AFRICAN CENTER OF EXCELLENCE IN DATA SCIENCE

COLLEGE OF BUSINESS AND ECONOMICS

A PREDICTICE MODEL FOR HEALTH INSURANCE PREMIUM RATES USING MACHINE LEARNING ALGORITHMS

By

ANGELA D. KAFURIA

Registration number: 220000008

A dissertation submitted in partial fulfillment of the requirements for the degree of Master of Science in Data Science in Actuarial Science

University of Rwanda, College of Business and Economics

Supervisor: Dr. Ignace H. KABANO

March 2022

DECLARATION

I declare that this thesis entitled **A predictive model for health insurance premium rates using machine learning algorithms** is the result of my own work and has not been submitted for any other degree at the University of Rwanda or any other institution.

Date: 20/03/2022

Names: Angela D. KAFURIA

Signature: 

APPROVAL SHEET

This dissertation entitled **A predictive model for health insurance premium rates using machine learning algorithms** written and submitted by **Angela D. KAFURIA** in partial fulfillment of the requirements of a degree for masters of science in Data Science majoring in **Actuarial Science** is hereby accepted and approved. The rate of plagiarism using TURNITIN is 12 % which is less than the rate accepted by ACE-DS.

Dr. Ignace H. KABANO

Supervisor

Dr. Ignace H. KABANO

Head of Trainings

DEDICATION

This work is dedicated to my late father who desired and supported me to always do my best in academic career. He gave me a profound base that I now stand where I am. My father is my hero.

R.I.P dad

ACKNOWLEDGEMENT

First, I will forever be gratefully to almighty God for the gift of life, each day and every step of my life is by his grace and blessings.

I want to express my sincere gratitude to my family for their unwavering support during my academic career. Their understanding and sacrifices have managed me to accomplish my course on time.

My heartfelt gratitude goes to my daughters, for it is with their understanding and patience that their mother could attend her studies in all academic years.

I give much respect and appreciation to all my lecturers and to the entire ACE-DS team for their sincere support to enlighten me with the new knowledge that has broaden my perspective towards Data Science. Mostly, I thank my supervisor Dr Ignace Kabano for his advice and endless provision of assistance during my studies. Dr Kabano has been a supportive lecturer and a mentor. I am proud to have him as my supervisor.

Lastly, I want to thank my sponsors because without them I would not manage to have taken this course. I will always be grateful to the Inter-University Council for East Africa (IUCEA) that have sponsored me for the two years my studies.

ABSTRACT

Universal health coverage is a crucial step to ensure the good health and wellbeing of members of any society. However, in developing countries like Tanzania, health care systems are highly reliant on out-of-pocket payments, a mechanism that is a barrier to universal health coverage because it contributes to inefficiencies, inequity, and cost. To solve this challenge, people are encouraged to enroll in health insurance schemes to reduce the burden of out-of-pocket payments whenever they suffer from an illness or have pre-existing disease conditions. On the other hand, insurance companies are advised to charge insurance premium rates that are affordable by many people to guarantee universal health care coverage. Thus, there is a strong need for insurance companies to develop models that accurately predict medical expenses for the insured population. This study used demographic and behavioral data to formulate a predictive model to determine health insurance charges using Machine learning algorithms techniques. Additionally, the study evaluated the performance of five machine learning models in predictive analysis; K-nearest Neighbors (KNN), Least Absolute Shrinkage and Selection Operator (LASSO), Multiple Liner Regression (MLR), eXtreme Gradient Boosting (XGboosting), and Random Forest Regression (RFR).

Multiple linear regression tests found that the following variables were significant; age ($p = 0.000$), BMI ($p = 0.001$), smoking ($p = 0.000$) and region (0.046). Therefore, these attributes can be said to be the determinants of health insurance charges. The model performance evaluation findings XGboosting and RFR were the best models in prediction with the following values $R^2 =$ 0. 855, MAE = 2688.2, RMSE = 4748.7 and $R^2 = 0.853$, MAE = 2726.4, RMSE = 4783.8 respectively. Insurance companies that seek to develop a model for prediction premiums are recommended to use XGboosting and RFR for a more accurate model.

Keywords: Premium rates, Machine Learning, Predictive model, Health insurance

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF SYMBOLS AND ACRONYMS

BMI	Body Mass Index
CART	Classification and Regression Trees
CHF	Community Health Fund
DT	Decision Trees
XGboosting	eXtreme Gradient Boosting
GNB	Gaussian Naive Bayes
HCHN	High-Cost High-Need
KNN	K-Nearest Neighbour
LASSO	Least Absolute Shrinkage and Selection Operator
LSLR	Least Squares Linear Regression
MAE	Mean Absolute Error
ML	Machine Learning
MLR	Multiple Linear Regression
NHIF	National Health Insurance Fund
RF	Random Forest
RMSE	Root Mean Squared Error
RNN	Recurrent Neural Networks
SDGs	Sustainable Development Goals
SVM	Support Vector Machines
UHC	Universal Health Coverage
UN	United Nations
UNDP	United Nations Development Programme
WHO	World Health Organization

CHAPTER ONE: INTRODUCTION

1.1 Introduction

This chapter provides background information on health insurance, as well as a statement of the problem, research aims, research hypotheses, study importance, the study scope, and thesis structure.

1.2 Background information

Universal Health Coverage is a crucial step to ensure good health and well-being of members of any society. Universal Health Coverage is defined as a coverage of good health services from health promotion to prevention, treatment, rehabilitation and palliation as well as coverage with a form of financial risk protection. A third feature is universality coverage should be for everyone (Evans et al., 2013). It is all about ensuring all people can use the promotive, preventive, curative, rehabilitative, and palliative health services they need, of sufficient quality to be effective, while also ensuring that the use of these services does not expose the user to financial hardship (Taylor, 2015).

Good health and wellbeing are aspects mentioned in one of the 17 Sustainable Development Goals (SDGs) designed by the United Nations (UN) to be a "blueprint to achieve a better and more sustainable future for all" (UN, 2015). According to the UN 2030 Agenda for Sustainable Development, nations were required to ensure healthy lives and promote well-being for all, at all ages (goal 3) (UN, 2010). Universal Health Coverage was proposed by World Health Organization (WHO) to ensure good health and wellbeing worldwide. It is being used as a way to reduce fragmentation of health insurance coverage and lead to a single national health insurer (Lee et al., 2018). It is expected to have a positive contribution to the 2030 agenda for SDGs which pledge not to leave anyone behind on good health provision (UNDP, 2019).

In impoverished nations like Tanzania, health-care systems rely significantly on out-of-pocket expenditures, the mechanism that is a barrier to universal health coverage, as it contributes to inefficiency, inequity, and cost (Tungu et al., 2020). One of the ways that people in various nations pay for their medical requirements is through health insurance, which protects against the possibility of incurring medical and related financial bills. (Ho, 2015).

Health insurance is one of the mechanisms that can be used to ensure people have good health and wellbeing and hence increase UHC. However, around the world, not all people are covered by health insurance. For example, In Tanzania as of 2019, only 32% of the entire population had access to health insurance coverage where there are several insurers such as the Community Health Fund (CHF) which covers 23% of the population, National Health Insurance Fund (NHIF) which covers 8% and the rest were covered by private health insurances (Manzi et al., 2012).

In order to increase coverage of health care insurance, health care premiums paid have to be realistic and attractive to people who will subscribe to the insurance scheme provided. This study adopted machine learning techniques to predict health care insurance premiums. Machine learning, a type of artificial intelligence (AI) is emerging data analytics in computer science that has a potential to improve predictions of healthcare premiums when large amount of data and variables are provided. Moreover, this technique can be widely applied in other aspects of healthcare sector such as medical imaging diagnostics, improved radiotherapy, personalised treatment, crowd sourced data gathering, smart health records, ML-based behavioural modification, clinical trials, and medical research. (Verma & Verma, 2022).

1.2 Motivation

Good health and wellbeing are a priority to every human being, and that makes it a worldwide priority to ensure that all people have access to health insurance coverage. However, due to the high rates that are charged many people, especially in developing countries are without health insurance and so fail to access health services which results in high death rates. Most people who have no permanent job for instance farmers, pastoralism, and small traders cannot afford to pay for good health insurance. Either people get coverage to less proper service due to the contribution they make or are forced to use cash which makes them not capable of attending regional hospitals for quality services (National Health Policy, 2017). Douven et al (2020) suggested that one way to encourage enrolment is to have rates that are affordable for many people and that give quality service to its clients. There comes a need for a fair premium calculation model that suits the unique population factors. In line with the above argument, I designed this study hoping that the findings will contribute to the efforts of developing accurate health insurance premiums that will eventually ensure universal health coverage to all.

1.3 Problem statement

Insurance companies need to make money by collecting more annual premiums than they spend on the medical expenses of their beneficiaries, hence making a profit and continuing to stay in insurance businesses. On the other hand, the premium charges need to be affordable for a large segment of the population to ensure universal health care coverage. Thus, there is a strong need for insurance companies to develop models that accurately predict medical expenses for the insured population.

The premiums are calculated based on the likelihood of certain occurrences occurring among a group of people (Greenlaw & Shapiro, 2011). However, the medical and other associated costs are difficult to estimate because medical conditions varies greatly from each other (Lantz, 2019). Another complex part of estimating medical expenses is that the occurrence of certain diseases differs from one person to another and from one segment of the population to the other. For example, people living in warm climates are more susceptible to diseases such as Malaria than those living in cold areas. Smokers are more likely to suffer from Lung cancer than non-smokers, and less exercising people are more likely to suffer from heart diseases than those who often exercise.

Thus, this study used demographic and behavioral data from the patients to predict health insurance premiums. The use of predictive analysis is expected to be able to improve premium pricing accuracy and build customized health insurance plans. The study used machine learning algorithms such as K-nearest Neighbors (KNN), Least Absolute Shrinkage and Selection Operator (LASSO), Multiple Liner Regression (MLR), eXtreme Gradient Boosting (XGboosting), and Random Forest Regression (RFR) to develop a predictive model. It compares the performance of several models to find the most suitable one.

1.4 Study objective

1.4.1 General objective

This study's main goal is to apply machine learning techniques to create a model for health insurance premiums based on demographic data and behavioral data.

1.4.2 Specific objectives

1) To analyze the determinants of health insurance charges among health insurance beneficiaries
2) To develop a model to predict health insurance premium using demographic and behavior data
3) To evaluate the performance of predictive models that use machine learning algorithms to predict health insurance premiums

1.5 Research questions

This study was designed to answer the following research questions

1) What demographic and behavioural variables influence health insurance premiums charged by health insurance companies?
2) Which models can be used in developing and predicting health insurance premiums charged by insurance companies?
3) Which Machine Learning models have high ability to predict health insurance premiums charged by insurance companies?

1.6 Study scope

The study evaluated five (5) ML prediction models to determine the most accurate model for the prediction of health insurance charges. The dataset contains patient data collected in a single hospital. Demographic and behavioral variables were used in the study.

1.7 Significance of the study

One of the most significant responsibilities of health insurance companies is to determine the policy premium and develop accurate premium plans for their customers. A proper premium plan is more likely to increase health insurance uptake, especially in developing countries such as Tanzania. Thus, the main goal of this study is to use regression models to predict insurance premiums based on demographic and behavioral data collected from health insurance members.

The study is expected to develop a model that gives better calculation of health premium rates which will be accepted and used by many people in Tanzania. Hence, the models evaluated will enable the insurers, including private insurance companies to make accurate premiums predictions and proper health insurance for customer segmentation.

The study will also contribute to the knowledge of machine learning in the prediction of health-related cost.

Forecasting health insurance expenses on a variety of criteria to aid insurance policy makers in attracting customers and save time when creating plans for each individual. (Hanafy & Mahmoud, 2021). The predictive modeling has tremendous benefits for the health insurance industry in determining how much the premium should be charged to the insured person based on his/her behaviors and health habits (Kaur, 2018).

This study can also be used by life insurance companies that provide life insurance schemes to their clients. The study will help policymakers come up with better policy that enforce majority of the population to have life insurance. More over this study contributes to health insurance literature.

