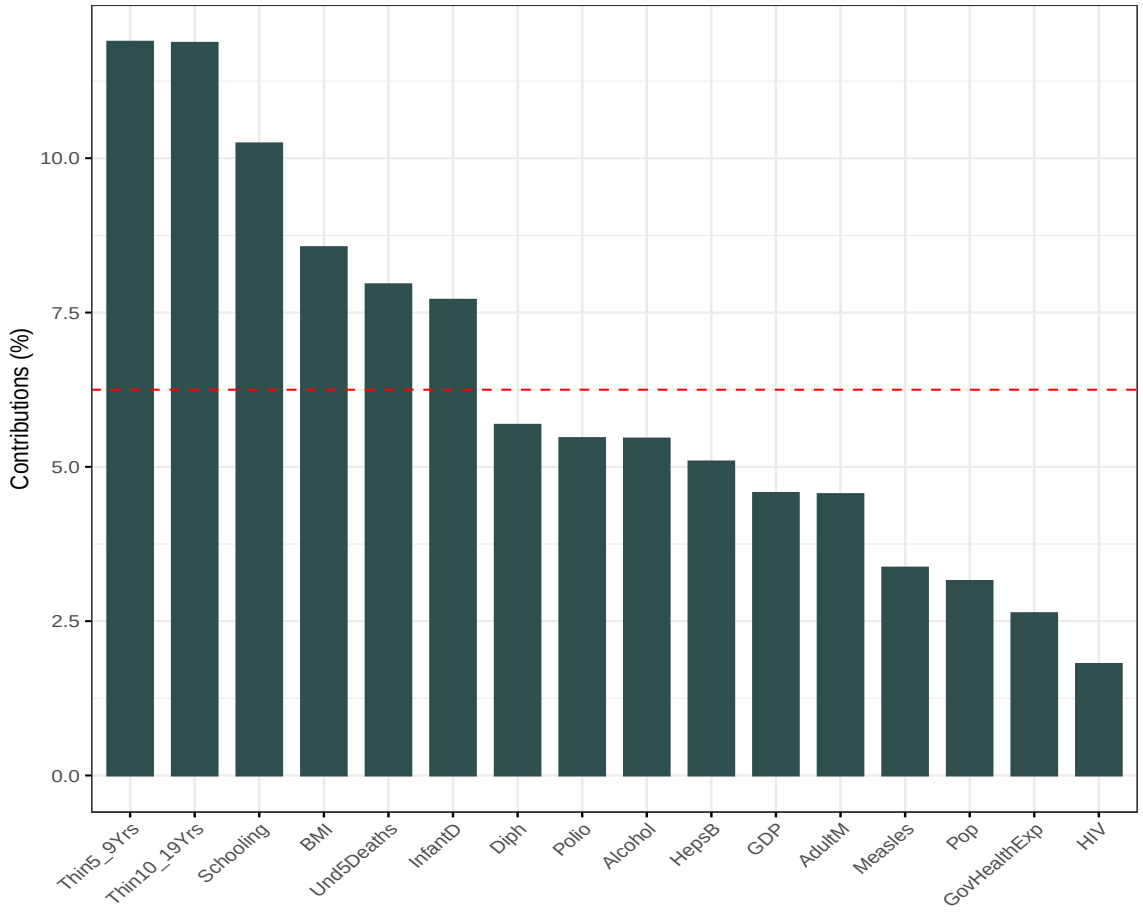


Figure 4.9: Contribution of the Numeric Predictors to the First Principal Component.

If the research variables contributed equally, the expected average value depicted as the cutoff in Figure 4.9 would be arrived at as shown in equation 4.1. A variable is deemed essential

for contributing to a specific principal component if its contribution exceeds this threshold (Kassambara, 2017).

$$\text{Expected Average Contribution}\,(\%) = \left[\frac{1}{\text{No. of Variables}}\right]\% = \frac{1}{16} \times 100 = 6.25\%. \quad (4.1)$$

Principal component one accounted for 32.6% of the variability in the dataset. From Figure 4.9, the percent of thinness among children aged 5-9 & 10-19, number of years at school, average body mass index, number of under-five deaths and infant deaths per 1000 population were important in contributing to the first principal component. These six variables exceed the threshold of 6.25%, hence considered important. As the research variables for further analysis, the six variables as well as region and income group were chosen.

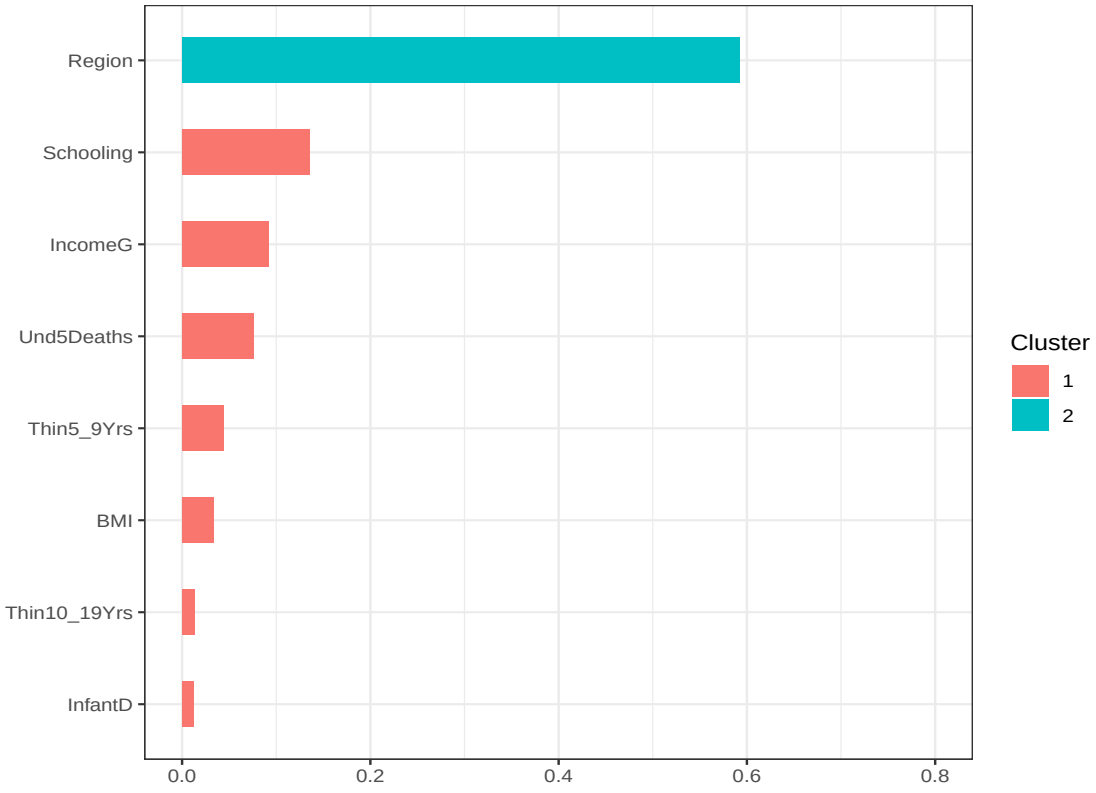


Figure 4.10: Selected Features' Importance Contribution Relative to the Whole Model.

The importance of the variables in the final model based on the gain measure is as depicted in Figure 4.10. Regional location, number of years at school, income group, number of under-five deaths per 1000 population, percent of thinness among children aged 5-9 and the average BMI are the key determinants of life expectancy, respectively.

4.3.3 Association of Life Expectancy and its Selected Predictors

Figure 4.11 visualizes the Spearman rank correlation matrix heatmap between life expectancy and the six significant predictors. Percent of thinness among children aged 5-9 and 10-19, the number of under five deaths, and the number of infant deaths are negatively associated with LE. The number of years spent in school and the average BMI are positively correlated with LE.

