AFRICAN CENTER OF EXCELLENCE IN DATA SCIENCE

COLLEGE OF BUSINESS & ECONOMICS

A PREDICTIVE MODEL OF DIARRHEA DISEASE AMONG UNDER-FIVE CHILDREN WITH MACHINE LEARNING ALGORITHMS: EVIDENCE FROM RWANDA DEMOGRAPHIC HEALTH SURVEY 2014-2015

By

UWAMAHORO Sandrine

Registration number: 220000879

A dissertation submitted in partial fulfilment of the requirements for the degree of Master of Data Science in Biostatistics

University of Rwanda, College of Business and Economics

Supervisor: Dr. Ignace H. KABANO

September, 2022

DECLARATION

I hereby declare that, except where indicated this dissertation entitled "A predictive model of diarrhea disease in under five children with machine learning algorithms: evidence from Rwanda Demographic Health Survey 2014-15' is entirely my work and has not been submitted for the award of degree in any other university.

STUDENT NAME: UWAMAHORO Sandrine

SIGNATURE: 

APPROVAL SHEET

This dissertation entitled ‘A predictive model of diarrhea disease in under five children: evidence from Rwanda Demographic Health Survey 2014-15’ written and submitted by UWAMAHORO Sandrine in partial fulfillment of the requirements of a degree for masters of science in Data Science majoring in Biostatistics is hereby accepted and approved. The rate of plagiarism using TURNITIN is 19% which is less than the rate accepted by ACE-DS.

Dr. Ignace H. KABANO

Supervisor

Dr. Ignace H. KABANO

Head of Trainings

DEDICATION

This dissertation is dedicated to my beloved parents, siblings and friends who have been on my side since day one.

ACKNOWLEDGMENT

Foremost I would like to thank the Almighty God; I wouldn't be here if it wasn't from him. My appreciation goes to the African Center of Excellence Data Science (ACEDS) for offering me this amazing opportunity of pursuing Master Degree and providing a platform for students to gain valuable skills.

My sincere gratitude goes to my supervisor Dr. Ignace H. KABANO for his continuous support and guidance throughout all the stages of writing the thesis. I would like to express my gratitude to my classmates for the great collaboration and hard work. Last but not least I would thank my parents and siblings for their encouragement, support and love.

ABSTRACT

Diarrhea disease is a worldwide burden since it is accounted as the second leading cause of death in children aged less than five and this is in line with the report of the Ministry of Health in Rwanda that identified childhood diarrhea as the second cause of morbidity in all health facilities in the period of June 2019 to June 2020. This research aimed to develop the best model to predict the occurrence of diarrhea disease among under-five children with machine learning techniques considering the socio-demographic variables and meteorological variables from RDHS 2014-2015.The target variable was dichotomous with class 0 of children with no diarrhea and class 1 representing children with diarrhea. Among all 7474 children considered in the study, only 905 (12%) experienced diarrhea episodes two weeks before the survey. Bivariate analysis has been performed where residence, age group, wealth index, type of toilet facility, main material floor, duration of breastfeeding, rotavirus vaccine and maternal education are associated with the childhood diarrhea status and the annual precipitation was found to be statistically significant. Six classifications algorithms including random forest, logistic regression, naïve Bayes, support vector machine, neural network, and gradient boosting were trained to find out the efficient model to predict diarrhea disease status among under-five children and Gradient boosting classifier was the best model with 86.3% of accuracy and this model identifies correctly 91.7% of children with diarrhea disease and can discriminate almost perfectly children with diarrhea and children without it. Feature importance test was performed to obtain relevant predictors that influenced the model to predict diarrhea disease status and high precipitation, children aged 12 to 24 months, household with earth and sand as main material floor, households with unimproved toilets, and children from poor households were identified as the most contributing predictors to predict diarrhea disease among children.