CHAPTER TWO: LITERATURE REVIEW

2.1 Introduction

In this fragment, I briefly review various research areas related to health insurance and the status of health insurance in developing countries including Tanzania. It first describes the key terms used in the study. Then it goes on to theoretical literature reviews where empirical studies relating to the evaluation of the predictive models are discussed. The last section of this chapter presents the study conceptual framework and hypothesis

2.2 Definition of key terms

Health

Health is a state of complete physical, mental and social well-being and not merely the absence of disease or infirmity (WHO, 2008). However, this definition of health has been challenged as being vague by the article by Huber et al (2011) that instead introduced a new concept of health as the ability to adapt and to self-manage, in the face of social, physical, and emotional challenges.

Insurance

Insurance refers to a method that households and firms use to prevent any single event from having a significant detrimental financial effect (Greenlaw & Shapiro, 2011). In a legal context, a contract of insurance is that whereby one party, the insurer, undertakes, for a premium or an assessment, to make a payment to another party, the policyholder or a third party, if an event that is the object of risk occurs (Outreville, 1998).

Insurance premiums

Insurance premiums are the regular payments that are made by households or firms with insurance to the insurance company (Greenlaw & Shapiro, 2011). Nurul (2013) defines premium as the price or amount of money an insurer collects from its clients to cover the client's unpredicted risk which are called claims. If the policyholder decides to pay a periodic premium, that arrangement is termed as a discrete contingent payment plan, meaning that the payment is from time to time for as long as the policyholder lives. If the policyholder decides to pay once, the arrangement is termed as a single initial premium (Bernard et al., 2017).

Universal Health coverage

The World Health Organization (WHO) defines Universal Health Coverage as ensuring that all people can use the promotive, preventive, curative, rehabilitative, and palliative health services they need, of sufficient quality to be effective, while also ensuring that the use of these services does not expose the user to financial hardship (Taylor, 2015; WB, 2021).

Machine learning

Machine learning is a type of artificial intelligence that provides computers with the ability to learn without being explicitly programed, making the process more accurate at predicting outcomes (Kalali et al., 2019). In this technique a computer is provided with large amounts of data to learn its own patterns, rather than the patterns and limits set by a human programmer, and therefore more improved results (Jokerst & Gotway, 2005).

2.3 Health Insurance process

Buhlmann (1984) viewed the insurance process as an input-output system and discussed how premium and interest add up a surplus but at the same time, claims and costs reduce that surplus. Meaning that for the stability of an insurance company the claims and cost should not be higher than the premium and interest. When assumed that the premium, claim, interest and cost are paid at the end of each year. The relationship between them is given below:

$$R = (I + P) - C - S$$

Where; R is Surplus or Equity per year, S is claims per year, I is the Interest rate per year, P is Premium per year, and C is Cost per year.

2.4 Health insurance schemes

The health coverage in developing countries is still very low. The main reasons are lack of awareness of the benefits of health insurance and health insurance premiums being too high for a majority to afford. However, the situation is improving in many countries as they are trying to move towards universal health coverage and social health insurance through the adoption of a various health insurance schemes.

Different countries have been utilizing various models of insurance and financing schemes to pay for medical services based on their respective socioeconomic realities and cultural contexts.

There are four main types of health insurance including National health insurance, community health insurance, Informal micro insurance schemes and Private health Insurance. These healthcare insurance plans are outlined by Mtei et al., (2007)

National Health Insurance Fund (NHIF)

This is type of healthcare service established by national insurance fund targeting employees working in both public and private sector. The enrollment in the NHIF is automatic and mandatory for all formal public sector employees (comprising civil servants, other government workers, and their dependents). As part of its benefits package, the NHIF provides both inpatient and outpatient care. The Tanzania NHIF was established in 1999, began its operations in 2001, and currently covers all public servants at both central and local government levels together with up to five family members.

Community Health Fund (CHF)

This the type of healthcare service established by national insurance fund targeting people living in rural areas and those working in an informal sector established by government insurance fund. It is a voluntary health insurance scheme, with members entitled to access services at the primary health facilities. The aim of establishing this type of scheme is not primarily to make profit from provision of health care service but rather to improve access to health care for the poor and vulnerable groups. In Tanzania, the CHF was established in 1996 as a possible mechanism granting access to basic health care services to populations in the rural areas and the informal sector.

Informal Micro Insurance and Community Based Health Financing Schemes

The micro insurance and community-based health financing schemes are schemes established by microfinance institutes and NGOs in both rural and urban areas to cover low-income individuals. The services that are covered by these schemes include primary health care, outpatient services, reproductive health, and minor surgery. Membership in these schemes is voluntary and the membership fee varies from one scheme to another. The number of smaller informal micro-insurance schemes has increased over time in Tanzania. Several schemes are now registered under the Tanzania Network of Community Health Funds (TNCHF), although many others choose not to register.

Private Health Insurance

The private health insurance is the type of insurance offered by the formally registered private companies. They might be local companies or multinational companies. Worldwide, there are many registered insurance companies, of which some have a health insurance component. In Tanzania, private health insurance coverage offered by these insurances is mostly at the hospital level rather than the dispensary or health Centre level.

2.5 Importance of health insurance

In every country, some people are unable to pay directly or out of pocket for the healthcare services they need, or financially they may be seriously disadvantaged by doing so (Ho, 2015). Thus, health insurance is very important as it ensures universal health care coverage. It is recommended in many countries as one of ensuring better health and wellbeing of people. It is especially beneficial in low-income areas since it saves insured people from paying excessive treatment costs in the event of disease by covering medical expenses incurred as a result of illness. (Wang et al, 2010). These charges could be related to drug costs, medical consultation fees, or hospitalization payments.

The purchase of health insurance reduces the risks and unpredictability inherent in a consumer's health care expenses. The consumer pays for a health insurance policy and then is subsequently (partly) reimbursed for his or her future expenditures on health care (Rapaport, 2015). The study by Tungu et al (2020) found that there was a positive statistical association between health insurance and the use of out-and inpatient services. Individuals and households were also shielded from catastrophic health costs through the application of both horizontal and vertical equality in the deployment of outpatient and inpatient care services.

The use of Health insurance has been proven to significantly improve maternal health. It is important to women due to their biological genetics, hence tend to be more helpful to them during pregnancy and when giving birth (Adebayo et al., 2015). A study by Kibusi et al (2018) found that women with health insurance were more likely to have the first antenatal appointment scheduled correctly and to have expert birth assistance at the time of delivery

The elderly can benefit from health insurance since it makes it easier for them to access medical treatments.

A study by Tungu et al (2020) found that in rural Tanzania, there is a positive substantial relationship between health insurance and the use of outpatient and inpatient care by the elderly.

2.6 Empirical literature review

Several studies have been done by other researchers evaluating several predictive regressions models on health insurance. Most of these studies have used demographic data as well as behavioural data. The demographic and behavioural characteristics were found to be among the important factors that influence health insurance premium rates. Some of them have a direct impact and some have an indirect impact on the health premium calculation.

Lakshmanarao, Koppireddy, and Kumar (2020) conducted a predicate analysis on the medical health insurance cost of a person based on gender age, smoking habit, body mass index (BMI), number of children, and region, using the medical information and costs dataset from Kaggle. Using machine learning techniques, the study applied four regression models to the dataset; Multiple Linear Regression, Support Vector Regression, Decision Tree Regression, and Random Forest regression. The study results indicated that, among the four algorithms, Random Forest Regression gives better results. Also, age and BIM were features with a strong influence on medical insurance charges.

Hanafy and Mahmoud (2021) employed machine learning regression models and deep neural networks to anticipate health insurance premiums based on age, sex, BIM, number of kids, smoking, and region of the person living for medical costs. The dataset was obtained from Kaggle.com. The models used were Multiple Linear Regression, Generalized Additive Model, Support Vector Machine, Random Forest Regressor, Classification and Regression Trees, XGBoost, k-Nearest Neighbors, Stochastic Gradient Boosting, and Deep Neural Network. The study demonstrated how different models of regression could forecast insurance costs. The findings showed that Stochastic Gradient Boosting offered the best efficiency.

Kaur (2018) predicted the insurance premium charge based upon other attributes (age, BMI, smoking, number of children a person has) using multiple linear regression, random forest, and Neural Network. The findings indicated that smoking has the highest impact on health insurance charges followed by BMI and age. The findings showed that the neural network did a better job of predicting the insurance charges.

Another predictive study that used machine learning models was the study by Yego, Kasozi, and Nkurunziza (2020). In this study seven (7) machine learning models (Logistic Regression Classifier Logistic, Support Vector machines (SVM), Gaussian Naive Bayes (GNB), K-Nearest Neighbor (KNN), Decision Trees (DT), Random Forest Regression (RFR) and XGboosting were compared for their performance in predicting health insurance premium. Attributes used in the study were sex, wealth quintile, region, education level, age group, household size, marital status, ownership of a phone, ownership of smartphone, most trusted providers, nature of residence, numeracy, having a set of an emergency fund, having electricity as a light source, having a bank product, urban versus rural and being a youth. This study used 2016 Kenya FinAccess Household Survey data that was used for comparison of performance in both over-sampled and under-sample data. For the over-sampled data, Random Forest showed the highest accuracy and precision but for under-sampled data, XGBoosting was optimal. The most important feature in prediction was 'having a bank product' followed by wealth quantile, region a person is living in, and education level.

Another study that used predictive machine learning models to forecast the expenditures, especially for the high-cost high-need (HCHN) patients was the study by Yang et al (2018). This study examined administrative insurance claims from the Medicaid program of the state of Texas, USA. Four predictive models were applied to forecast the patients' expenditures based on the previous periods, including ordinary least squares linear regression, Least Absolute Shrinkage and Selection Operator (LASSO), gradient boosting machine (GBM), and recurrent neural networks (RNN). The study used multiple features including Demographic variables (age, sex, race/ethnicity, and disabled status), diagnoses, medical procedures, and medications. Findings showed that additional information such as clinical information and demographics are useful to improve prediction performance. LASSO and GBM were found to be more effective in generating interpretable contributions and finding.

In Killada (2017) four regression models were evaluated for individual health insurance expenses. The models were Multiple Linear Regression, Decision tree Regression, AdaBoost Regression, and Gradient Boosting Decision Tree Regression. The study used health insurance marketplace data from 2014, 2015, and 2016 to develop the four regression models, and the predicted premiums. The data were drawn from marketplace Public Use Files released by the Center for Consumer

Information and Insurance Oversight, USA. To test and verify the model, the study used the data from 2014, 2015, and 2016 as inputs for training the models, and the predicted premiums were compared with the actual 2017 data. Features used in the study were age, family size, region a person is living (county), maximum out-of-pocket, and metal level. The findings indicated that family size and age had significant effects on premiums. The maximum out-of-pocket and deductible showed a negative sign, indicating that these variables negatively correlated with premiums. That means the higher the premium lower the deductible and the higher the maximum out of pocket the lower the premium. Also, the study found that the Adaboost model which is built upon a decision tree is the best performing model.

According to Adebayo et al (2015), there is a relationship between age, past medical history, body mass index, and health premium calculation. Old people are more exposed to diseases than the youth ones and for that old people will be willing to pay a higher premium for their health coverage than the young ones. Also, those with a long history of suffering from a certain illness or having pre-existing disease conditions are often charged higher than those who do not have. Strawiński and Celińska-Kopczyńska (2019) presented occupation as another factor that affects health insurance premiums. A person working in the office has less accident risk compared to those working in construction so higher premium for those in construction. Besides occupation, other factors discussed in that study are economic status and the type of plan chosen. Regarding economic status, some people are wealthier than others, hence they are more likely to choose first-class health insurance, they pay high premiums rates to cover their health care services.

CHAPTER THREE: METHODOLOGY

3.1 Introduction

This chapter explains the research methods employed in this study. Firstly, the chapter presents the variables that were used in the study. Secondly, it defines the type of data used and the dataset source. Thirdly, the data preprocessing and analysis procedures are explained in the preceding subsections.