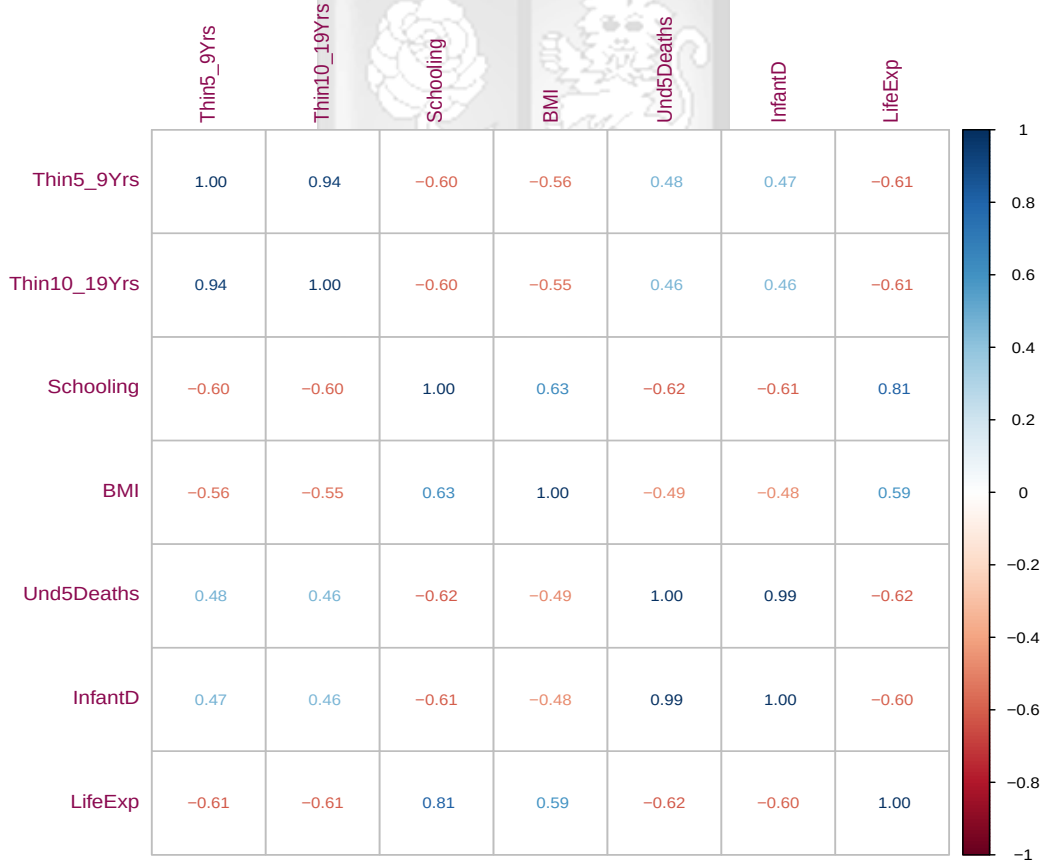


Figure 4.11: A Correlation Matrix Heatmap of the Selected Numeric Predictors with LE.

The percent of thinness among children aged 5-9 (t = -26.701, df = 2654, p-value < 2.2e-16) and 10-19 (t = -27.092, df = 2654, p-value < 2.2e-16) were significantly correlated with LE. Similarly from Figure 4.11, the number of years at school and life expectancy were significantly correlated (t = 58.709, df = 2654, p-value < 2.2e-16). Furthermore, the average BMI (t = 35.733, df = 2654, p-value < 2.2e-16), number of under five deaths (t = -10.72, df = 2654, p-value < 2.2e-16) and infant deaths (t = -9.3523, df = 2654, p-value < 2.2e-16) per 1000 population were established to have a significant correlation with life expectancy.

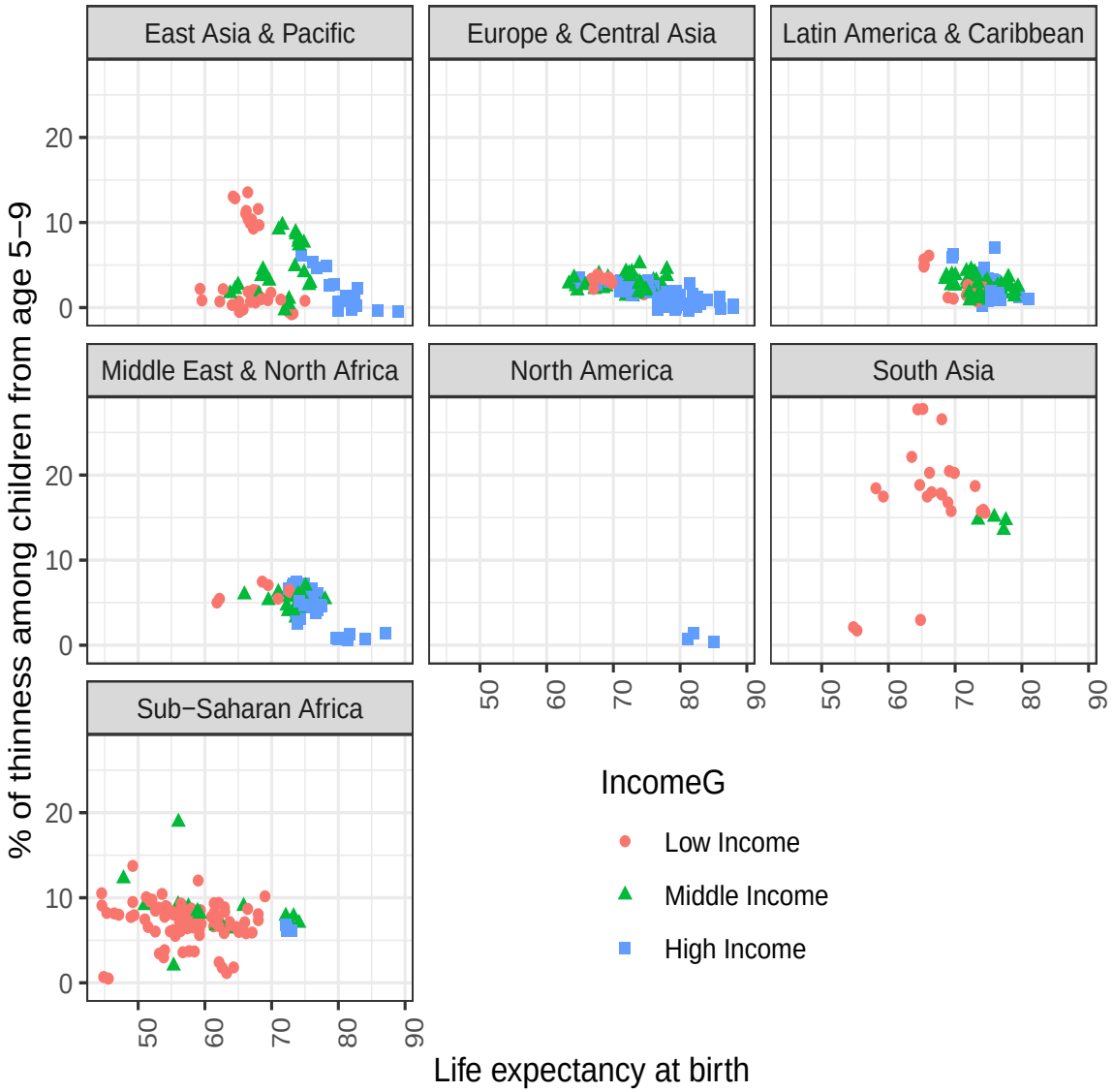


Figure 4.12: Scatter Plots of the Association Between Thinness for 5-9 Year Olds & LE.

It is evident from Figure 4.12 that life expectancy improves as the percent of thinness among children aged 5-9 years reduces.

The scatter plot in Figure 4.13 depicts a general trend of increasing life expectancy rates with a decrease in the percent of thinness among children aged 10-19 years. Low income countries seem to have higher percentages of thinness among 10-19 year olds in comparison to their middle and high income counterparts.