This model was valuable to identify accurately a vulnerable group of children at risk and it can be used at health facilities level and by community health workers to detect earlier the likelihood of diarrhea among children and set preventive measures to hinder diarrhea which could lead to severe diarrhea and dehydration, and this can lessen the morbidity and the number of hospital admissions due to diarrhea.

Keywords: Diarrhea, Under five children, machine learning, model, RDHS

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

MOH: Ministry of Health

WHO: World Health Organization

NISR: National Institute of Statistics of Rwanda

RDHS: Rwanda Demographic Health Survey

GAPPD: Integrated Global Plan of Action for the Prevention and Control of Pneumonia and Diarrhea

RPHC: Rwanda Population and Housing Census

EA: Enumerations Areas

HH: Household

ROC: Receiver Operating Characteristic

AUC: Area Under Curve

CHAPTER ONE: INTRODUCTION

1.1 Background

Diarrhea is defined as the passing of at least three loose or watery stools per day (or passing more frequently than normal for the individual). It is usually a symptom of an infection in the intestinal tract, which is originated from a variety of bacterial, viral, and parasitic organisms. The infection can be spread through contaminated food or drinkable water, or from person to person because of poor hygiene.

The World Health Organization stated three types of clinical diarrheas such as acute watery diarrhea, acute bloody diarrhea known as dysentery, and persistent diarrhea. Acute watery diarrhea can be referred to diarrhea that lasts not more than fourteen days. Nevertheless, dysentery and persistent diarrheas are types of diarrheas that can last more than fourteen days. (Brian A Maponga, 2013).

Rotavirus is said to be the primary cause of acute diarrhea mainly for children under five. It is a viral infection in the intestines that leads to symptoms such as diarrhea and being sick or vomiting (Starr, 2017). There are also other common causes of childhood diarrhea such as Campylobacter spp. and Salmonella spp, Shigella spp, and Campylobacter jejuni and they differ according to the geographical locations. (Lei Tian, 2016).

Globally, diarrhea disease is the second leading cause of death in children under five years old, it is responsible for killing around 525,000 children every year and it accounts for almost 1.7 billion cases every year among children (WHO, 2017). It accounts worldwide for nine percent of all deaths among under-five children and a high number of those deaths occur among children under two years living in South Asia and Sub-Saharan Africa. (Liliana Carvajal-Vélez, 2016).

(Ibrahim Khalil, 2017) has stated that repeated episodes of diarrheal lead to intestinal inflammation which can lead to malnutrition, long-term cognitive impairment, and increased vulnerability to opportunistic infections like pneumonia.

According to the Integrated Global Plan of Action for the Prevention and Control of Pneumonia and Diarrhea (GAPPD), there is a solid approach to ending pneumonia and diarrhea deaths by 2025 and it is comprised of both essential services and interventions to make healthy environment, it encourages practices that protect children from disease, and guarantees that

every child has access to established and suitable prevention and treatment measures, therefore the approach has purposed to reduce diarrhea mortality in children under 5 to less than 1 per 1,000 live births. (UNICEF, 2013).

In Africa, Diarrhea disease is the third leading cause of disease and death in children younger than 5 years of age and were accountable for 30 million cases of severe diarrhea and 330,000 deaths (95% credible interval, 270,000 to 380000) in 2015. (Boby Reiner, 2018).

According to Rwanda Demographic and Health Survey (RDHS) 2014-15, 12% of children under the age of 5 had diarrhea within the two weeks before the survey. (NISR, 2014-2015). However, the 2019-20 Demographic and Health Survey uncovered that 14% of children under the age of five had diarrhea in two weeks before the survey. (NISR M. o., 2020).