3.2 Study variables

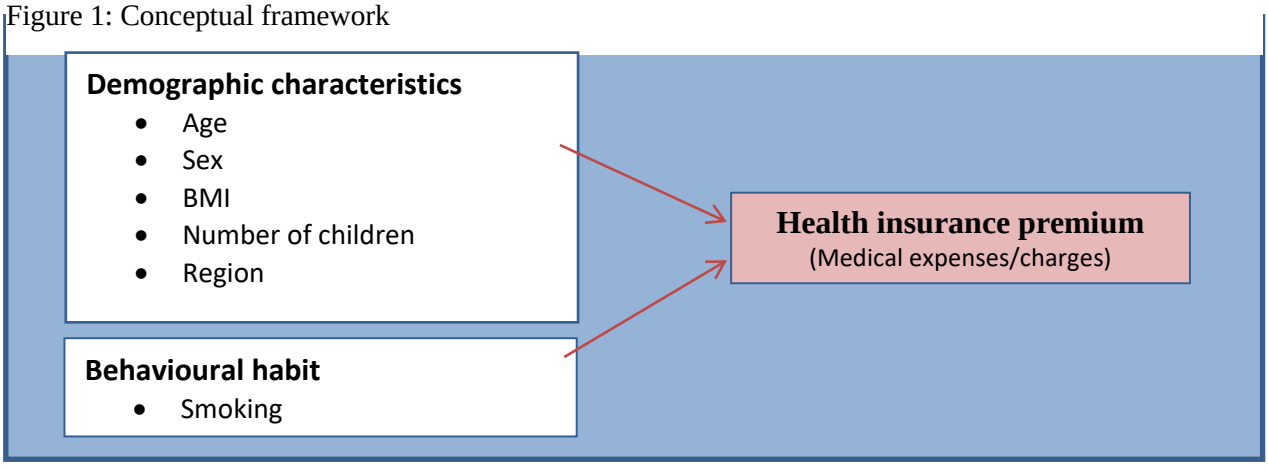
Previous studies have identified several variables that determine the health insurance premiums payment. Based on the reviewed studies, several direct and indirect variables were identified as presented in the table 1.

Table 1: Variables affecting premium payment according to previous studies

Direct variables	Indirect variables
Age	Occupation
Sex	Economic status
BMI	Smoking
Past medical history	Types of plans chosen
Education level	Region a person live
	Number of children/Family size

Among the variables presented above (Table 1), six (6) variables were selected to form a conceptual framework for this study; smoking, age, sex, BMI, number of children, and region a person lives as the independent variable while health premium paid by a person who is insured stood as dependent variable. Therefore, the study conceptual framework was developed based on the literatures (Figure 1).

Figure 1: Conceptual framework



3.3 Data source

The focus of our thesis is to come up with a model that will assist health insurance companies to calculate better premium for people of different categories. Our study proposes to use secondary data from health insurance companies. The researcher planned to use data from health insurance companies or Tanzania public health insurance schemes, NHIF. However, due to the limitation of the availability of data from these sources, it was later decided to use secondary data from online sources. Thus, the dataset used for experiments was collected from Kaggle.com (machine learning repository). The dataset contained medical information and costs billed by the health insurance company. It had 1339 rows and 7 columns. The following are the columns (variables) in the dataset; age, gender, BMI, number of children, smoking, region, and insurance charges (Table 2). In regression analysis, independent variables are used to predict the value of a dependent variable. While the age, gender, BMI, number of children, smoking, and region are treated as independent variables, insurance charge was an independent variable.

Table 2: Description of the variables in the dataset

Variable	Description	Data type		R - Data type (atomic class)
		Data type	Categories	
Age	The principal beneficiary's age	Continuous		Integer
Sex	The major beneficiary's/sex Contractor's (male or female)	Categorical (binary)	Male Female	Character
BMI	Body Mass Index, providing an understanding of body weights that are relatively high or low relative to height, objective index of body weight using the ratio of height to weight	Continuous		Numeric
Smoking	The smoking habit of insurance beneficiary (smoking or not)	Categorical (binary)	Yes No	Character
Children	Number of dependents, number of children covered by health insurance	Continuous		Integer
Region	The residential area of the beneficiary	Categorical	Northeast Southeast Southwest Northwest	Character
Charges/ Expenses	Individual insurance premiums billed by health insurance	Continuous		Numeric

3.4 Data analysis

The data analysis was conducted using R-software. R is powerful data analysis software and has been widely applied in regression analysis.

3.4.1 Data preparation

3.3.1.1 Dealing with missing data and duplicated values

The presence of missing data leads to wrong results while performing any functions. Availability of missing values prevents the application of machine learning algorithms. Also, duplicated values can lead to accurate results so we need to eliminate them. In our data, there were no missing values but one value was duplicated, so we only took the distinct values.

3.4.1.2 Checking for multicollinearity

The relation between predictors or independent variables was explored. Whenever an independent variable is highly correlated with one or more of the other independent variables, it can be said that a Multicollinearity problem exists (Allen, 2007). The 'Pearson' t-test was conducted to find out if the correlation among the predictors (independent variables) was significant. The test results showed no significant correlation among the independent variables. Therefore, it was concluded that the issue of multicollinearity in the dataset did not exist.

3.4.1.3 Categorical data conversion

The following variables were nominal categorical data; sex, smoker, and region. The Linear and KNN models require that all predictor variables be numeric, categorical data cannot be properly handled by this model. To get better performance of this model the categorical data were transformed into numerical data by using a dummy encoding technique which leaves one group out (the first level of the factor) and creates a new column for all other groups coded 1 or 0 depending on whether the original variable represented that value or not. The tree-based models; Random Forest and Gradient boosting model naturally handle numeric or categorical predictors. However, even tree-based models can benefit from preprocessing categorical predictors.

3.4.2 Exploratory data analysis

To explain the basic properties of the data in a study, descriptive statistics were used. Simple descriptions of the sample and metrics were provided in the findings.

Following descriptive statistics, the causal relationship between the independent variables and the dependent variable was investigated.

3.4.3 Predictive Modelling

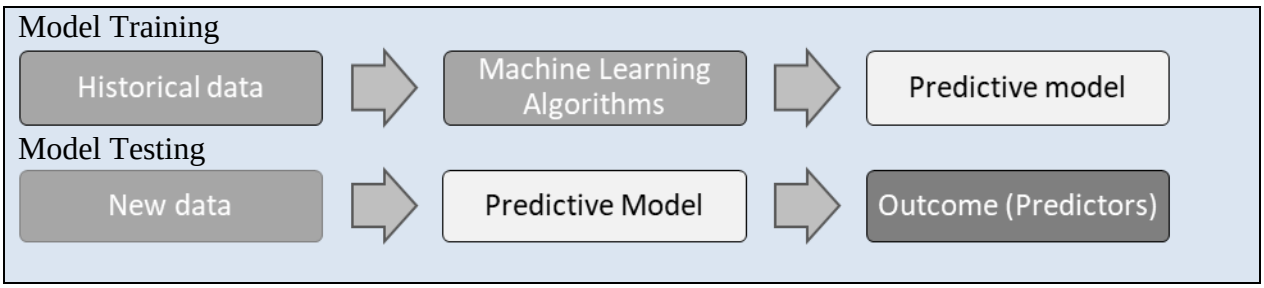
In this study predictive analysis was done using the ML regression analysis. ML regression models allowed the study to concentrate on prediction by using general-purpose learning algorithms and to find patterns, hence increase accuracy of the predictions (Bzdok et al., 2018; Edgar & Manz, 2017). Other options that were available were conventional statistics such as simple liner regression analysis, logistic regression analysis, ANOVA and t-tests. The main potential pitfall of the conventional statics methods is that the link between input and output is user chosen and may result in less accurate prediction model if the actual input–output association is not well represented by the chosen model (Ley et al., 2022).

3.5 Machine Learning Algorithms

3.5.1 Structure of Machine Learning-based predictive model

Machine learning algorithms build a model based on the "training data", to make predictions or decisions without being explicitly programmed to do so. The model is trained from historical data and the outcome is generated for the new test data. It involves two phases; Model training and Model testing (Bhadja & Abhangi, 2018; Sarker, 2021). This is how it was conducted; The dataset was divided into two parts, with the first being used for model training whereby 80% of the total data was used as training data. This data was also termed as historical data. The training data contain input and target values. The rest of the data, termed as new data, was used for model testing. The algorithm picked up the pattern and map the input values to the output and use it for prediction (Figure 2)

Figure 2: General Structure of a Machine Learning-based predictive model



Source: Sarker (2021)

3.5.2 Types of machine learning algorithms used in the study

The following predictive models were used in this study; Multiple regression analysis (MRL), K-nearest Neighbors (KNN), Least Absolute Shrinkage and Selection Operator (LASSO), Extreme Gradient Boosting, and Random Forest Regression (RFR).

3.5.2.1 Multiple Linear Regression (MLR)

The MLR model helped to analyze the relationship between insurance charges which is here treated as dependent variable (outcome) and independent variables (predictors) such as age and sex. This model will get us a more precise and accurate understanding of the association of each predictor with the outcome. We will see a linear relationship between the dependent variable (Y) and independent variables (x) (also known as a regression line) using the best fit straight line. The general equation for linear regression is as follows

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots\ldots + \beta_n x_n \ldots\ldots (2)$$

Where a is the intercept, b is the slope of the line, and e is the error term.

Multiple linear regression (MLR) is an extension of simple linear regression where there is one dependent variable (Y) and two or more independent variables (x_1, x_2, x_3, …, x_n). For this study, the values of x (independent variables) were as follows; x_1= Age, x_2 = Sex, x_3 = BMI, x_4 = children, x_5 = smoker, x_6 = region

$$\text{Insurance charges } (y) = \beta_0 + \beta_1 \text{Age} + \beta_2 \text{ Sex} + \beta_3 \text{ BMI} + \beta_4 \text{ Children} + \beta_5 \text{ Smoker} + \beta_6 \text{ Region} \ldots\ldots (3)$$

In this dataset, the dependent variable is medical charges and independent variables are age, gender, smoker, BMI, children, region. Most of the researchers have been using a generalized linear model (GLM) for the health premium prediction because of its simple interpretability of the fitted parameters. This study used a supervised learning technique under machine learning. This is because with supervised ML there is a more accurate model compared to GLM.

3.5.2.2 K-nearest Neighbours (KNN)

KNN is non-generalizing learning. Instead of constructing a general internal model, it stores all instances corresponding to training data in n-dimensional space. Here, the two parameters considered are the value of K which is a parameter that refers to several nearest neighbours (in our case we used 10 neighbours), and the distance function whereby the distance between the new point and each training point is calculated, then the closest points are picked. There are various methods of calculating distance, the common ones are three; Euclidean distance, Manhattan distance, and Hamming distance. Euclidean distance (D_H) is the squire root of the sum of the squired of the distance differences between a new point (x) and an existing point (y)

$$D_E = \sqrt{\sum_{i=1}^{k}(x_i - y_i)^2} \quad \text{..... (4.a)}$$

Manhattan distance is the distance between real vectors using the sum of their absolute difference (of a new point (x) and an existing point (y)

$$D_m = \sum_{i=1}^{k}|x_i - y_i| \quad \text{..... (4.b)}$$

Hamming distance is used for categorical variables. If the value of the new point (x) and the value of the existing point (y) is the same, then the distance D = 0, otherwise D = 1

$$D_m = \sum_{i=1}^{k}|x_i - y_i| \quad \text{..... (4.c)}$$

$$x = y \Rightarrow D = 0$$

$$x \neq y \Rightarrow D = 0$$

Where K is defined as some points to be considered.

This model was used since it does not require a training period and so makes it a faster algorithm, unlike other regression models. With this model, the training dataset is stored and used during real-time prediction. Since our data is huge, this becomes one of the good prediction models. KNN uses data and classifies new data points based on similarity measures (Sarker, 2021).

3.5.2.3 Least Absolute Shrinkage and Selection Operator (LASSO)

LASSO is the regression analysis technique that reduces the absolute value of the regression coefficients (Kiang, 2018). The objective function that is minimized by the LASSO algorithm can be expressed as;

$$\min_{w} \frac{1}{2n}\|Xw - Y\|_2^2 + \alpha\|w\|_1 \quad \text{.....(6)}$$

where w is the coefficient vector, which contains coefficients associated with model parameters; X is the feature vector; Y is the target vector; n is the number of depth samples in the training dataset; and the hyperparameter is the penalty parameter that balances the importance of the sum of squared errors term and the regularization term, which is the norm of the coefficient vector.

This model is chosen because there is difference in coefficients of predictor variables in our data set. The main advantage of this model on this data set is that it learns the linear relationship and shrinks the regression coefficient towards zero by penalizing the regression model using regularization terms together to ensure the sparsity of the coefficients (Misra et al., 2020).