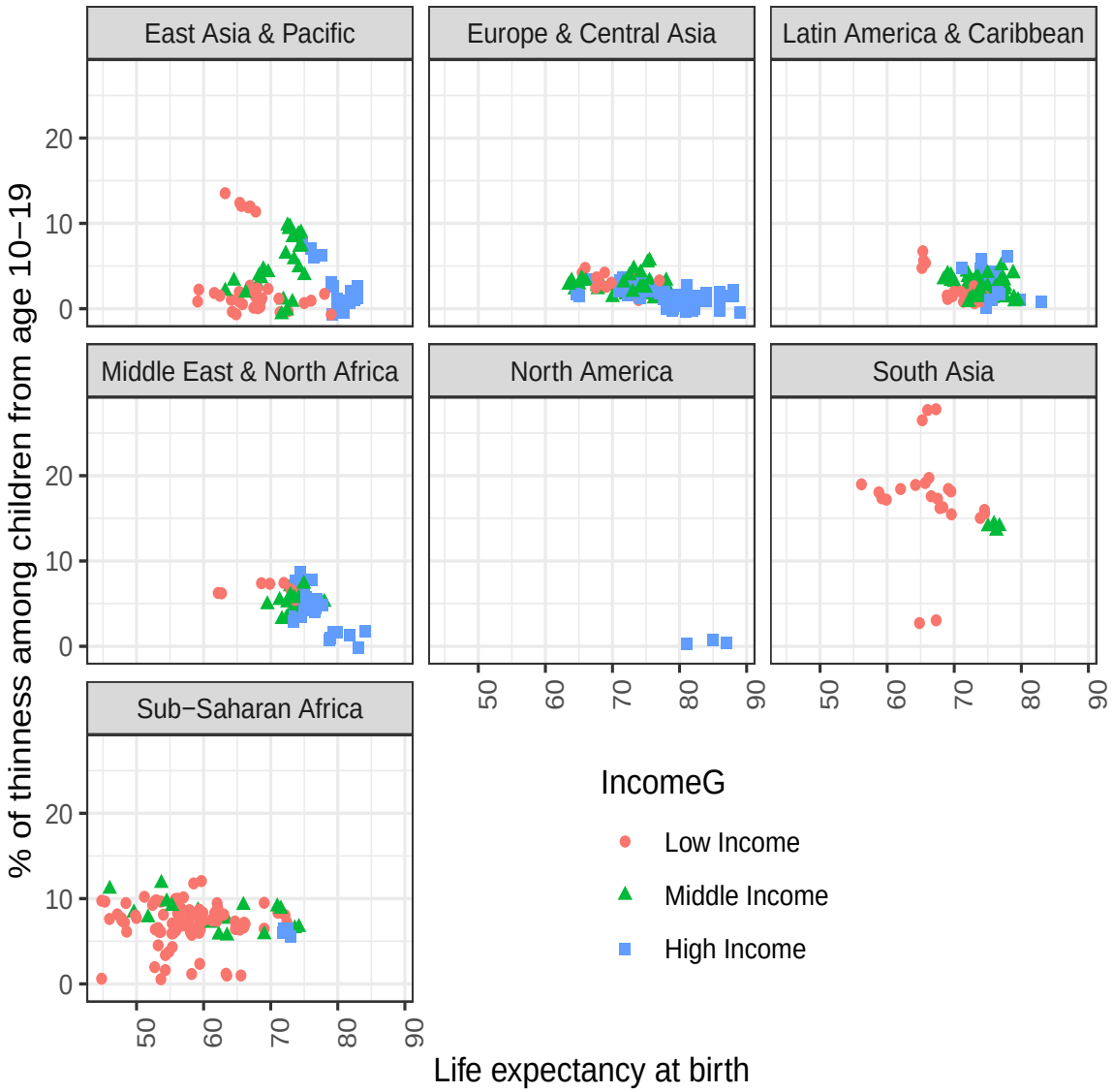


Figure 4.13: Scatter Plots of the Association Between Thinness for 10-19 Year Olds & LE.

From Figure 4.14, it is evident that life expectancy increases with the number of years spent in school across all the seven regions. High income countries have a majority of their populations spending more years in school compared to middle and low income countries.

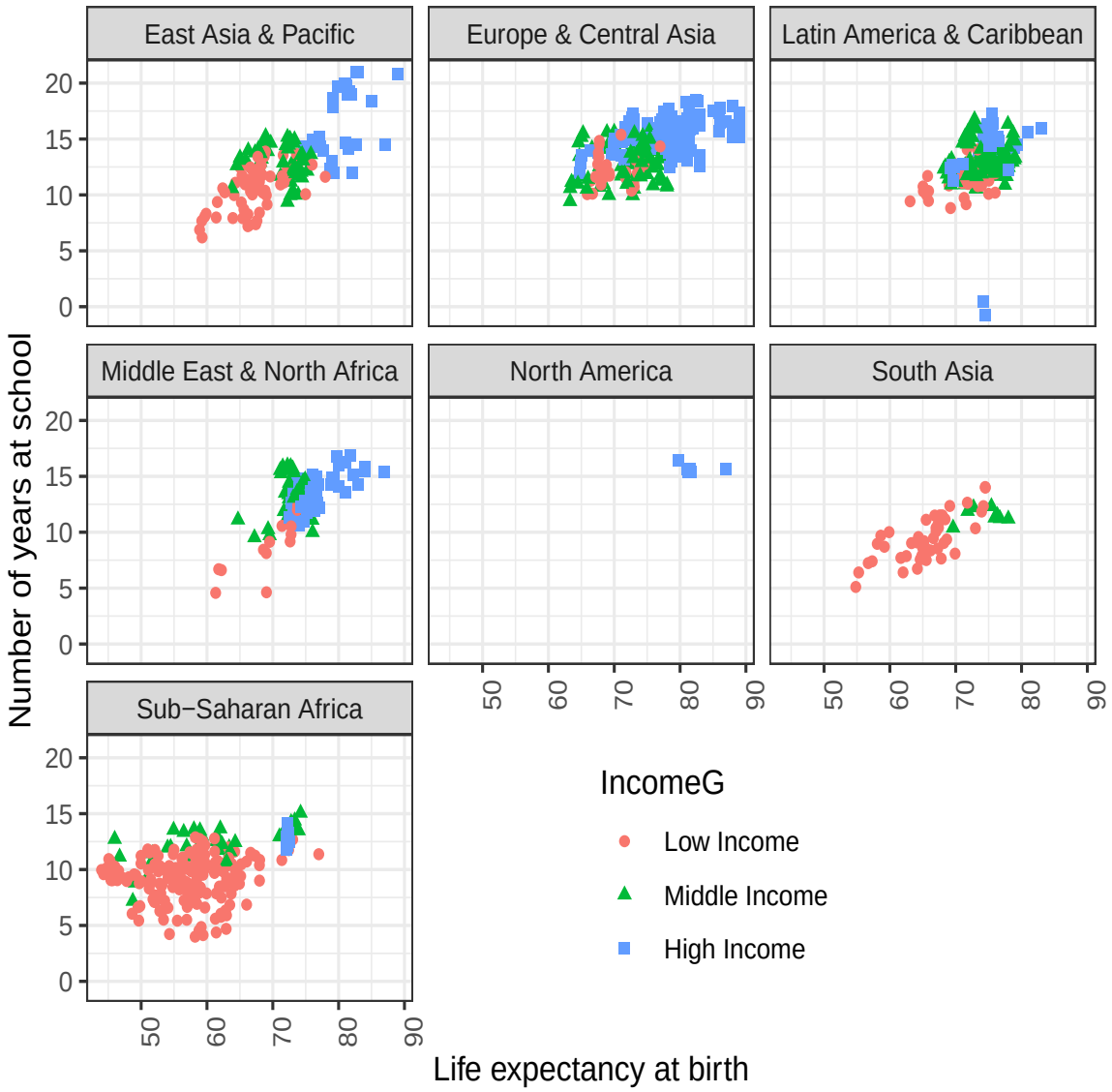


Figure 4.14: Scatter Plots of the Association Between Number of Years at School & LE.

Figure 4.15 suggests an increase in life expectancy rates with average body mass index of the entire population. Majority of the countries in the Sub-Saharan Africa and South Asia have BMI within the healthy and underweight ranges. Most of the populations in Europe

& Central Asia, Latin America & Caribbean, and Middle East & North Africa regions are obese.