The National Institute for Statistics of Rwanda also reported that gastrointestinal disease was the ninth leading cause of death in 2014-2015 and the second leading cause of morbidity among children under five in health centers after acute respiratory infection in 2016 (Rwanda, 2018). Conforming to World Life Expectancy, deaths caused by diarrhea diseases in Rwanda reached 4036, or 7.11% of total deaths in 2018/

In Rwanda, data on diarrhea among children under five has been collected from all health facilities and community health workers situated in 30 districts for five years starting from January 1st, 2014, to December 31st, 2018, and 1,012,827 new diarrheal diseases episodes were observed in outpatient consultations by Community health workers during that period and per 100,000 population, the annual incidence rate was 12669.

However, there was a significant difference in rate from districts and years. The top incidence rate was noticed in the eastern province especially in Kirehe District in 2017 (329.3/1000) and the lowest incidence of 48.5/100 was observed in Kamonyi District. (Ladislas NSHIMIYIMANA, 2019). From June 2019 to June 2020, the ministry of health in Rwanda reported diarrhea as a second cause of morbidity in all health facilities among children under five. (Health, 2020).

1.2 Problem statement

Diarrhea disease in children under five years is a major threat yet considerable changes have been made in the prevention and treatment of the disease. It has been proved to be a leading cause of childhood mortality and morbidity. (Archana B Patel, 2011).

According to Rwanda's Fourth Health Sector Strategic Plan, the MoH targets to reduce the prevalence of diarrhea diseases to 9% by 2024. (health, 2018). However, 2019-20 RDHS highlighted an increase in the occurrence of diarrhea disease where 14% of children have experienced diarrhea in two weeks before the survey while it was 12% in 2014-15 RDHS (NISR M. o., 2020). It is evident that the occurrence of diarrhea disease is increasing, and this could hinder achieving the target of the ministry of health. There is a need to conduct deep analysis on the occurrence of diarrhea among under five children considering important variables and robust techniques.

The diarrheal disease remains a public concern even though preventive measures have been put in place. Although many studies on diarrheal disease have been conducted in Rwanda, they focused on classical methods to determine socio-demographic risk factors associated with diarrhea disease and the latter investigate only on the association based on hypothesis testing rather than extracting useful insights and accurate predictions as machine learning techniques do hence there is a gap of its application.

On the other hand, meteorological factors are recognized to influence the morbidity of diarrhea in different areas, (Nan-nan HUANG, 2021) uncovered that diarrhea disease is quietly associated with temperature, and (Ruixue Li, 2020) explained also that seasons are identified as one of the main factors influencing diarrhea among children.

Nowadays, the healthcare sector generates a large amount of data about patients and disease diagnostic, and when such data is well processed and analyzed with robust techniques it provides important knowledge that can be used efficiently in decision making, healthcare management, pharmaceutical firms, disease detection and diagnosis.

Therefore, this raised the interest of the researcher thus this study proposed to consider both socio-demographics and meteorological factors and the application of machine learning in

developing a robust predictive model of diarrheal disease in under-five children and that will enable decision-makers to set measures accurately and effectively to improve the quality of life of a child and the predictive model can be operationalized and be used to prevent earlier the disease. No similar method has been used to study diarrhea disease in Rwanda.

1.3 Relevance of the study

This study is important because it will provide a deep understanding of the factors that contribute to predicting diarrhea. It will enable health-related organizations to use the generated model to make informed decisions in preventing diarrheal diseases or significantly decrease its impact on children and establish control measures effectively.

The study will allow the government and other stakeholders to identify evidently where they can direct resources for better prevention with early intervention of proven efficacy. It will highlight the importance of a machine learning-based approach to predict diseases and this will help data scientists and other scientists for further research.

Furthermore, this research will help health facilities to detect children with diarrhea or not, provide treatment on time to patients, suggest early preventive measures and this will decrease unwanted complications such as dehydration and reduce the number of hospital admissions. Health workers will be aware of predominant factors that contribute to predicting diarrhea and this will help them to draw focus on vulnerable households.

1.4 Research objectives

The main objective of this study is to build a valid model that predicts diarrhea disease among under-five children in Rwanda.