3.5.2.4 Extreme Gradient Boosting tree (XGBoost)

The boosted tree is the ensemble method that constructs more than one decision tree. It is an additive regression model in which individual terms are simple trees. Boosting combines classifiers made from weighted versions of the learning sample with weights that are adaptively altered at each step to provide more weight to cases that were misclassified in the previous stage (Sutton, 2005). In this study, XGboosting was applied. XGboosting sequentially adds predictors to the ensemble and follows the sequence in correcting preceding predictors to arrive at an accurate predictor.

This model is chosen because it combines the strengths of two algorithms: regression trees (models that relate a response to their predictors by recursive binary splits) and boosting (an adaptive method for combining many simple models to give improved predictive performance) (Elith, Leathwick, and Hastie, 2008).

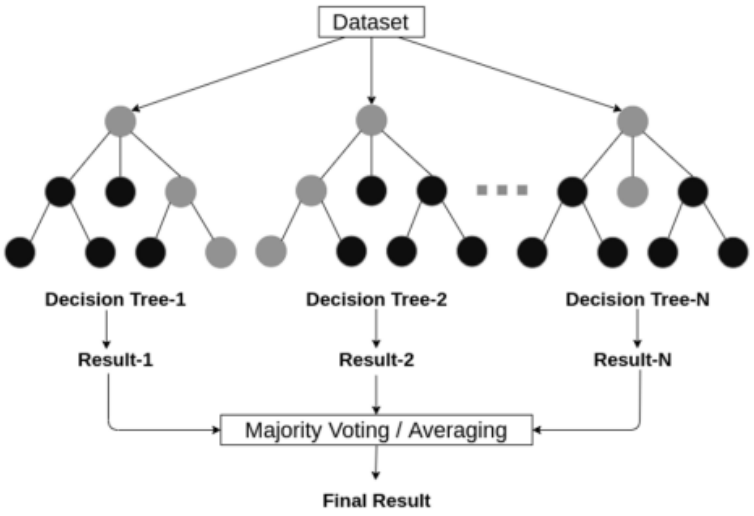
3.5.2.5 Random Forest Regression (RFR)

RFR is another ensemble method that constructs more than one decision. It is a tree-based algorithm whose trees (that are independently trained) are assembled by bagging (Figure 3). According to Hanafy and Mahmoud (2021), the following is an example of a random model for forest regressors:

$$g(x) = f_0(x) + f_1(x) + f_2(x) + \ldots\ldots + f_n(x) \quad (5)$$

where g is the final model, which is the sum of all models $f(x)$ is the decision tree

Figure 3: An example of random forest structure in consideration of multiple decision tree



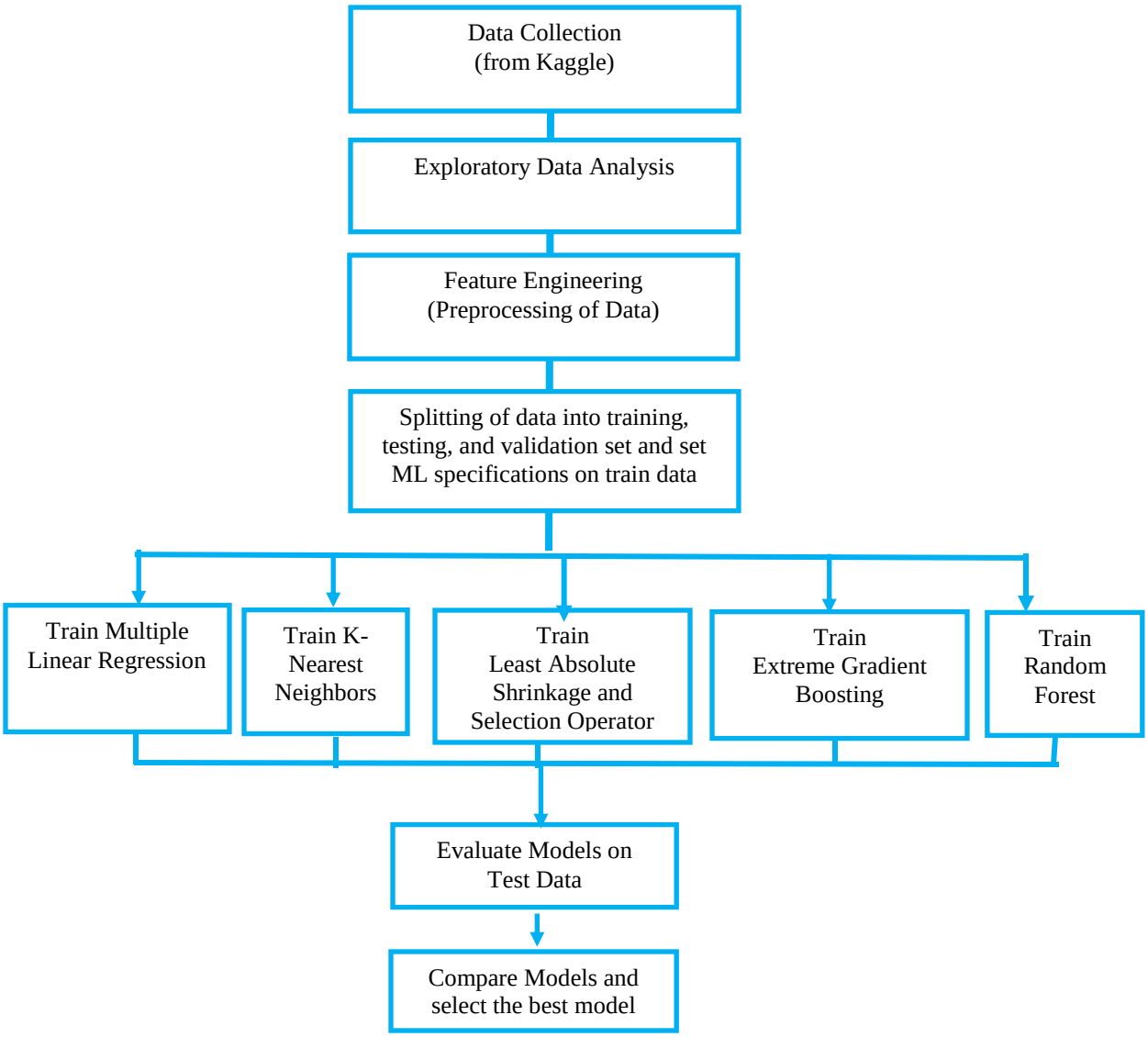
Source: Sarker (2021)

RFR model was chosen for the given dataset because it combines many decision trees to predict a more accurate outcome. Each tree is used to generate a prediction for a new random sample, then the predictions are averaged to form the forest's prediction (Yego et al., 2020). This model reduces the problem of over fitting on our dataset.

3.5.3 Predictive modelling phases

The modelling started with data collection, followed by exploratory data analysis. The last stage was models' comparison. Figure 4, illustrated the stages that were carried out in this study.

Figure 4: Predictive modelling



3.6 Estimation of the accuracy of the prediction

The estimation of accuracy of the prediction of the ML models was evaluated by R squared (R^2), Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE).

3.6.1 R-squared

R^2 is the coefficient of decision. The value of R^2 is between 0 and 1. When R^2 value is higher, the better the model output. This indicates that the model has drifted from real-world values less. The best possible value of R^2 is 1.0 and it is given by

$$R^2 = \frac{Explained\ variance}{Total\ variance} \quad \text{..... (6.a)}$$

3.6.2 Root Mean Squared Error (RMSE)

The square root is used to calculate the RMSE of the difference between expected and real values. The lesser the root mean square error, the better (means are less variance among the expected values and the real values)

$$RMSE = \frac{1}{N}\sum_{n=1}^{N}(\hat{Y} - Y)^2 \quad \text{..... (6.b)}$$

Where N denotes the total number of observations, E denotes the predicted premium value, and Y denotes the actual insurance premium value.

3.6.3 Mean Absolute Error (MAE)

The MAE is the difference between the original and forecast values, which is calculated by averaging the absolute difference over the whole data set. The MAE should be as low as possible.

$$MAE = \frac{1}{N}\sum_{n=1}^{N}\left|\hat{Y} - Y\right| \quad \text{..... (6. c)}$$

CHAPTER FOUR: DATA ANALYSIS

4.1 Introduction

This chapter summarizes the study's findings. The first part of the chapter presents the results of the exploratory data analysis. In this subsection, several graphs are presented which show correlation matrix, univariate and bivariate analysis as well as the relationship between independent variables and insurance premium. The second part of the chapter presents the results of predictive analysis.

4.2 Exploratory Data Analysis

4.2.1 Descriptive Statistics

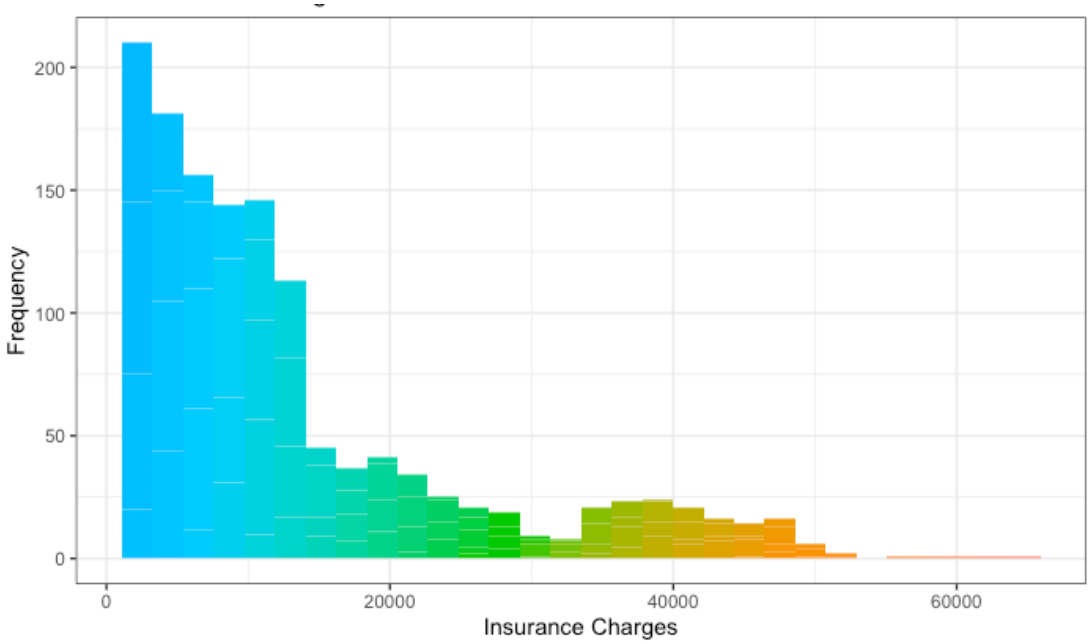
The number of male and female respondents was nearly the same, and more than half of them (79.5%) were non-smokers. The average age and BMI were 39 years and 30.67 respectively. The average medical charge was $ 13,270 (Table 3).

Table 3: Descriptive statistics for categorical data - Sex, Smoking, and Location variables

Categorical data		
Variable	**Frequency**	**Percent (%)**
Sex		
Female	662	49.5
Male	676	50.5
Smoking		
Smoker	274	20.5
Non-smoker	1064	79.5
Location		
Northeast	324	24.2
Northwest	325	24.3
Southeast	364	27.2
Southwest	325	24.3
Continuous data		
Variable	**Median**	**Mean**
Age	39	39
BMI	30.4	30.67
Number of children	1	1

The charges vary greatly between $1,000 to $64,000. Many respondents were charged less than $2,000 by health insurance companies (figure 5).