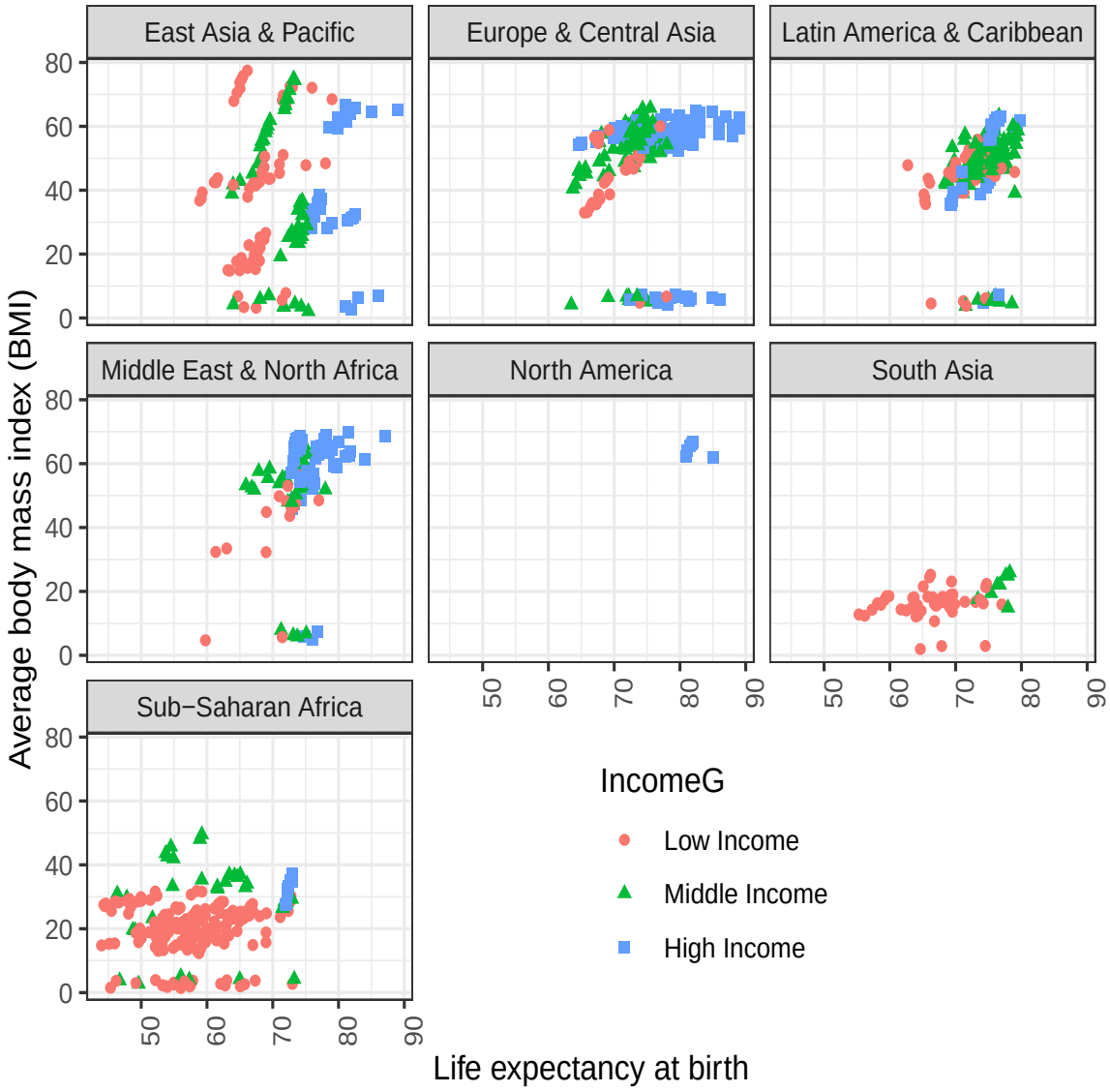


Figure 4.15: Scatter Plots of the Association Between Average BMI & LE.

Notably from Figure 4.16, life expectancy improves as the number of under five deaths per 1000 population decreases. High income countries appear to have the lowest numbers of under-five deaths. Low income countries in the Sub-Saharan Africa and South Asian regions have higher numbers of under-five deaths.

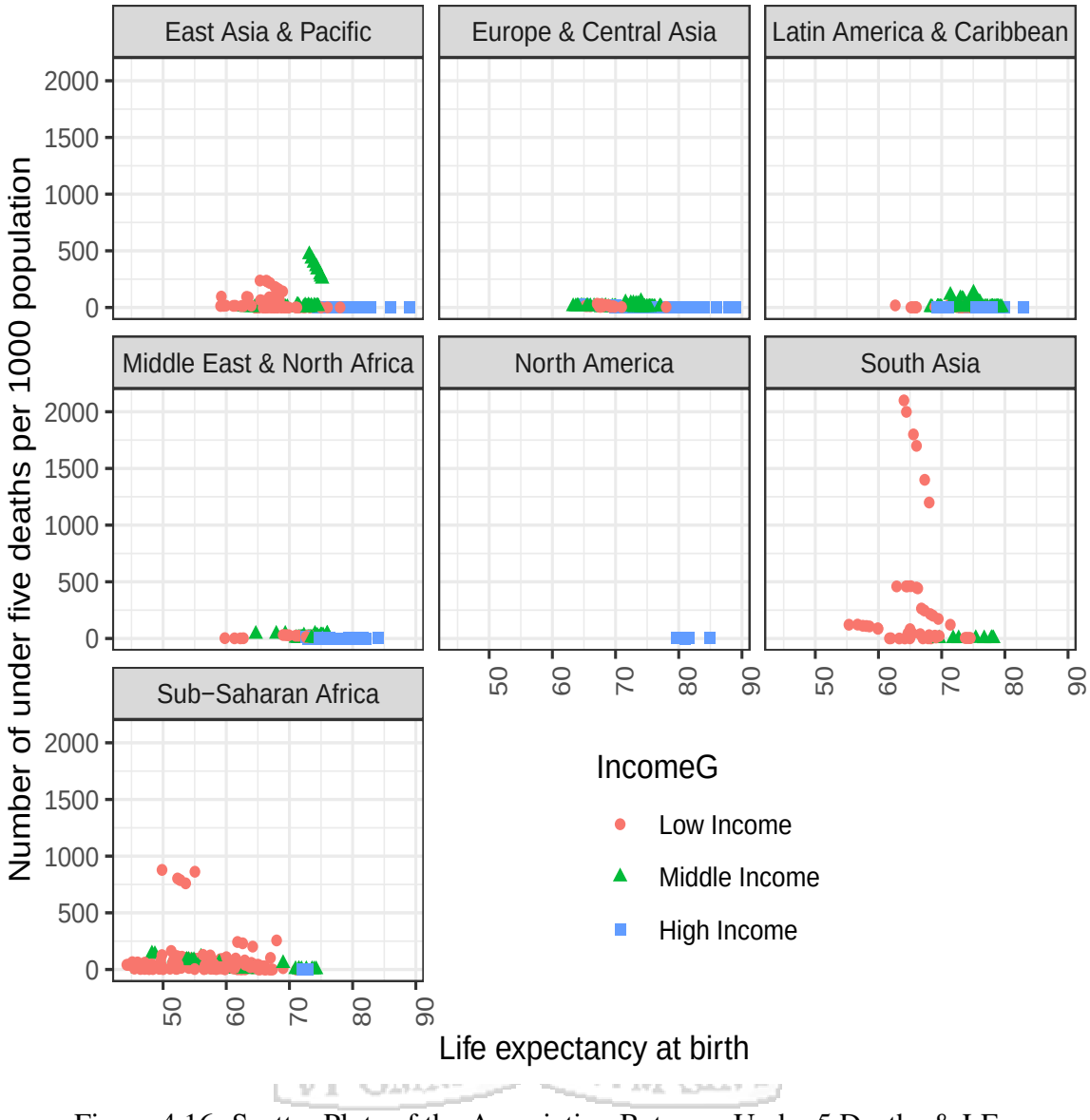


Figure 4.16: Scatter Plots of the Association Between Under 5 Deaths & LE.

It can further be observed from Figure 4.17 that life expectancy increases with a decrease in the number of infant deaths per 1000 population. High income countries seem to experience the lowest numbers of infant mortality across all the regions. On the other hand, low income countries in the Sub-Saharan Africa and South Asian regions, and some of the middle income countries in East Asia & Pacific have higher numbers of infant mortality.

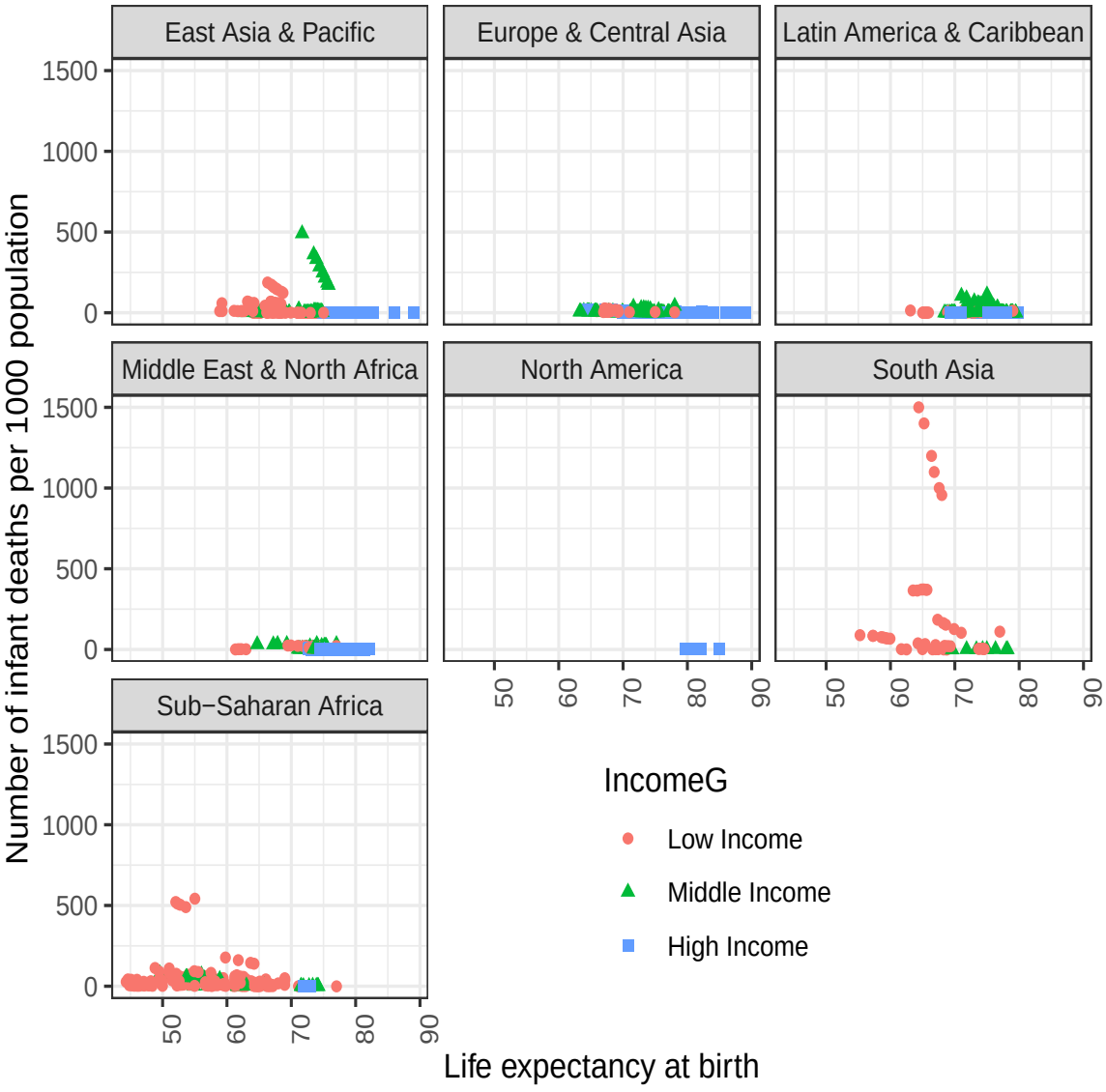


Figure 4.17: Scatter Plots of the Association Between Number of Infant Deaths & LE.