The specific objectives are:

1. To identify risk factors associated with diarrhea disease among under-five children in Rwanda.

2. To develop different supervised learning models that predicts the occurrence diarrhea disease among under-five children in Rwanda.
3. To choose the best model that accurately predicts the occurrence of diarrhea disease among under five children in Rwanda.
4. To find out important features contributing to the best model for predicting diarrhea disease among under-five children in Rwanda.

1.5 Research questions

The following are the research questions that this study intends to respond to:

1. Which risk factors are associated significantly with diarrhea diseases among under five children?
2. What are the supervised learning models developed to predict diarrheal disease among under-five children?
3. What is the model that predicts accurately the occurrence of diarrhea disease among children under five?
4. What are the best important features that contribute to predicting the occurrence of diarrhea disease among under-five Children?

1.6 Definition of key terms

Diarrhea: is a passing of at least three loose or watery stools per day (or passing more frequently than normal for the individual).

Risk factor: is a characteristic, a behavior, or a condition that increases the chance of an event to happen.

Machine learning: is a branch of Artificial Intelligence that can be referred to as a study of computer programs that influence algorithms and statistical models to learn via inference and pattern regardless of being explicitly programmed and making informed decisions.

CHAPTER TWO: LITERATURE REVIEW

This section comprises a clear picture of diarrhea disease among children under five children, its causes, and symptoms, it highlights as well the risks factors associated with the diarrhea disease according to different researchers and the methodology used to evaluate them.

2.1 Definition and causes of diarrhea among children

Diarrhea is referred to the passing of the stools more frequently than for a normal individual per day, in other words, the passing is at least three stools per day. The etiology of diarrhea varies across regions and countries. However, few causes of diarrhea disease have been revealed such as Rotavirus, Norovirus, Adenovirus, Enteroaggregative E.Coli, Campylobacter spp. and Salmonella spp, Shigella spp, and Campylobacter jejuni.

A case-control study conducted by Aldo A. in six cities from Brasil indicated that six enteropathogenesis including norovirus, adenovirus, Enteroaggregative E. Coli, giardia, and STEC are associated with childhood diarrhea where Enteroaggregative E Coli is associated with high diarrhea severity. (Aldo A. M. Lima, 2019). In Sub-Saharan countries where Rwanda is located identified rotavirus as the main cause of diarrhea among children where it is accounted as a leading cause for morbidity and mortality of diarrhea among children less than 5 years. (Christopher Troeger, 2018).

2.2 Analysis of factors associated with diarrhea diseases in under-five children using classical methodologies

This section portrays researchers that have under-taken scientific studies on risk factors associated with the diarrheal disease among less than five years children and the prediction of the disease with logistic regression analysis.

(Sisay Shine, 2020) explained using bivariate analysis in a study conducted in Ethiopia that there was a significant association between diarrhea disease and birth order, age of the child, vaccination against rotavirus, age to starting complementary food, and feeding children by the

hand. With multiple logistic regression analysis, the researcher observed that children in the age group (7-11 months) are more likely to have diarrhea than their counterparts (48-59 months).

With the same methodology, (Getachew Yismaw Workie, 2019) revealed as well that children in the age group (6 to 23 months) are more likely to get diarrhea disease than children with less than 6 months. He highlighted as well that children living in rural area, children from a household with no latrine facility, and children whose households had no hand washing facility, children from a household with an unprotected drinking water source, and children from households with an openly dumped waste around the house are more likely to develop diarrhea disease.

Considering (Malachie Tuyizere, 2019), wealth index, age of the child, mother's education and household floor material have a significant association with diarrhea disease among under-five children. Logistic regression was performed and found that children in age group 6 to 11 months, children from poor families are at higher risk of diarrhea disease**.** (Sokhna Thiam, 2017) revealed that the predictor "children number exceeds two" is significantly associated with the occurrence of diarrhea.