Figure 5: Distribution of insurance charges



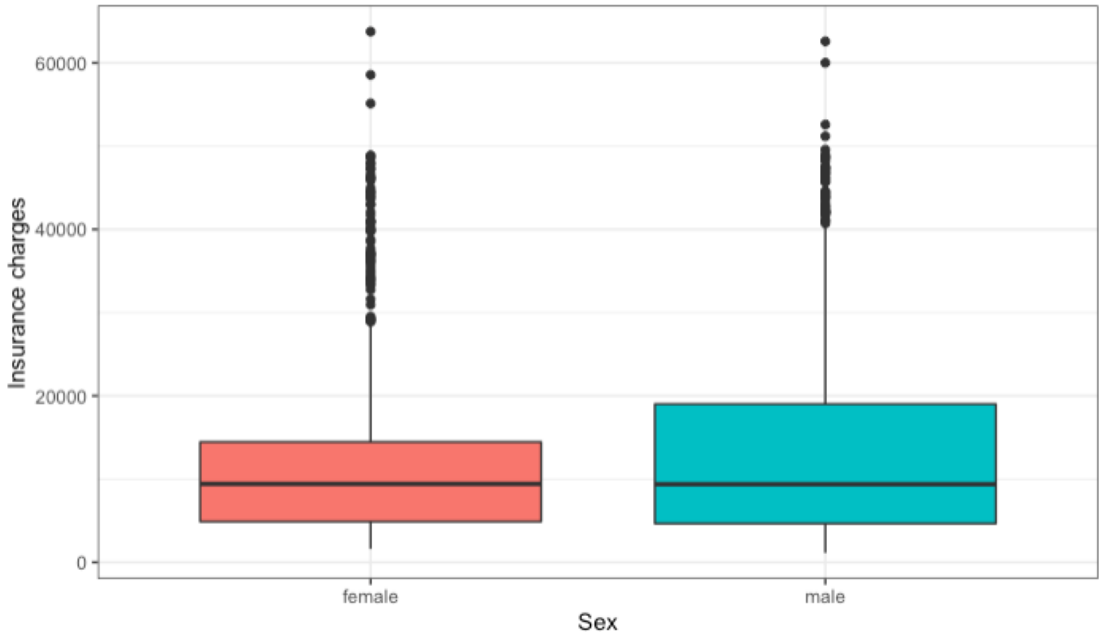
4.3 Multicollinearity test

The Pearson t-test was conducted to find out if the correlation among the predictors (independent variables) was significant. The test results showed no significant correlation among the independent variables. Therefore, it was concluded that the issue of multicollinearity in the dataset did not exist.

4.2.3 Relationship between insurance charges and predictor variables

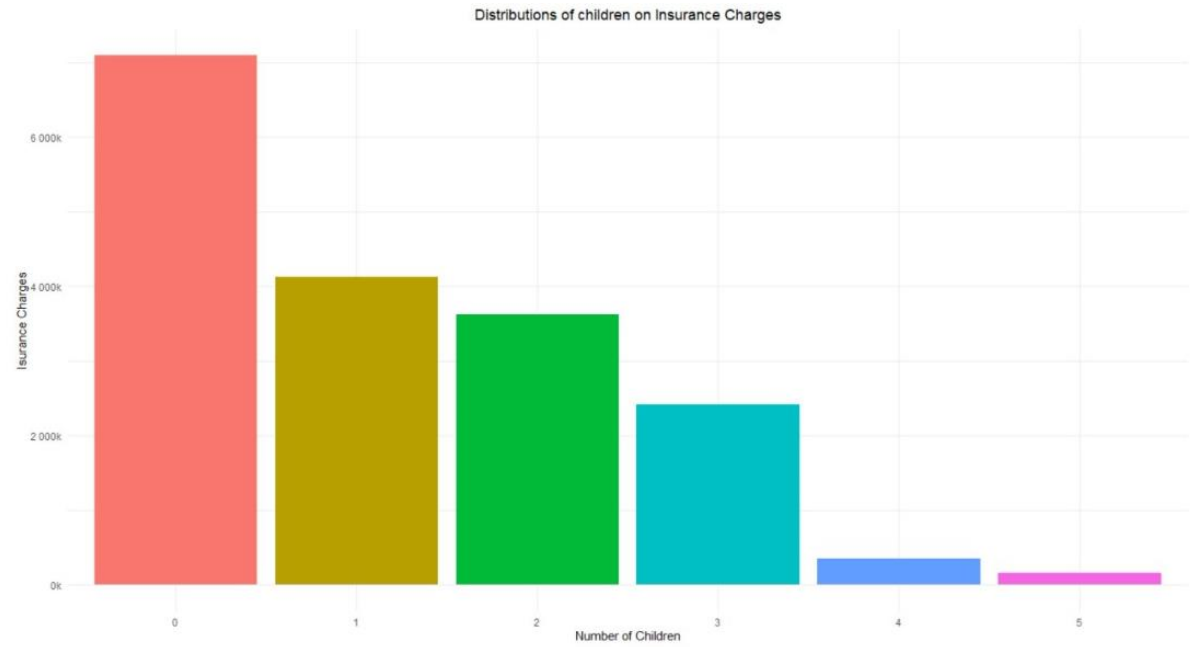
The distribution of insurance charges on sex shows that there is no difference between male and female. The distribution is nearly the same for both sex categories (figure 6).

Figure 6: Boxplot of insurance charges per sex



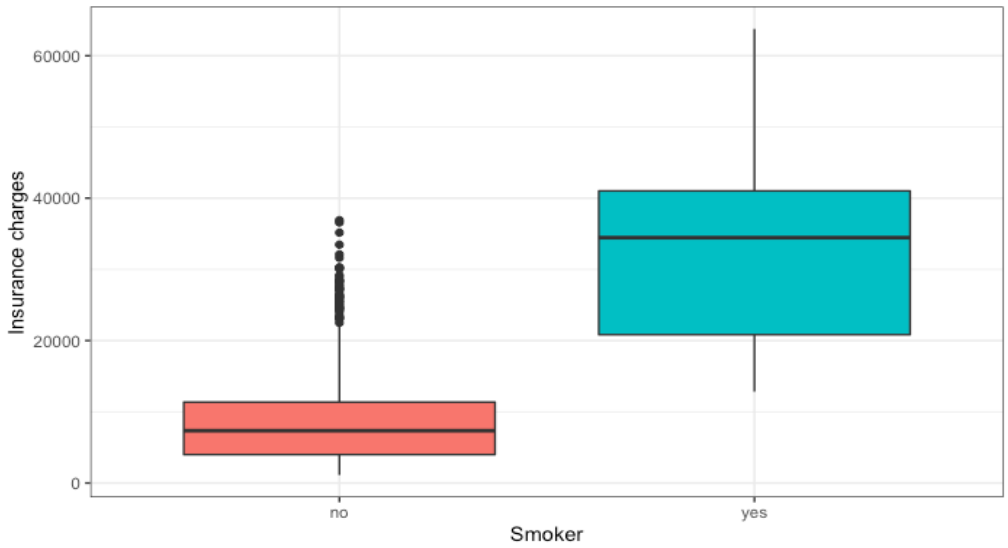
The number of children has a small effect on the insurance charges. The distribution of charges on the number of children shows that most patients that did not have children were charged less than those with children as shown on the diagram below (Figure 7).

Figure 7: Histogram showing the distribution of insurance charges per child



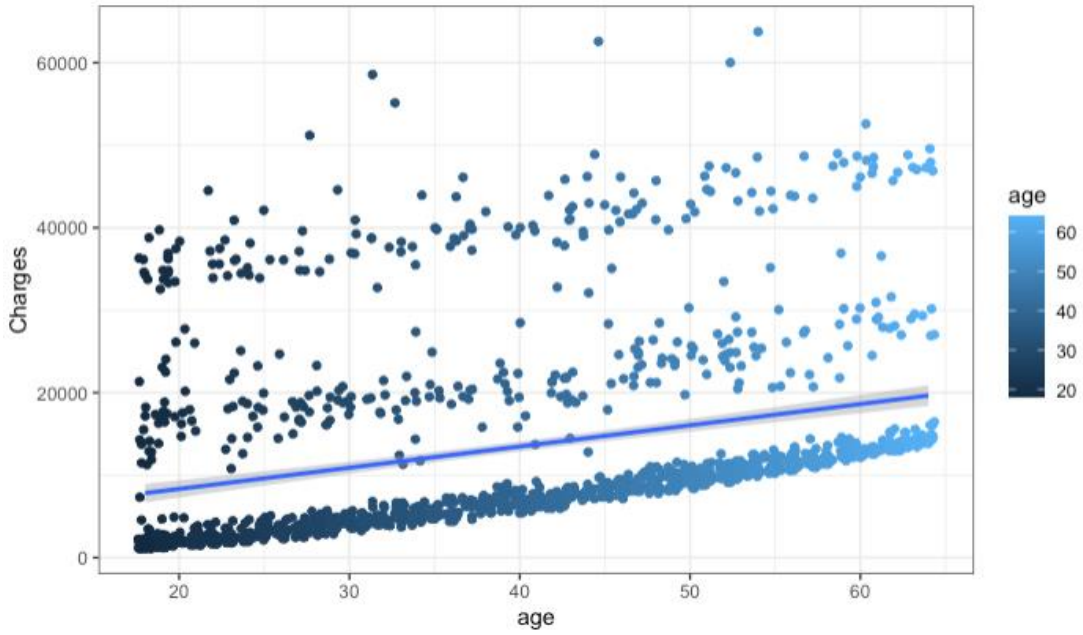
Regarding charges and smoking, the boxplot (figure 8) indicates that there is a high increase in insurance charges for people who smoke compared with people who do not smoke.

Figure 8: Boxplot of insurance charges per smoking



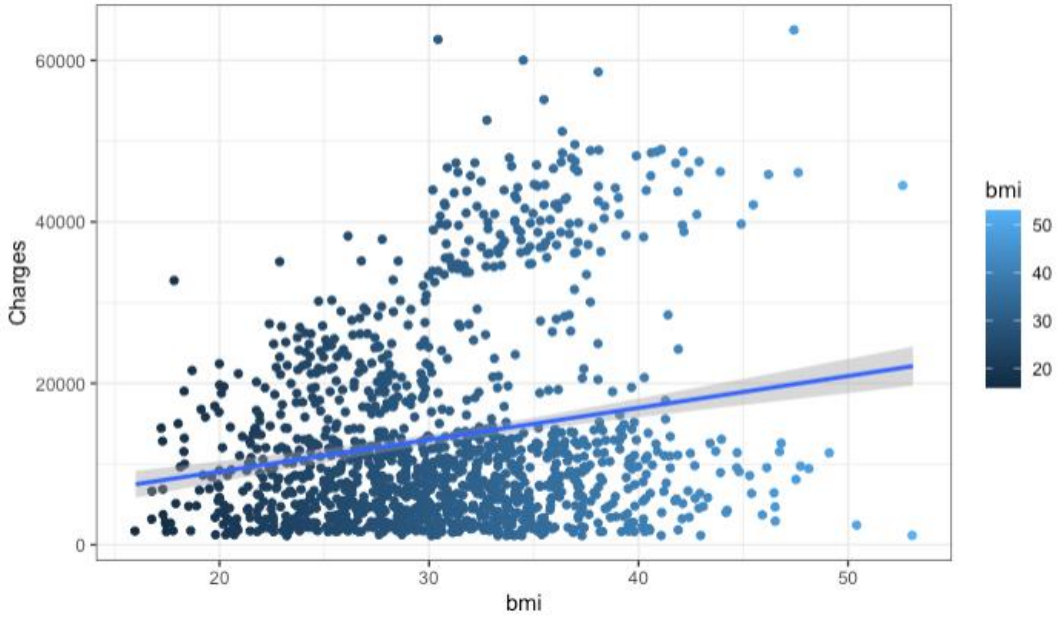
The plot of age on insurance charges shows that insurance charges increase as age increases (figure 9)

Figure 9: Relationship between insurance charges and age



The plot of BMI on insurance charges shows insurance charges increases as BMI increases (figure 10). This means people with high BMI are charged more by the insurance company than people with low BMI.