As a general observation, life expectancy seems to be higher in the high income countries and lower in the low income countries across all the regions of the world.

4.3.4 Model Performance Results

This study compared the performance of XGBoost model with that of two prominent machine learning models used by earlier studies on life expectancy data as shown in Table 4.4.

Table 4.4: RF, ANN & XGBoost Models' Predictive Performance Results.

Model	**MAE**	**RMSE**	**Runtime (in minutes)**
Random Forest (RF)	7.938	11.304	3.029
Artificial Neural Network (ANN)	3.86	5.002	51.057
eXtreme Gradient Boosting (XGBoost)	1.554	2.402	3.049

The XGBoost model outperformed the ANN and RF models, respectively, as indicated in Table 4.4 above.

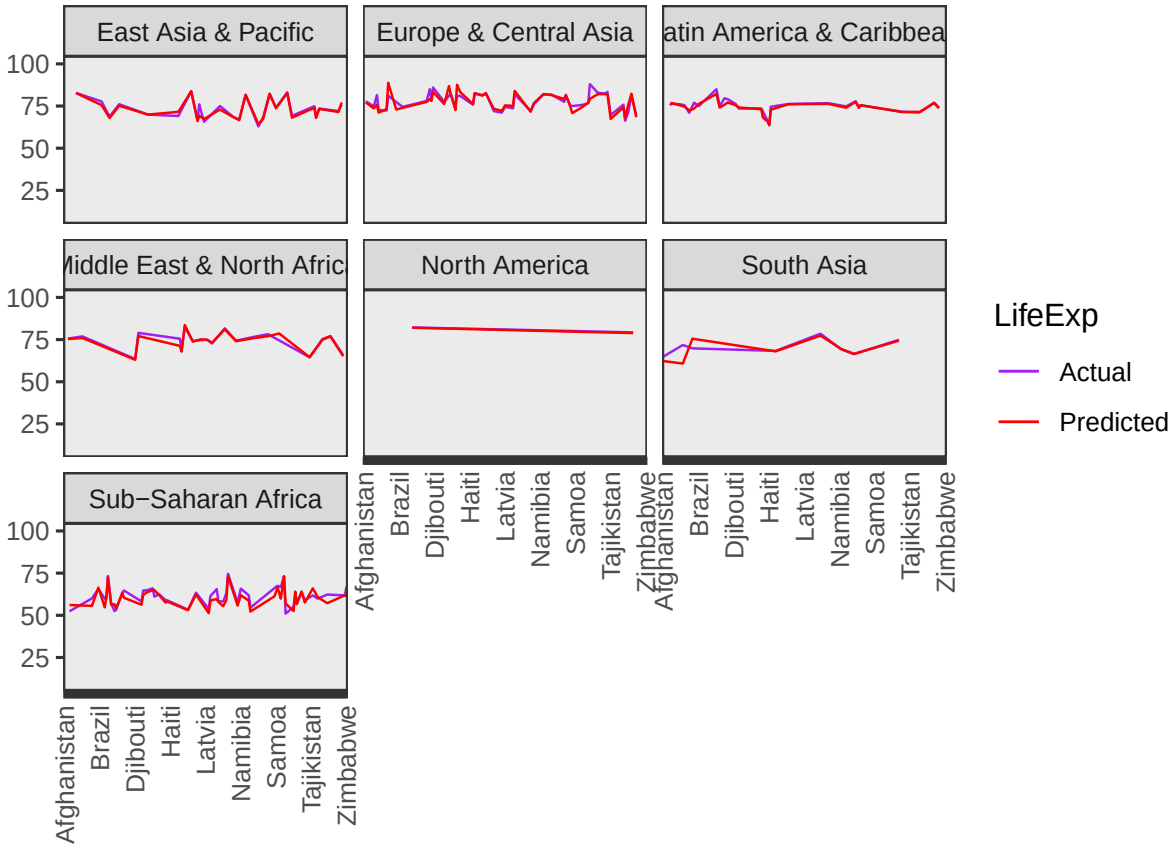


Figure 4.18: Regional Actual Vs. Predicted LE Values for the Year 2015 per Country.

The XGBoost and RF models were efficient in terms of the model training run-time, as shown by the model performance results in Table 4.4. On this front, the ANN model performed poorly.

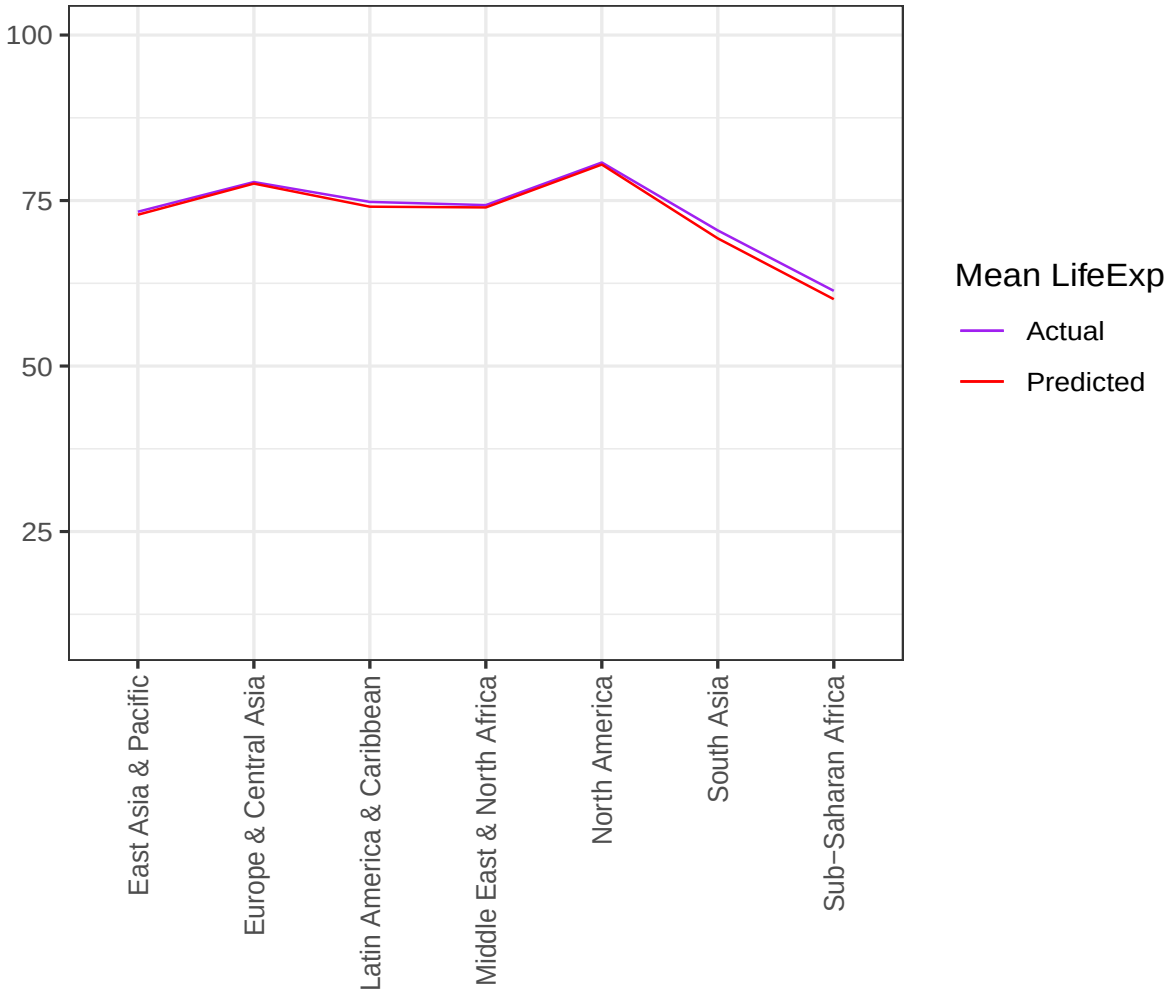


Figure 4.19: Actual Vs. Predicted Mean LE Values for the Year 2015 per Region.