In the North West of Ethiopia, (Thomas Sinmegn Mihrete, 2014) found through a logistic regression that child age and birth orders are statistically related to the diarrheal disease among children. He uncovered as well that mother's education, father's education; mother's occupation had a significant association with the morbidity of diarrhea among children.

Moreover, it showed that children from families with non-improved water sources are two times more likely to get diarrhea compared to children from families with an improved water source. The researcher showed that children from families with no toilet facility and children whose waste is not disposed of safely are more likely to have diarrhea.

(Ruixue Li, 2020) conducted a spatiotemporal analysis of diarrhea among under-five children taking into account the Nepal Demographic Health Survey, he found using Bayesian logistic regression that child age, child gender, mother's education year, and seasons are the main relevant factors influencing diarrhea among children under five. The risk of diarrhea decreases with age, the year of education of the mother and girls.

W.-P. SCHMIDT highlighted that diarrhea and other infectious diseases are known to have a strong dependence on the season and the researcher emphasized the fact that in the wet season, malaria and diarrhea increase (W.-P. SCHMIDT, 2009).

A study conducted in China revealed a significant association between meteorological factors and the incidence of infectious diarrhea since the disease varies according to seasons and periods. It was found that the incidence of infectious diarrhea is the highest in every autumn and winter. A distributed lag no linear model was performed in this study and factors such as daily maximum temperature, daily minimum temperature, daily average temperature, daily average air pressure, daily average relative humidity, and daily precipitation were considered where the maximum temperature was found to have the most significance lag effect on the incidence of infectious disease. This method is quite powerful to evaluate the non-linear relationship of meteorological factors and the lag effects. (Nan-nan HUANG, 2021).

2.3 Machine learning algorithms

Classification is one of the main tasks in machine learning and data mining and different researchers used it to classify diarrhea diseases.

Machine learning techniques have been proved to learn the pattern in data and predict similar patterns in new data efficiently and accurately rather than the classical methods since they follow predefined rules, the latter investigates only the relationship between covariates while machine learning algorithms analyze predictors automatically, they reveal unseen trends because they don't consider priori assumptions such as the type of error distribution, the additivity of parameters, and they have the capabilities of tuning parameters. Moreover, they provide a robust model with high accuracy hence they deliver more accurate results in predicting disease (Hema Sekhar Reddy Rajula, 2020).

A study by (Md. Maniruzzaman, 2020) identified that a mother's education, region, age of the child, and household wealth index have a statistical significance towards childhood diarrhea. He reported that mothers with no education had a high prevalence of childhood diarrhea compared

with mothers with secondary or higher education and children from age group 12-23 and 24-59 months have less prevalence of diarrhea.

The researcher continued applying machine learning algorithms to predict childhood diarrhea in Bangladesh. Support vector machine, naïve Bayes, linear discriminant analysis, quadratic discriminant analysis were used for prediction, and support vector machine was found to predict well the diarrhea disease than the other algorithms considered.

Xinyu Fang compared a random forest model to autoregressive integrated moving average (ARIMA) models to predict infectious diarrhea disease in Jiangsu Province in China where meteorological factors such as precipitation, relative humidity, atmospheric pressure were considered. ARIMA models assume a linear relationship between the dependent variable and independent variables, therefore, it failed to predict the incidence of diarrhea since meteorological factors have no linear relationship, however, the Random Forest model fitted well the data and predicts the infectious disease with good accuracy (Xinyu Fang, 2020).

He recommended that other factors associated with infectious diarrhea might also be considered as good predictors and be studied in the future. He continued proposing further studies to be conducted to investigate a random forest model with meteorological variables and other variables for the development of a functional tool for predicting other major infectious diseases.

Abdullah Zahirzda used a predictive model of childhood diarrhea with a cross-sectional study of Afghanistan Demographic Health Survey where naïve Bayes, random forest, and support vector machine (SVM) algorithms were considered for that particular task and the findings revealed that the best classifier was Random Forest with 81.84 % of accuracy (Abdullah Zahirzda, 2021).