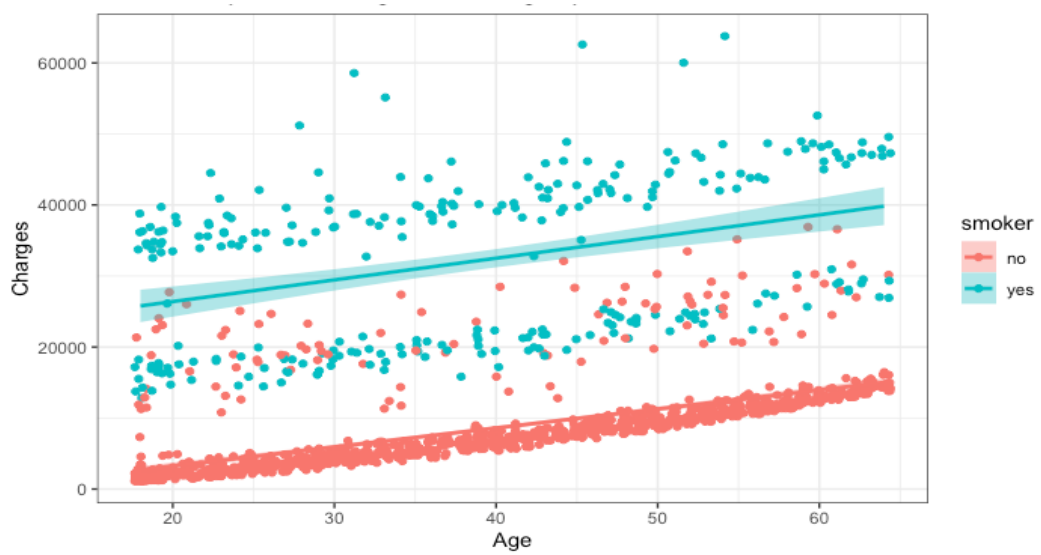
Figure 10: Relationship between insurance charges and BMI



4.2.4 Combined influence of smoking and other predictors on insurance charges

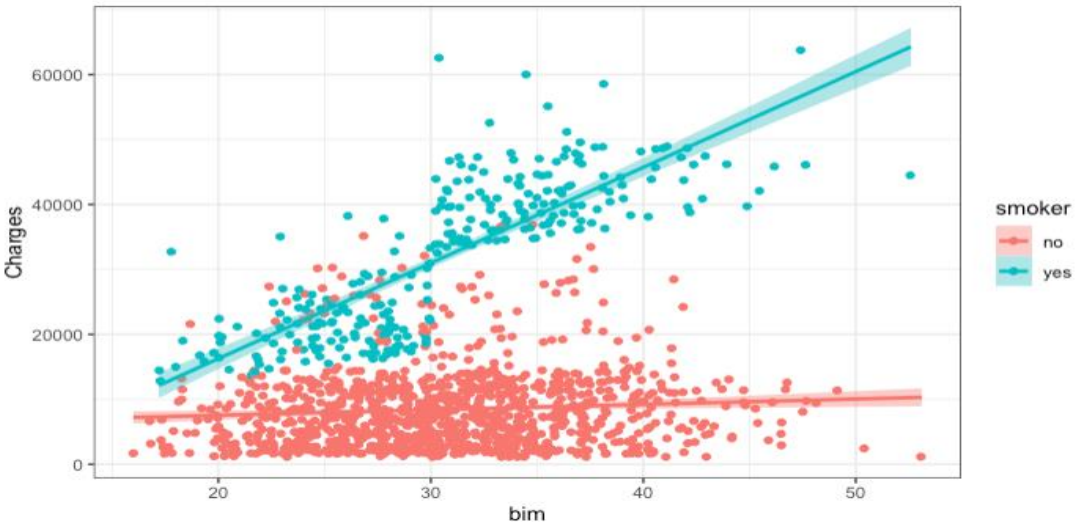
Insurance charges were plotted against combined smoking and other variables, to visualize the combined effect of smoking and other predictors on insurance charges. The distribution of age and insurance charges for smokers and non-smokers shows insurance charges increase as age increases for both smokers and non-smokers. Insurance charges are relatively higher on old smokers than young smokers (Figure 11)

Figure 11: Distribution Age and insurance charges for smokers and non-smokers



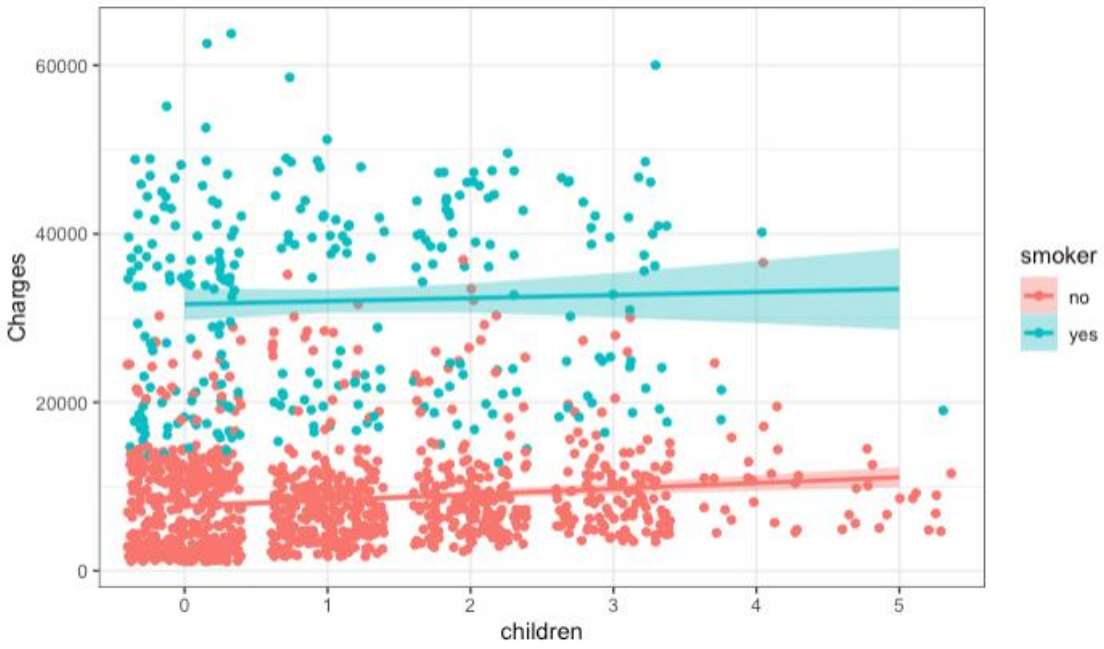
A plot distribution of BMI and insurance charges for smokers and non-smokers shows although there is an increase of insurance charges as BMI increases, smokers have a sharp increase of charges as their BMI increases compared with non-smokers (Figure 12)

Figure 12: Distribution of BMI and insurance charges for smokers and non-smokers



The relationship between number of children and insurance charges for both smokers and non-smokers show that charges among smokers are higher than charges among non-smokers (Figure 13).

Figure 13: Relationship between children and insurance charges for smokers and non-smokers

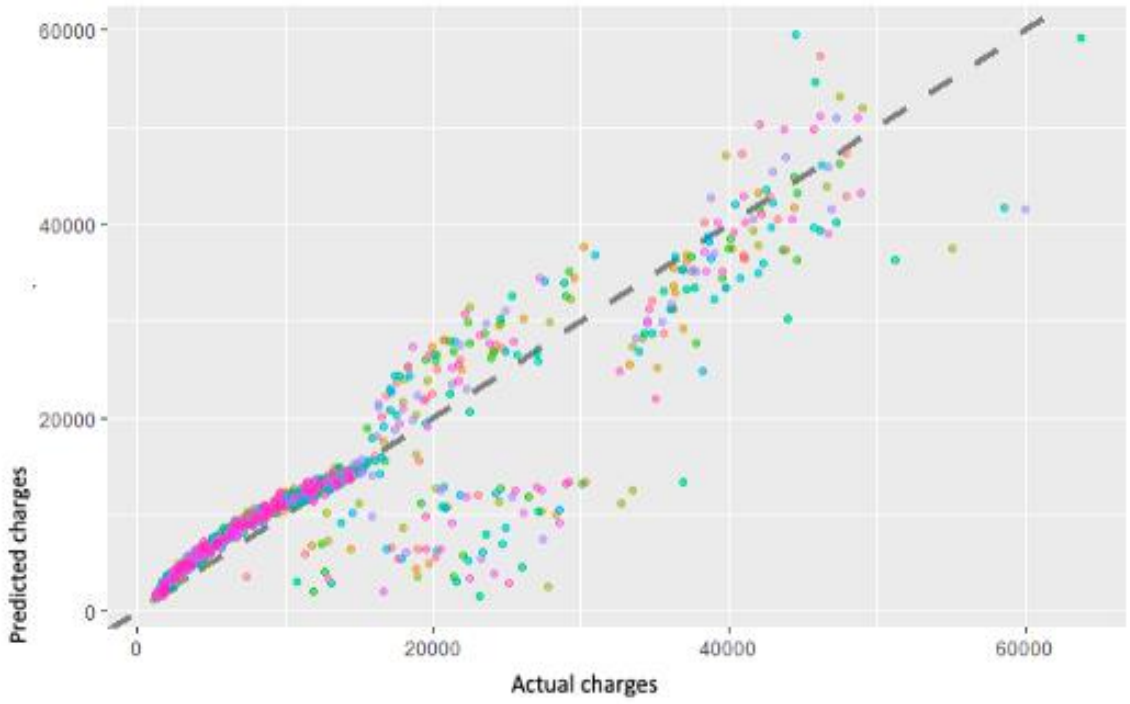


4.3 Predictive Modelling

4.5.1 Multiple liner regression analysis

The most common machine learning regression model used is the MLR. The plot of actual versus predicted values for MLR illustrates how well the model fits the data (Figure 14). The model did well in estimating the value for smaller charges, as can be observed. However, the prediction performance of the values for higher charges was not well performed.

Figure 14: Plot of Actual vs Predicted values for MLR



The coefficients obtained from MLR are tabulated in table 4, and a p-value of 0.05 is used as a cut-off point to determine the significance of the variables. The bigger value coefficients represent higher relevance to the model. Based on the results, the following values had a statistically significant influence on the insurance prices; Age, BMI, Smoking and Region. The older people pay higher premiums for health coverage than young ones. Also, people with high BMI pay higher proportional premiums for health insurance compared to people with low BMI. Regarding Smoking, the insurance companies charge smokers higher premiums for health insurance compared to non-smokers. Regarding region, individuals living in the Northeast have higher insurance charges when compared with individuals living in the Southeast region (reference variable) (Table 4)

Table 4: Linear Regression Variables' coefficients

	estimate	Std. Error	t- value	Pr (>I t I)
(intercept)	3.86147262	0.02992126	129.0544873	0.000***
Age	0.21956219	0.01298423	16.9099069	0.000***
Sex_male	-0.03380324	0.02489352	-1.3579133	0.175
BMI	0.03218319	0.01304476	2.4671357	0.001**
Smoker_yes	0.70958159	0.03130212	22.6688045	0.000***
Children	0.47568900	0.01378000	3.4520000	0.100
Northeast	0.03462494	0.03554844	0.9740211	0.033*
Southwest	-0.03571256	0.03624123	-0.9854124	0.325
Southeast	-0.07255654	0.03629282	-1.9991982	0.046

Sign. Codes: 0 '***' 0.001 '**' 0.01 '*'. 0.05 '-' 0.1 ' ' 1

Based on the estimates and significant values presented in table 4, the MLR predictive model (equation 2) can be presented as follows

$$\text{Insurance charges} = \beta_0 + \beta_1 \text{Age} + \beta_3 \text{BMI} + \beta_4 \text{smoker} + \beta_5 \text{ Northeast (location)}$$

$$\text{Insurance charges} = \beta_0 + 0.219\ \text{Age} + 0.032\ \text{BMI} + 0.709\ \text{Smoker} + 0.033\ \text{Northeast}$$

4.5 Evaluation of the Performance

To test the effectiveness of the five (5) machine learning algorithms, the values of MAE, RMSE, and R^2 for each of the models were compared. MLR has much lower performance on test data and leads to overfitting on this data and would not be preferred for this problem. The findings are tabulated in Tables 5.

Table 5: Model's comparison

Model	R^2	MAE	RMSE
Evaluation metrics for train data			
K-nearest Neighbours (KNN)	0.8754706	2691.375	4299.594
Least Absolute Shrinkage and Selection Operator (LASSO)	0.7722119	4136.346	5733.035
Extreme Gradient Boosting (XGboosting)	0.8901612	2224.492	3990.497
Random Forest Regression (RFR)	0.8978284	2196.060	3847.263

Evaluation metrics for test data			
K-nearest Neighbours (KNN)	0.8252715	3221.427	5236.708
Least Absolute Shrinkage and Selection Operator (LASSO)	0.7397349	4516.719	6384.349
Extreme Gradient Boosting (XGboosting)	0.8553049	2688.200	4748.673
Random Forest Regression (RFR)	0.8530681	2726.356	4783.827

Based on test data, by looking at the value of R^2, the XGboosting model was able to explain 85.5% of the variation, followed by RFR which explained 85.3%. The RMSE estimate for Extreme Gradient Boosting is 4748.673, which is much better than other models. Based on the findings above, the comparison graph was developed. Figure 15 and 16 presents a comparison of the four (4) models, on three (3) performance measures used in this study (R^2, MAE, RMSE).