The study set aside data for all nations for the year 2015 to be used in the confirmatory test of the best model. Figures 4.18 and 4.19 compare predicted and actual values in terms of life expectancy and mean life expectancy for the year 2015, respectively, and show that there is little fluctuation.

Chapter 5

Discussion of Research Findings

5.1 Introduction

The development, hyperparameter optimization, and validation components of the model, as well as the study outcomes in relation to the research objectives, are covered in this chapter. Model development is the process of generalizing and learning from a training dataset to create a mathematical representation. Hyperparameter optimization, on the other hand, is the process of determining the best collection of values for managing the machine learning process. Model validation is the process of evaluating a trained algorithm on a test dataset to determine how effective it is at predicting a result for fresh, unknown observations (Kassambara, 2018).

On 80% of the dataset, the XGBoost algorithm was utilized to create the model. The default parameter values were used to create the baseline model. The *gbtree* booster method was used, with 0.3 for learning rate (*eta*), 6 for maximum tree depth (*max_depth*), 0 for gamma, 1 for minimum child weight (*min_child_weight*), 1 for row subsample ratio (*subsample*), and 1 for column subsample ratio (*colsample_bytree*) as tree booster parameter values. Regression with squared loss objective (*reg:squarederror*) was used for the learning task parameter. A total of 100 trees were included in the model.

For hyperparameter tuning, the researcher used the grid search repeated cross-validation approach. A 10-fold cross-validation with two partitioning repeats yielded the best bias-variance trade-off. For model training efficiency, 6 cores were used to run cross-validation in parallel. After partitioning, the remaining 20% of the dataset was used to evaluate the model. The model accuracy was assessed using RMSE and MAE.

5.2 Key Determinants of Life Expectancy

Regional location, number of years at school, income group, number of under-five fatalities per 1000 population, percent of thinness among children aged 5-9, and average BMI were shown to be the major drivers of life expectancy in this study. These findings are consistent with earlier research. Kaplan et al. (2014) and Luy et al. (2019) discovered that rising educational levels were a determinant of rising life expectancy.

According to Szwarcwald et al. (2016), life expectancy at birth for women and men living in the wealthiest regions was 5 years higher than for those living in the poorest regions. Trpkova Nestorovska and Levkov (2019) discovered that gross national income had a favorable and statistically significant influence on life expectancy. According to a research conducted by Miladinov (2020), the country's population health and socioeconomic development had a significant impact on life expectancy at birth.

5.3 Association of Life Expectancy and its Predictors

Across all regions, the current study finds that life expectancy is greater in high-income nations and lower in low-income ones. These findings are explained by the fact that nations in the richest geographical areas, such as North America, have longer life expectancies. This is due to increased economic growth in such regions, as a result of better health care, social well-being, industrialization, and educational attainment levels.

Across all seven regions, life expectancy rises as the number of years spent in school increases. When compared to medium and low income nations, high income countries have a majority of their inhabitants spending more years in school. This is because having a high level of literacy allows people to make better life decisions, such as having more job prospects, having more negotiating power in terms of remuneration, and having healthy food and lifestyle habits, to name a few. This elevates the standard of living for citizens in a country.

As a metric of economic development, improved GDP per capita boosts life expectancy at birth through promoting economic growth and development in a country, resulting in a longer lifespan. High-income nations outperform their middle and low-income peers in terms of economic growth and development. This is due to increasing economic growth, higher living standards, and improved health in the first world countries.

Life expectancy rises when the number of under five deaths per 1000 people drops. High-income nations tend to have the fewest deaths among children under the age of five. This is due to increased access to better healthcare in advanced economies, particularly prenatal and postnatal care, as well as dietary issues. As a result, the risk of death for children under the age of five is quite low in these nations.

Low-income nations, on the other hand, have a higher rate of mortality among children under the age of five. The highest estimations are seen in Sub-Saharan Africa and South Asia. This is due to low-income nations' lag in terms of economic growth, employment rates, access to better healthcare, improved sanitation, and overall living standards.

As the percentage of thinness among children aged 5-9 years decreases, life expectancy increases. High-income countries have lower percentages of children in this cohort, with North American countries having the lowest percentages. Low-income countries, on the other hand, continue to have higher rates of childhood thinness, with Sub-Saharan Africa and South Asia topping the list. Thinness is connected to medical, societal, and economic difficulties (Suder et al., 2020).

Poor nutrition, caused by insufficient food and beverage consumption, a lack of available food and drink, chronic famine and food insecurity, is a key contributor to this occurrence, particularly in Sub-Saharan Africa and South Asia. The situation in high-income countries may be explained by the ambition of young girls to achieve a dreamy beauty of thinness, shown by the beauty modelling business (Tambalis et al., 2019).

This research reveals that when the average BMI of a population rises, so does its life expectancy. This contradicts the findings of prior studies that looked at the effect of BMI on life expectancy. This is an intriguing finding as obesity has been strongly linked to

an increased risk of cancer, hypertension, diabetes, and plenty of other noncommunicable illnesses.

Obesity is linked to a higher risk of disease and mortality, especially from heart disease and cancer. Obesity raises the risk of breast cancer recurrence and death in both premenopausal and postmenopausal women, as well as the risk of developing insulin resistance, and leads to decreased productivity, unemployment, and direct healthcare expenses (Abdelaal et al., 2017).

In Sub-Saharan Africa and South Asia, the majority of countries have BMIs that are between healthy (18.5–24.9 kg/m^2) and underweight (<18.5 kg/m^2). Poor nutrition in low-income countries may be responsible for the majority of underweight people in different parts of the world. Other causes may be accountable for the underweight populations in the few high and moderate income nations in East Asia and Pacific, Europe and Central Asia, as well as Latin America and the Caribbean, which are beyond the scope of this study.

Obesity (BMI > 29.9 kg/m^2) affects the majority of inhabitants of Europe and Central Asia, Latin America and the Caribbean, as well as those in the Middle East, and North Africa. This may be due to an increase in physical inactivity caused by the more sedentary character of many types of job, changing means of transportation, and rising urbanization in advanced economies. Apart from the fact that obesity is prominent in these locations, particularly in high-income countries, the increase in life expectancy might be due to better weight management via the adoption of active lifestyles and medical interventions.

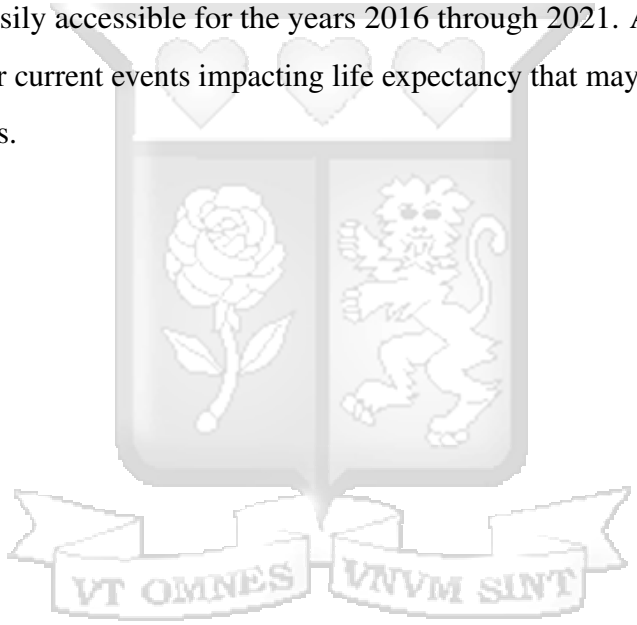
5.4 XGBoost Model Performance

On the test set, the XGBoost model attained MAE and RMSE values of 1.554 and 2.402, respectively. The number of optimal iterations attained was 500, with a learning rate of 0.3 and a maximum tree depth of 7. These findings outperform those published by Meshram (2020), that used the Random Forest regression model achieving an MAE value of 1.58. The

results from this research support the XGBoost regressor as a reliable and efficient model for estimating life expectancy across the globe.

5.5 Limitations of the Study

There are two limitations to this study. To begin with, only health, socioeconomic, and behavioral aspects were included in this study. Environmental issues such as pollution and climate change have an influence on life expectancy. Even so, inclusion of these components was outside the scope of this study. Furthermore, data for all of the factors included in this analysis was not easily accessible for the years 2016 through 2021. As a result, the model does not account for current events impacting life expectancy that may have happened in the UN member nations.



Chapter 6

Conclusion and Recommendations

6.1 Conclusion

Using the eXtreme Gradient Boosting algorithm, this research proposes a unique method for forecasting life expectancy. The results of this work speak for XGBoost as an efficient and trustworthy model for life expectancy estimation, having been proven in numerous industry applications. Additional studies may be explored by integrating environmental components in the model, in addition to updating the model with fresh data. Moreover, the average BMI could be investigated further to gain a deeper understanding of its favorable relationship with life expectancy beyond the approved healthy range.