2.4 Conceptual framework

Several risk factors associated with diarrheal diseases have been found by different researchers and were discussed in this part.

Brian A Maponga found that there is a statistical significance between family sourcing water outside the home, hand washing in a single bowl, garbage near home, flies near home, and contracting diarrhea disease among under children. Therefore, he identified some protective

factors for getting diarrhea disease such as using municipal water sources treating water with chlorine-containing water purification tablets, boiling water, keeping water in a closed container, having a hand washing facility in the family, and breastfeeding exclusively for 6 months. (Brian A Maponga, 2013).

Getachew discovered that the age of the child, maternal education, household income, hygiene of feeding practices, breastfeeding condition, malnutrition, personal hygiene, environmental sanitation, water availability and quality, and latrine utilization are determinants that influence the occurrence of diarrhea among children. (Getachew Yismaw Workie, 2019).

Among explored determinants of diarrhea disease, (Shyam Sundar Budhathok, 2016) identified that the age of the child, the sex, nutrition status of the child, washing hand practice and education of the mother, and other socio factors such as water and sanitation, cultures/society values, wealth index, and healthcare services are the main factors of Diarrhea in Nepal.

This part of the literature review depicts the linkage between the outcome variable and the predictors. In this study, the researcher has considered the occurrence of diarrhea disease or not as the outcome variable and the considered predictors are the age of the child, sex of child, residence, mother's education, maternal employment, household wealth status, type of toilet facilities, toilet facilities shared with other households, source of drinking water, stool disposal of a child, main material floor, breastfeeding practices, full received rotavirus vaccine, number of children under five living in the same household, annual precipitation and mean temperature.

Below is a figure illustrating the linkage between the outcome variable and the independent variables