Figure 15: Comparison of the four (4) models on R^2 performance measure

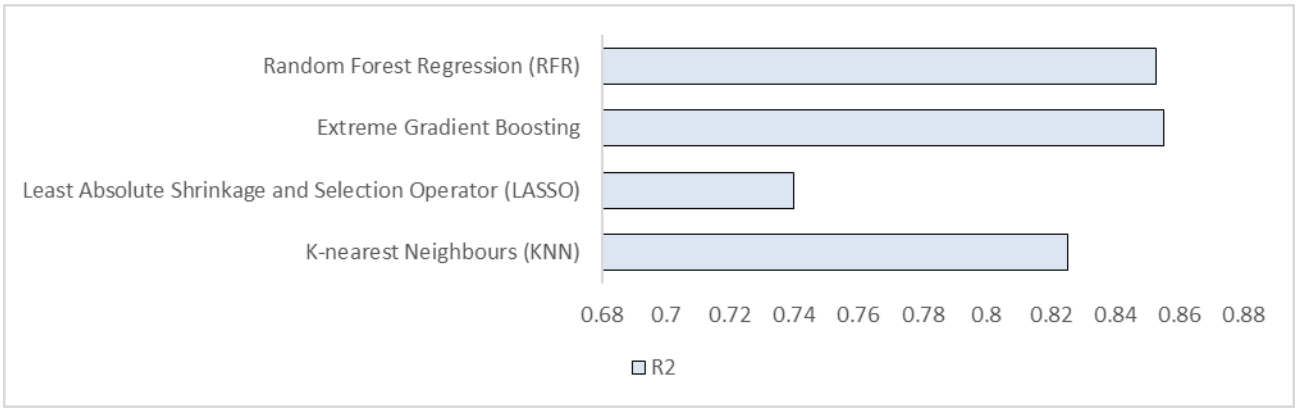


Figure 16: Comparison of the four (4) models on two performance measures (RMSE and MAE)

CHAPTER FIVE: DISCUSSION OF FINDINGS

5.2 Discussion

The number of male and female respondents was nearly the same and the distribution of the data based on the location of the respondents was also nearly the same. More than half of the respondents were non-smokers. The correlation test showed that all numeric predictor variables are less correlated to the response variables, meaning that the variables were suitable for the prediction of the insurance charges.

The plot distributions revelled the following variables have relationship with medical insurance charges; Age, BMI, number of children and smoking habit.. Sex is less likely to influence the charges, as the premiums changes had similar distribution for both males and females. Strong habit has strong influence on charges especially when combined with other variables. For example, the study revealed that old people who were also smokers are charged higher compared with old people who are non-smokers. Moreover, smokers who had higher BMI are also charged higher compared with smokers with low BMI.

The MLR was used to analyze the determinants of health insurance charges among the health insurance beneficiaries. The MLR results show that older individuals pay more than younger ones. The significant influence of age on medical insurance charges was also found in previous studies Similar findings were also obtained from the studies by Adebayo et al (2015), Kaur (2018), Kodiyan & Francis (2020) and Lakshmanarao et al (2020). Moreover, MLR findings indicate that the insurance company charges higher the smokers than non-smokers. Smoking was found to be one of the factors that strongly influence health insurance charges by other studies including study by Kaur (2018) and a study by Kodiyan & Francis (2020). Other previous studies that have similar findings are Adebayo et al., (2015) and Lakshmanarao et al., (2020).

The MLR findings show that people with higher BMI are charged more by insurance companies compared with people with relatively lower BMI. Similarly, the study by Kodiyan & Francis (2020) and Lakshmanarao et al (2020) also found a significant influence of BIM on medical insurance charges. Location is another factor that has a significant influence on medical charges insurance.

The medical insurance company tend to charge differently based on the beneficiary's residential area. These findings are contrary to the findings by Kaur (2018), Kodiyan & Francis (2020) that found Region/location has no significant influence on medical charges.

The relationship between the sex of the patient and medical insurance charges was also analysed using MLR. The MLR findings indicated that there is no or little influence on sex on medical insurance charges. Similarly, Kodiyan and Francis (2020) also found no significant influence of sex on medical charges. Contrary to these findings, Hanafy and Mahmoud (2021) found a significant influence of sex on medical insurance charges. Number of children was found to have no significant influence on the medical charges of the beneficiary.

To evaluate the performance of predictive models that use machine learning algorithms to predict health insurance premiums, the performance of five (5) machine learning regression models was evaluated. Those models are MLR, XGBoost, KNN, LASSO, RFR. The MLR had lower performance on test data and lead to overfitting based on the given data, for that reason, this model was not included in the comparison. The performance of the two models was better than all other models used. These models are XGBoost and RFR. The XGBoost model was able to explain high percentage of the variation (R^2), the highest among all other models followed RFR. Moreover, the root means square error estimate for XGBoost was much better than that of other models. These findings indicate that XGBoost and RFR have high ability to appropriately capture linear and non-linear relationships between the dependent and independent variables (as compared to other models evaluated). These findings are supported by other studies that found XGBoost and RFR as the best predictive models when compared with other models (Lakshmanarao et al., 2020; Yego et al., 2020). Generally, these results give us more reason to use the XGBoost and RFR models in the prediction of health premium rates.

CHAPTER SIX: CONCLUSION AND RECOMMENDATIONS

6.1 Conclusions

The dataset from *Kaggle* used included the data collected from individuals who were charged insurance premiums by the insurance company based on their demographic characteristics as well as behavioral characteristics. These data were then used to analyze the determinants of health insurance premiums as well as assess the performance of the machine learning regression models in the data set.

Demographic characteristics and behaviour characteristics are the main factors that influence the premium charges among the patients. The analysis of the main determinants of health insurance charges among the health insurance beneficiaries found that Age, BMI, Smoking, and region play significant roles in predicting the insurance changes. Thus, the above mentioned- demographic characteristics are also useful in developing a model to predict health insurance premium using demographic and behavior data. This means the health insurance companies can use these variables to develop a predictive model that might insure fair premium charges to their members. Regarding the evaluation of the performance of the machine learning regression models; XGBoost and RFR models have better performance than KNN and LASSO models. This means XGBoost and RFR models can appropriately capture linear and non-linear relationships between the dependent and independent variables. It is expected that premiums predictions based on these models would give realistic payment models for to insurance companies. Since the insurance rates that are affordable and give quality services encourage many people to purchase health insurance products (Douven et al., 2020), usage of these models have a potential to expand UHC.

6.2 Recommendations

The XGboosting and RFR are recommended as the best model for predicting health insurance premiums. The insurance companies that seek to develop a model for prediction premiums are recommended to use either XGBoost or RFR. To get a more precise prediction of premium charges, it is recommended to use datasets that are composed of data collected in several years.

Moreover, using a large dataset, larger than what was used in this study is more likely to increase the accuracy of the model.

Lastly, in order to provide more insights into the determinants of health insurance premiums, it is recommended that future studies should use more attributes in the predictive models in addition to age, sex, BIM, number of children, smoking, and region. More attributes will help to develop a comprehensive predictive model that can be applied in many contexts.

6.3 Study limitations

The first limitation is the lack of primary data from the Tanzania health insurance sub-sector. The study planned to use primary data from insurance companies in Tanzania. However, due to an outbreak of the COVID 19 pandemic, Collection of primary data from Tanzanian's insurance companies was not possible. The researcher had to use secondary data. This has limited the researcher an opportunity to learn the factors that influence health insurance premium charges in the Tanzania context. Despite of these fact, it is believed that the findings obtained from this study will contribute to the efforts to improve the health insurance sub-sector in Tanzania and other countries.

Another limitation is the existence of few variables in the dataset. The data set used had only six variables; age, sex, BIM, number of children, smoking, and region. Previous studies found there are additional attributes that have a significant influence on health insurance charges such as wealth quantile, education level (Yego, Kasozi and Nkrunziza, 2020), clinical information and disabled status (Yang et al, 2018), occupation, type of plan used (Strawiński and Celińska-Kopczyńska, 2019) and past medical history (Adebayo et al, 2015). Future studies can use these attributes to further evaluate the performance of the ML models.

BIBLIOGRAPHY

Adebayo, E. F., Uthman, O. A., Wiysonge, C. S., Stern, E. A., Lamont, K. T., & Ataguba, J. E. (2015). A systematic review of factors that affect uptake of community-based health insurance in low-income and middle-income countries. In *BMC Health Services Research* (Vol. 15, Issue 1, p. 543). BioMed Central Ltd. https://doi.org/10.1186/s12913-015-1179-3

Allen, M. P. (2007). Chapter 37 - The problem of multicollinearity. In *Understanding Regression Analysis* (pp. 176–180). Springer. https://doi.org/10.1007/978-0-585-25657-3_37

Bernard, C., Cui, Z., & Vanduffel, S. (2017). Impact of Flexible Periodic Premiums on Variable Annuity Guarantees. *North American Actuarial Journal*, *21*(1), 63–86. https://doi.org/10.1080/10920277.2016.1209119

Bhadja, N. D., & Abhangi, P. A. A. (2018). A review Of Machine Learning Methodology in Big data. *International Journal of Scientific Development and Research - IJSDR*, *3*(5), 361–368.

Buhlmann, H. (1984). *P r e m i u m c a l c u l a t i o n from top d o w n by*. *2*.

Bzdok, D., Altman, N., & Krzywinski, M. (2018). Points of Significance: Statistics versus machine learning. *Nature Methods*, *15*(4), 233–234. https://doi.org/10.1038/nmeth.4642

Cohen, S. (2021). The basics of machine learning: strategies and techniques. In *Artificial Intelligence and Deep Learning in Pathology* (pp. 13–40). Elsevier Inc. https://doi.org/10.1016/b978-0-323-67538-3.00002-6

Conn, C., & Walford, V. (1998). An Introduction to Health Insurance for Low Income Countries. *DfID Health Systems Resource Centre, UK*.

Douven, R., van der Heijden, R., McGuire, T., & Schut, F. (2020). Premium levels and demand response in health insurance: relative thinking and zero-price effects. *Journal of Economic Behavior and Organization*, *180*, 903–923. https://doi.org/10.1016/j.jebo.2019.02.030

Edgar, T. W., & Manz, D. O. (2017). *Chapter 6 - Machine Learning* (T. W. Edgar & D. O. B. T.-R. M. for C. S. Manz (eds.); pp. 153–173). Syngress. https://doi.org/https://doi.org/10.1016/B978-0-12-805349-2.00006-6

Elith, J., Leathwick, J. R., & Hastie, T. (2008). *A working guide to boosted regression trees*. *Ml*, 802–813. https://doi.org/10.1111/j.1365-2656.2008.01390.x

Evans, D. B., Hsu, J., & Boerma, T. (2013). Universal health coverage and universal access. *Bulletin of the World Health Organization*, *91*(8), 10–11. https://doi.org/10.2471/BLT.13.125450

Greenlaw, S., & Shapiro, D. (2011). *Principles of Economics 2e*.

https://d3bxy9euw4e147.cloudfront.net/oscms-prodcms/media/documents/Economics2e-OP_s2jF42u.pdf

hanafy, M., & Mahmoud, O. M. A. (2021). Predict Health Insurance Cost by using Machine Learning and DNN Regression Models. *International Journal of Innovative Technology and Exploring Engineering*, *10*(2), 137–143. https://doi.org/10.35940/ijitee.c8364.0110321

Ho, A. (2015). Health Insurance. *Encyclopedia of Global Bioethics*. https://doi.org/10.1007/978-3-319-05544-2_222-1

Hong Wang, Kimberly Switlick, Christine Ortiz, C., & Connor, and B. Z. (2010). *Africa Health Insurance Hand Book: How to make it work. June*. www.healthsystems2020.org

Huber, M., André Knottnerus, J., Green, L., Van Der Horst, H., Jadad, A. R., Kromhout, D., Leonard, B., Lorig, K., Loureiro, M. I., Van Der Meer, J. W. M., Schnabel, P., Smith, R., Van Weel, C., & Smid, H. (2011). How should we define health? *BMJ (Online)*, *343*(7817). https://doi.org/10.1136/bmj.d4163

Jokerst, C. E., & Gotway, M. B. (2005). Thoracic Radiology: Noninvasive Diagnostic Imaging. In R. J. Mason, J. F. Murray, & J. A. Nadel (Eds.), *Murray and Nadel's Textbook of Respiratory Medicine* (7th ed., pp. 272–299). Elsevier Saunders. https://books.google.co.tz/books?id=DCUjuQAACAAJ

Kalali, A., Richerson, S., Ouzunova, E., Westphal, R., & Miller, B. (2019). Chapter 16 - Digital Biomarkers in Clinical Drug Development. In G. G. Nomikos & D. E. B. T.-H. of B. N. Feltner (Eds.), *Translational Medicine in CNS Drug Development* (Vol. 29, pp. 229–238). Elsevier. https://doi.org/https://doi.org/10.1016/B978-0-12-803161-2.00016-3

Kaur, T. (2018). *Factors affecting health insurance premiums : Explorative and predictive analysis Factors Affecting Health Insurance Premiums : Explorative and Predictive Analysis Creative Component Project Report By*.