6.2 Policy Recommendations

According to this study, economic growth, years spent in school, and a drop in under-five mortality as well as percent of thinness among 5-9 year olds all had a positive influence on life expectancy at birth. As a result, policy decisions that encourage economic development, higher literacy, improved nutrition, and decreased child mortality should be prioritized by governments.

Furthermore, in order to boost life expectancy estimates, this study recommends that governments and a wide spectrum of industry participants should use the formulated model to influence policy decisions in healthcare, education, and socioeconomic development.

References

Abdelaal, M., le Roux, C. W., and Docherty, N. G. (2017). Morbidity and mortality associated with obesity. *Annals of translational medicine*, 5(7).

Abdu-Aljabar, R., Dhia'a, Awad, O., and A (2021). A comparative analysis study of lung cancer detection and relapse prediction using XGBoost classifier. In *IOP conference series: materials science and engineering*, volume 1076, page 012048. IOP Publishing.

Alballa, N. and Al-Turaiki, I. (2021). Machine learning approaches in COVID-19 diagnosis, mortality, and severity risk prediction: A review. *Informatics in Medicine Unlocked*, 24:100564.

Alpaydin, E. (2020). *Introduction to machine learning*. MIT press.

Anderson, S., Auquier, A., Hauck, W., Oakes, D., Vandaele, W., and Weisberg, H. (1980). Statistical methods for comparative studies. New York, Chichester, Brisbane.

Arpaci, I., Huang, S., Al-Emran, M., Al-Kabi, M. N., and Peng, M. (2021). Predicting the COVID-19 infection with fourteen clinical features using machine learning classification algorithms. *Multimedia Tools and Applications*, 80(8):11943–11957.

Ayuso, M., Bravo, J. M., and Holzmann, R. (2021). Getting life expectancy estimates right for pension policy: period versus cohort approach. *Journal of Pension Economics & Finance*, 20(2):212–231.

Bakas, I. and Kontoleon, K. J. (2021). Performance Evaluation of Artificial Neural Networks (ANN) Predicting Heat Transfer through Masonry Walls Exposed to Fire. *Applied Sciences*, 11(23):11435.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer.

Chen, T. and Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 785–794.

Chen, Y. Q. and Cheng, S. (2006). Linear life expectancy regression with censored data. *Biometrika*, 93(2):303–313.

Clancy, C., Hecker, J., Stuntebeck, E., and O'Shea, T. (2007). Applications of machine learning to cognitive radio networks. *IEEE Wireless Communications*, 14(4):47–52.

Cruz, J. A. and Wishart, D. S. (2006). Applications of machine learning in cancer prediction and prognosis. *Cancer informatics*, 2:117693510600200030.

Cui, B. (2020). *DataExplorer: Automate Data Exploration and Treatment*. R package version 0.8.2.

Dias, N., Sucharitharathna, C., et al. (2017). Prediction of Life Expectancy. *American Scientific Research Journal for Engineering, Technology, and Sciences (ASRJETS)*, 34(1):252–260.

Donges, N. (2021). A Complete Guide to the Random Forest Algorithm.

Duda, R. O., Hart, P. E., and Stork, D. G. (2001). Pattern Classification second edition John Wiley & Sons. *New York*, 58:16.

El Naqa, I. and Murphy, M. J. (2015). What is machine learning? In *machine learning in radiation oncology*, pages 3–11. Springer.

Fanyin, H., Mazumdar, S., Tang, G., Bhatia, T., Anderson, S. J., Dew, M. A., Krafty, R., Nimgaonkar, V., Deshpande, S., Hall, M., et al. (2017). Non-parametric MANOVA approaches for non-normal multivariate outcomes with missing values. *Communications in Statistics-Theory and Methods*, 46(14):7188–7200.

Freeman, T., Gesesew, H. A., Bambra, C., Giugliani, E. R. J., Popay, J., Sanders, D., Macinko, J., Musolino, C., and Baum, F. (2020). Why do some countries do better or worse in life expectancy relative to income? An analysis of Brazil, Ethiopia, and the United States of America. *International journal for equity in health*, 19(1):1–19.

Girum, T., Muktar, E., and Shegaze, M. (2018). Determinants of life expectancy in low and medium human development index countries. *Medical Studies/Studia Medyczne*, 34(3):218–225.

Goals, G. (2022). The global goals.

Guillaume, W., Michel, M., and Duchene, J. (2002). *The Life Table: Modelling Survival and Death*. European Studies of Population 11. Springer Netherlands, 1 edition.

Gulland, A. (2016). Global life expectancy increases by five years.

Ho, J. Y. and Hendi, A. S. (2018). Recent trends in life expectancy across high income countries: retrospective observational study. *bmj*, 362.

Hou, N., Li, M., He, L., Xie, B., Wang, L., Zhang, R., Yu, Y., Sun, X., Pan, Z., and Wang, K. (2020). Predicting 30-days mortality for MIMIC-III patients with sepsis-3: a machine learning approach using XGBoost. *Journal of translational medicine*, 18(1):1–14.

Joseph, V. R. (2022). Optimal Ratio for Data Splitting. *Statistical Analysis and Data Mining: The ASA Data Science Journal*.

Josse, J. and Husson, F. (2016). missMDA: A Package for Handling Missing Values in Multivariate Data Analysis. *Journal of Statistical Software*, 70(1):1–31.

Kaplan, R., Spittel, M., and Zeno, T. (2014). Educational Attainment and Life Expectancy. *Policy Insights from the Behavioral and Brain Sciences*, 1:189–194.

Karacan, I., Sennaroglu, B., and Vayvay, O. (2020). Analysis of Life Expectancy Across Countries Using a Decision Tree. *Eastern Mediterranean Health Journal*, 26(2):143–151.

Kassambara, A. (2017). *Practical Guide to Principal Component Methods in R: PCA, M (CA), FAMD, MFA, HCPC, factoextra*, volume 2. STHDA.

Kassambara, A. (2018). *Machine Learning Essentials: Practical Guide in R*. STHDA.

Ketenci, N. and Murthy, V. N. (2018). Some determinants of life expectancy in the united states: results from cointegration tests under structural breaks. *Journal of Economics and Finance*, 42(3):508–525.

Korkmaz, S., Goksuluk, D., and Zararsiz, G. (2014). MVN: An R Package for Assessing Multivariate Normality. *The R Journal*, 6(2):151–162.

Kuhn, M. (2021). *caret: Classification and Regression Training*. R package version 6.0-88.

Lesnussa, Y. A., Rumlawang, F. Y., Risamasu, E., and Fhilya, C. (2020). Prediction of life expectancy in maluku province using artificial neural networks backpropagation. *J. Mat. Integr*, 16(2):75–82.

Li, S. and Zhang, X. (2020). Research on orthopedic auxiliary classification and prediction model based on XGBoost algorithm. *Neural Computing and Applications*, 32(7):1971–1979.

Li, W., Yin, Y., Quan, X., and Zhang, H. (2019). Gene expression value prediction based on XGBoost algorithm. *Frontiers in genetics*, page 1077.

Luy, M., Zannella, M., Wegner-Siegmundt, C., Minagawa, Y., Lutz, W., and Caselli, G. (2019). The impact of increasing education levels on rising life expectancy: a decomposition analysis for Italy, Denmark, and the USA. *Genus*, 75(1):1–21.

Mamidanna, S. K., Reddy, C., and Gujju, A. (2022). Detecting an Insider Threat and Analysis of XGBoost using Hyperparameter tuning. In *2022 International Conference on Advances in Computing, Communication and Applied Informatics (ACCAI)*, pages 1–10. IEEE.

Martín Cervantes, P. A., Rueda López, N., and Cruz Rambaud, S. (2019). A causal analysis of life expectancy at birth. evidence from spain. *International journal of environmental research and public health*, 16(13):2367.

Martin Cervantes, P. A., Rueda Lopez, N., and Cruz Rambaud, S. (2020). Life expectancy at birth in Europe: An econometric approach based on Random Forests methodology. *Sustainability*, 12(1):413.

Meshram, S. S. (2020). Comparative Analysis of Life Expectancy between Developed and Developing Countries using Machine Learning. In *2020 IEEE Bombay Section Signature Conference (IBSSC)*, pages 6–10. IEEE.

Miladinov, G. (2020). Socioeconomic development and life expectancy relationship: Evidence from the EU accession candidate countries. *Genus*, 76(1):1–20.

Minaee, S. (2019). An Introduction to the Most Important Metrics for Evaluating Classification, Regression, Ranking, Vision, NLP, and Deep Learning Models: Part 1-Classification and Regression Evaluation Metrics. *Towards Data Science*.

Mitchell, T. (1997). *Machine Learning*. McGraw Hill.