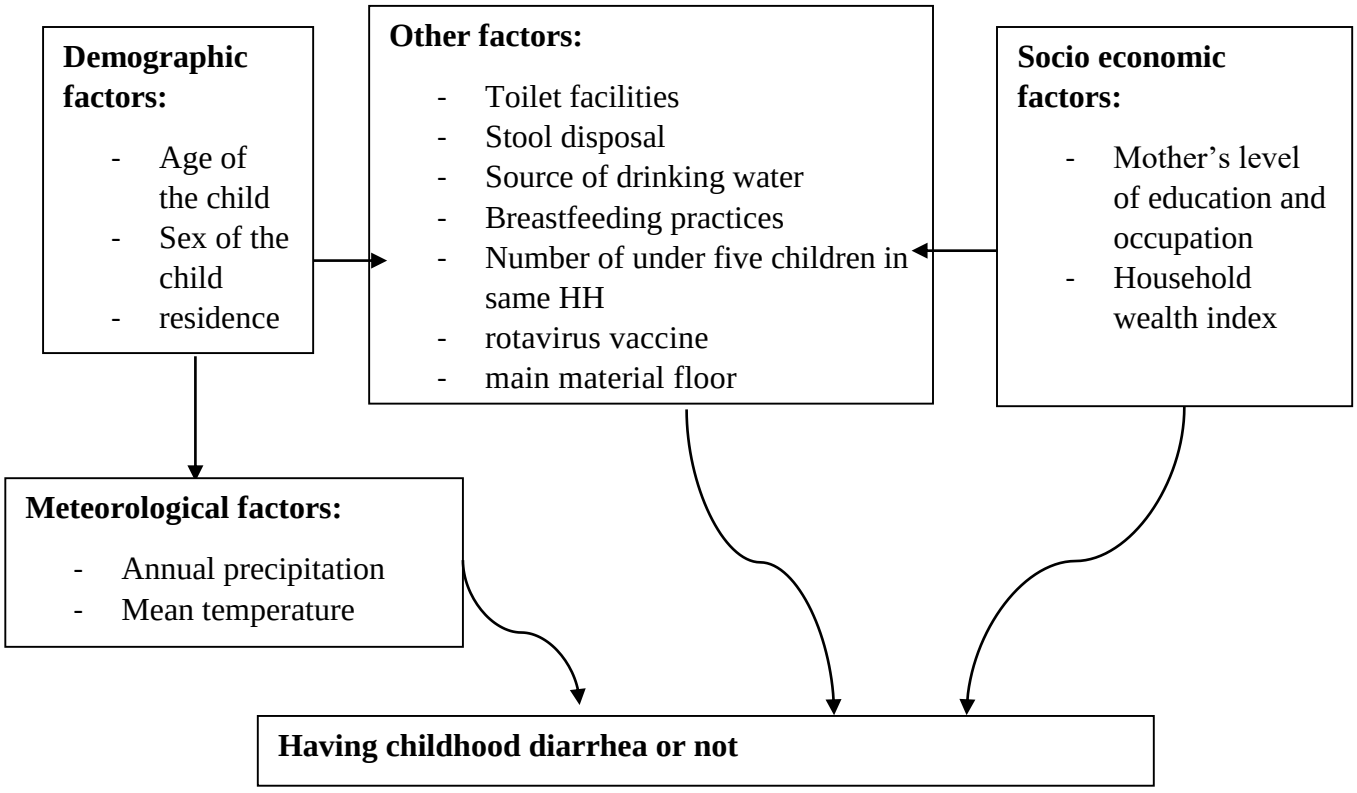


Figure 1: Conceptual framework diagram

CHAPTER THREE: METHODOLOGY

This section explains the source of data, the sampling design and sampling frame, detailed description of data and methodologies used in data pre-processing and analysis.

3.1 Data source

This study has used secondary data from the Rwanda Demographic Health survey 2014-15 and spatial data repository which gives geographically related health and demographic data. It is conducted through the Ministry of Health (MOH), National Institute of Statistics of Rwanda (NISR) together with the technical support of ICF International. DHS is a cross-sectional study that gathers information related to demographic and health indicators.

3.2 Sampling design and sampling frame

The sampling frame considered for the 2014-2015 Demographic Health survey is the fourth Rwanda Population and Housing census (RPHC), the latter was carried out in 2012 by the National Institute of Statistics of Rwanda (NISR). The sampling frame comprises an entire list of Enumerations Areas (EA) covering the whole country. An EA is a natural village, or a part of a village produced for 2012 RPHC and they are considered as the counting unit of the census.

A two-stage sample design was followed in the 2014-15 RDHS, it allowed to make estimates of main indicators at the national level, urban/rural areas, 5 provinces, and thirty districts for some limited indicators. In the first stage, clusters were selected considering the sampling frame and 492 clusters were formed where 379 are in rural areas and 113 in urban areas.

The second stage considered a systematic sampling of households, random sampling was performed on a list of households and twenty-six households were selected from each cluster, therefore 12792 households were considered as sample size, nevertheless, one household was found to be two households hence the sample size increased to 12793 households. Since the sample size in every district is approximately equal, the sample is not self-weighing at the national level and weighting factors were included in the data file for the results will be proportional at the national level.

For the target population, all women and all men in half of the household whose age is between15-49 and who were either permanent residents of the household or visitors the night preceding the survey were eligible for the interview in half of the households, all men aged 15-59 who were either permanent residents or visitors the night preceding the survey were eligible for the interview.

3.3.Sample size

In this study, the children recode which characterizes every child of interviewed women born in the five years before the survey was considered, therefore the total sample size of alive children aged 0 to 59 months is 7474 and those women were asked if their children have experienced diarrhea in two weeks preceding the survey. The children recode was merged with DHS geospatial covariates to get meteorological variables such as annual precipitation and mean temperature.