Kiang, Y.-H. (2018). Chapter 2.- Model development and validation methodology: A classical big data application. In *Fuel Property Estimation and Combustion Process Characterization* (pp. 11–39). https://doi.org/10.1016/B978-0-12-813473-3.00002-7

Kibusi, S. M., Sunguya, B. F., Kimunai, E., & Hines, C. S. (2018). Health insurance is important in improving maternal health service utilization in Tanzania - Analysis of the 2011/2012 Tanzania HIV/AIDS and malaria indicator survey. *BMC Health Services Research*, *18*(1), 1–10. https://doi.org/10.1186/s12913-018-2924-1

Killada, P. (2017). *Data Analytics using Regression Models for Health Insurance Market place Data* (Vol. 11, Issue 1). University of Toledo.

Kodiyan, A. A., & Francis, K. (2020). *Linear regression model for predicting medical expenses based on insurance data*. *December 2019*. https://doi.org/10.13140/RG.2.2.32478.38722

Lakshmanarao, A., Koppireddy, C. S., & Kumar, G. V. (2020). Prediction of medical costs using regression algorithms. *Journal of Information and Computational Science*, *10*(5), 751–757.

Lantz, B. (2019). *Machine Learning with R: Expert techniques for predictive modeling, 3rd Edition*. Packt Publishing. https://books.google.co.tz/books?id=iNuSDwAAQBAJ

Lee, B., Tarimo, K., & Dutta, A. (2018). Tanzania's Improved Community Health Fund An Analysis of Scale-Up Plans and Design. *HP Policy Brief, October*.

Ley, C., Martin, R. K., Pareek, A., Groll, A., Seil, R., & Tischer, T. (2022). Machine learning and conventional statistics: making sense of the differences. *Knee Surgery, Sports Traumatology, Arthroscopy*, *30*(3), 753–757. https://doi.org/10.1007/s00167-022-06896-6

Manzi, F., Schellenberg, J. A., Hutton, G., Wyss, K., Mbuya, C., Shirima, K., Mshinda, H., Tanner, M., & Schellenberg, D. (2012). Human resources for health care delivery in Tanzania: A multifaceted problem. *Human Resources for Health*, *10*, 3. https://doi.org/10.1186/1478-4491-10-3

Misra, S., Li, H., & He, J. (2020). Chapter 5 - Robust geomechanical characterization by analyzing the performance of shallow-learning regression methods using unsupervised clustering methods. In *Machine Learning for Subsurface Characterization* (pp. 129–155). Elsevier Inc. https://doi.org/10.1016/B978-0-12-817736-5.00005-3

Mtei, G., Mulligan, J., Ally, M., Palmer, N., & Mills, A. (2007). An Assessment of the Health Financing System in Tanzania: Implication for Equity and Social Health Insurance. *Framework, May*.

Nurul, mas'ud waqiah. (2013). 済無No Title No Title. In *Persepsi Masyarakat Terhadap Perawatan Ortodontik Yang Dilakukan Oleh Pihak Non Profesional* (Vol. 53, Issue 9).

Outreville, J. F. (1998). Theory and Practice of Insurance. *Theory and Practice of Insurance, June 2016*. https://doi.org/10.1007/978-1-4615-6187-3

Rapaport, C. (2015). An Introduction to Health Insurance : What Should a Consumer Know ? *Congressional Research Service*, 7–5700.

Sarker, I. H. (2021). Machine Learning: Algorithms, Real-World Applications and Research Directions. *SN Computer Science*, *2*(3), 1–21. https://doi.org/10.1007/s42979-021-00592-x

Strawiński, P., & Celińska-Kopczyńska, D. (2019). Occupational injury risk wage premium. *Safety Science*, *118*, 337–344. https://doi.org/10.1016/j.ssci.2019.04.041

Sutton, C. D. (2005). Classification and Regression Trees, Bagging, and Boosting. In *Handbook of Statistics* (Vol. 24, Issue 04). Elsevier Masson SAS. https://doi.org/10.1016/S0169-7161(04)24011-1

Taylor, R. M. (2015). Approaches to Universal Health Coverage and Occupational Health and Safety for the Informal Workforce in Developing Countries. In *Approaches to Universal Health Coverage and Occupational Health and Safety for the Informal Workforce in Developing Countries*. https://doi.org/10.17226/21747

Tungu, M., Amani, P. J., Hurtig, A. K., Dennis Kiwara, A., Mwangu, M., Lindholm, L., & San Sebastiån, M. (2020). Does health insurance contribute to improved utilization of health care services for the elderly in rural Tanzania? A cross-sectional study. *Global Health Action*, *13*(1). https://doi.org/10.1080/16549716.2020.1841962

UN. (2010). *United Nations General Assembly: Resolution adopted by the General Assembly on 25 September 2015 - 70/1. Transforming our world: the 2030 Agenda for Sustainable Development* (Vol. 25, Issue 2). https://doi.org/10.1163/157180910X12665776638740

UN. (2015). *Take Action for the Sustainable Development Goals – United Nations Sustainable Development*. https://www.un.org/sustainabledevelopment/sustainable-development-goals/

UNDP. (2019). *Universal Health Coverage for Sustainable Development - Issue Brief*. https://www.undp.org/publications/universal-health-coverage-sustainable-development-issue-brief#modal-publication-download

Verma, V. K., & Verma, S. (2022). Machine learning applications in healthcare sector: An overview. *Materials Today: Proceedings*, *57*, 2144–2147. https://doi.org/https://doi.org/10.1016/j.matpr.2021.12.101

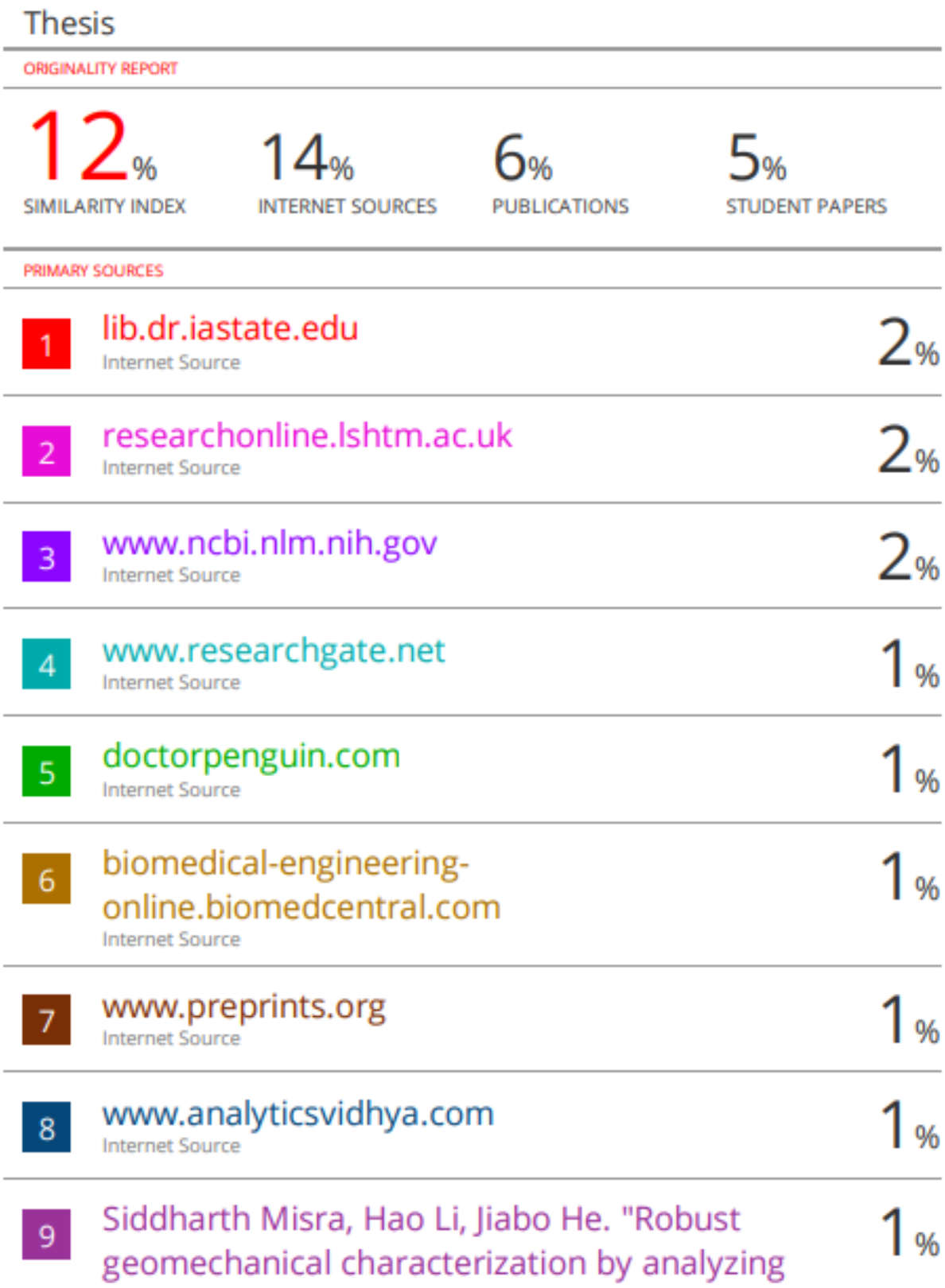
WB. (2021). *Universal Health Coverage*. Understanding Poverty. https://www.worldbank.org/en/topic/universalhealthcoverage

WHO. (2008). Consitution of the World Health Organization. In *The World Health Organization (WHO)* (Issue October). https://doi.org/10.4324/9780203029732

Yang, C., Delcher, C., Shenkman, E., & Ranka, S. (2018). Machine learning approaches for predicting high cost high need patient expenditures in health care. *BioMedical Engineering Online*, *17*(S1), 1–20. https://doi.org/10.1186/s12938-018-0568-3

Yego, N., Kasozi, J., & Nkrunziza, J. (2020). A Comparative Analysis of Machine Learning Models for Prediction of Insurance Uptake in Kenya. *MDPI*, *October*. https://doi.org/10.20944/preprints202010.0186.v1

APPENDIX



Thesis

ORIGINALITY REPORT

12%	14%	6%	5%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	lib.dr.iastate.edu Internet Source	2%
2	researchonline.lshtm.ac.uk Internet Source	2%
3	www.ncbi.nlm.nih.gov Internet Source	2%
4	www.researchgate.net Internet Source	1%
5	doctorpenguin.com Internet Source	1%
6	biomedical-engineering-online.biomedcentral.com Internet Source	1%
7	www.preprints.org Internet Source	1%
8	www.analyticsvidhya.com Internet Source	1%
9	Siddharth Misra, Hao Li, Jiabo He. "Robust geomechanical characterization by analyzing	1%

the performance of shallow-learning regression methods using unsupervised clustering methods", Elsevier BV, 2020
Publication

10	Submitted to Edison State College Student Paper	1 %

Exclude quotes	Off	Exclude matches	< 1%
Exclude bibliography	On		