Nielsen, D. (2016). Tree boosting with XGBoost-why does XGBoost win" every" machine learning competition? Master's thesis, NTNU.

Nkalu, C. N. and Edeme, R. K. (2019). Environmental hazards and life expectancy in Africa: evidence from GARCH model. *Sage Open*, 9(1):2158244019830500.

Nketiah-Amponsah, E. (2019). The impact of health expenditures on health outcomes in sub-Saharan Africa. *Journal of Developing Societies*, 35(1):134–152.

Nyuytiymbiy, K. (2020). Parameters and hyperparameters in Machine Learning and Deep Learning.

OECD (2022). Health Status: Life expectancy at birth - OECD Data.

R Core Team (2022). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Raftery, A. E., Chunn, J. L., Gerland, P., and Ševčíková, H. (2013). Bayesian probabilistic projections of life expectancy for all countries. *Demography*, 50(3):777–801.

Raftery, A. E., Lalic, N., and Gerland, P. (2014). Joint probabilistic projection of female and male life expectancy. *Demographic research*, 30:795.

Roser, M., Ortiz-Ospina, E., and Ritchie, H. (2013). Life Expectancy.

Saygili, G. and Rathje, E. M. (2008). Empirical predictive models for earthquake-induced sliding displacements of slopes. *Journal of Geotechnical and Ggeoenvironmental Engineering*, 134(6):790–803.

Sedgwick, P. (2014). Spearman's rank correlation coefficient. *Bmj*, 349.

Shahbaz, M., Shafiullah, M., and Mahalik, M. K. (2019). The dynamics of financial development, globalisation, economic growth and life expectancy in sub-Saharan Africa. *Australian Economic Papers*, 58(4):444–479.

Shang, H. L. (2012). Point and interval forecasts of age-specific life expectancies: A model averaging approach. *Demographic Research*, 27:593–644.

Sonal (2021). Importance of Exploratory Data Analysis Before ML Modelling.

Suder, A., Jagielski, P., Piórecka, B., Płonka, M., Makiel, K., Siwek, M., Wronka, I., and Janusz, M. (2020). Prevalence and factors associated with thinness in rural Polish children. *International journal of environmental research and public health*, 17(7):2368.

Sun, X. (2021). Application and Comparison of Artificial Neural Networks and XGBoost on Alzheimer's Disease. In *Proceedings of the 2021 International Conference on Bioinformatics and Intelligent Computing*, pages 101–105.

Szwarcwald, C. L., Souza Júnior, P. R. B. d., Marques, A. P., Almeida, W. d. S. d., and Montilla, D. E. R. (2016). Inequalities in healthy life expectancy by Brazilian geographic regions: findings from the National Health Survey, 2013. *International journal for equity in health*, 15(1):1–9.

Tafran, K., Tumin, M., and Osman, A. F. (2020). Poverty, income, and unemployment as determinants of life expectancy: Empirical evidence from panel data of thirteen Malaysian states. *Iranian journal of public health*, 49(2):294.

Tambalis, K., Panagiotakos, D., Psarra, G., and Sidossis, L. (2019). Prevalence, trends and risk factors of thinness among Greek children and adolescents. *Journal of preventive medicine and hygiene*, 60(4):E386.

Trpkova Nestorovska, M. and Levkov, N. (2019). Determinants of Life Expectancy: Analysis of Southeastern European Countries. *Knowledge International Journal*, 31.

Tuli, S., Tuli, S., Tuli, R., and Gill, S. S. (2020). Predicting the growth and trend of COVID-19 pandemic using machine learning and cloud computing. *Internet of Things*, 11:100222.

UN (2021). The sustainable development goals report.

Valletta, J. J., Torney, C., Kings, M., Thornton, A., and Madden, J. (2017). Applications of machine learning in animal behaviour studies. *Animal Behaviour*, 124:203–220.

Vamathevan, J., Clark, D., Czodrowski, P., Dunham, I., Ferran, E., Lee, G., Li, B., Madabhushi, A., Shah, P., Spitzer, M., et al. (2019). Applications of machine learning in drug discovery and development. *Nature reviews Drug discovery*, 18(6):463–477.

Wade, C. (2020). *Hands-On Gradient Boosting with XGBoost and scikit-learn: Perform accessible machine learning and extreme gradient boosting with Python*. Packt Publishing.

Wang, H., Naghavi, M., Allen, C., Barber, R. M., Bhutta, Z. A., Carter, A., Casey, D. C., Charlson, F. J., Chen, A. Z., Coates, M. M., et al. (2016). Global, regional, and national life expectancy, all-cause mortality, and cause-specific mortality for 249 causes of death, 1980–2015: a systematic analysis for the global burden of disease study 2015. *The lancet*, 388(10053):1459–1544.

Wang, M.-X., Huang, D., Wang, G., and Li, D.-Q. (2020). SS-XGBoost: a machine learning framework for predicting newmark sliding displacements of slopes. *Journal of Geotechnical and Geoenvironmental Engineering*, 146(9):04020074.

Wang, S. and Ren, Z. (2019). Spatial variations and macroeconomic determinants of life expectancy and mortality rate in China: a county-level study based on spatial analysis models. *International journal of public health*, 64(5):773–783.

Wang, Y. and Ni, X. S. (2019). A XGBoost risk model via feature selection and Bayesian hyper-parameter optimization. *arXiv preprint arXiv:1901.08433*.

WHO (2022). Life Expectancy at Birth (years).

World Health Organization (2021). World Health Statistics 2021: Monitoring health for the SDGs, sustainable development goals. *The Global Health Observatory*, pages 1–121.

WorldBank (2022). Life Expectancy at Birth, Total (years).

Zhang, Y., Wang, Y., Xu, J., Zhu, B., Chen, X., Ding, X., and Li, Y. (2021). Comparison of prediction models for acute kidney injury among patients with hepatobiliary malignancies based on XGBoost and LASSO-logistic algorithms. *International Journal of General Medicine*, 14:1325.

Appendix A

A.1 Ethical Review Committee Report

Strathmore UNIVERSITY

30th May 2022

Mr Lipesa Brian,
brian.lipesa@strathmore.edu

Dear Mr Lipesa,

RE: Machine Learning Based Prediction of Life Expectancy

This is to inform you that SU-IERC has reviewed and **approved** your above **SU Masters'** research proposal. Your application reference number is **SU-IERC1331/22.** The approval period is **30th May 2022 to 29th May 2023.**

This approval is subject to compliance with the following requirements:

i. Only approved documents including (informed consents, study instruments, MTA) will be used
ii. All changes including (amendments, deviations, and violations) are submitted for review and approval by SU-IERC.
iii. Death and life-threatening problems and serious adverse events or unexpected adverse events whether related or unrelated to the study must be reported to SU-IERC within 48 hours of notification
iv. Any changes, anticipated or otherwise that may increase the risks or affected safety or welfare of study participants and others or affect the integrity of the research must be reported to SU-IERC within 48 hours
v. Clearance for export of biological specimens must be obtained from relevant institutions.
vi. Submission of a request for renewal of approval at least 60 days prior to expiry of the approval period. Attach a comprehensive progress report to support the renewal.
vii. Submission of an executive summary report within 90 days upon completion of the study to SU-IERC.

Prior to commencing your study, you will be expected to obtain a research license from National Commission for Science, Technology, and Innovation (NACOSTI) https://research-portal.nacosti.go.ke/ and obtain other clearances needed.

Yours sincerely,

for: **Dr Ben Ngoye,**
Secretary; SU-IERC

Cc: Prof Fred Were,
Chairperson; SU-IERC

A.2 Similarity Report

Curiginal

Document Information

Analyzed document	136562_MSc_Thesis_V1.pdf (D138430976)
Submitted	2022-05-30T05:55:00.0000000
Submitted by	
Submitter email	Brian.Lipesa@strathmore.edu
Similarity	1%
Analysis address	library.strath@analysis.urkund.com

Sources included in the report

W	URL: https://www.oecd-ilibrary.org/life-expectancy-at-birth_5kg23tkz1w40.pdf;itemId=/content/chapter/health_glance-2011-4-en Fetched: 2021-12-01T19:10:49.4230000	1
W	URL: https://www.researchgate.net/publication/342270212_SS-XGBoost_A_Machine_Learning_Framework_for_Predicting_Newmark_Sliding_Displacements_of_Slopes Fetched: 2020-10-26T08:09:10.8830000	4
W	URL: https://gwang.people.ust.hk/Publications/Wang.et.a.(2020)_XGBoost_NewmarkSlope_ASCE_JGGE_AcceptedMarch2020.pdf Fetched: 2022-05-30T05:57:08.5300000	5
SA	**airThs.pdf** Document airThs.pdf (D136202840)	1
SA	**Crowdfunding_Machine_Learning.pdf** Document Crowdfunding_Machine_Learning.pdf (D132355300)	3
SA	**MLReport_Group116.pdf** Document MLReport_Group116.pdf (D132354410)	1