3.4 Ethical consideration

The access on 2014-15 RDHS was done through online registration on the DHS program where the access and use of the dataset were approved. The DHS data should only be used for the registered study and statistical analysis and reporting. DHS data should be treated anonymously and no attempt has to be put in identifying any respondents interviewed in the survey and the data shouldn't be handed to other researchers. Moreover, the users of DHS data must submit a copy of the report or any publication resulting from DHS data files.

3.5 Description of variables

This study has considered the children Recode (KR) dataset and below are the variables that were included in the study. The dependent variable is "a child had diarrhea or not in two weeks preceding the survey" and the independent variables are rotavirus vaccine, education and employment of the mother, residence place, source of drinking water, type of toilet facility, toilet facility shared with other households, main material floor, wealth index, breastfeeding practices, disposal of child's stool, number of under-five children living in the same household, sex of the child, age of the child in months, annual precipitation and mean temperature.

Table 1: Variables Description

VARIABLES	DESCRIPTION	CATEGORY
The child had diarrhea or not (H11)	If a child experienced diarrhea or not in two weeks preceding the survey	0: had no diarrhea 1: had diarrhea
Maternal education (V106)	Education level of the mother	0: no education 1: primary 2: secondary 3: higher
Residence (V102)	Residence place of the household	0: urban 1: rural
Rotavirus vaccine (rotavirus)	Full immunization with rotavirus vaccine	0: didn't receive the vaccine 1: received the vaccine
Source of drinking water (v113)	Source of drinking water	0: unimproved source 1: improved source
Type of toilet facility(v116)		0: unimproved toilet facility 1: improved toilet facility
Toilet facility shared(v160)	Toilet facility shared with other HH	0: no 1: yes
Main material floor (v127)	The main material floor of the household	0: earth, sand 1: dung 2: ceramic tiles 3: cement 4: other
Wealth index (v190)	Wealth status of the household	0: middle 1: poor 2: rich
Disposal of child's stool (v465)	Where the child stool is disposed of when not using toilets	0: unsanitary disposal 1: sanitary disposal
Maternal employment(v714)	If the mother is currently employed	0: no 1: yes
Sex of the child(b4)		0: male 1: female
Age group	Age of the child in months	0:0-11 1:12-24 2:25-36 3:37-47 4:48+
Duration of breastfeeding (m4)		0: ever breastfed, not currently breastfeeding 1: never breastfed 2: still breastfeeding

Number of children under 5 under a household(v136)
Annual precipitation
Mean Temperature

3.6 Data pre-processing

Some variables such as the source of drinking water, toilet facility, and child waste disposal have been categorized. According to the Guide to DHS Statistics 7, the researcher created new categories for some variables, where the source of drinking water has two categories: improved and unimproved source, the improved source includes piped into dwelling, piped to yard/plot, public tap or standpipe, piped to a neighbor, tube well or borehole, protected well/spring and rainwater. Unimproved water source comprises unprotected well/spring, river/lake, tanker truck, cart with a small tank and others.

For toilet facility, the improved facility contains flush to a piped sewer system, flush to the septic tank, flush to pit latrine, flush unspecified, pit latrine VIP, pit latrine with slab and composting toilet. On the other hand, unimproved facilities are flush to somewhere else, flush to an unknown place, pit latrine without slab/ open pit, no facility/bush/field, bucket toilet, hanging toilet, and others.

For child stool disposal, sanitary disposal includes putting the feces in the toilet/latrine, using the toilet, and burying the feces. Moreover, for unsanitary disposal, there are put/ rinsed into drain or ditch, feces thrown into the garbage, feces left or buried in the open and to unknown places.

To increase the performance of the model, processing variables play a key role therefore K Nearest Neighbor imputer has been used to handle missing values in variables. KNN imputer has been proved to be effective, and the missing value is replaced with the nearest neighbor estimated values. To handle class imbalance, SMOTE (Synthetic Minority Oversampling Technique) function has been applied, this method creates a new example of the minority class with the nearest neighbor and under-sample the majority class. All categorical variables have been transformed with dummies function since they were nominal so that the model can be capable to comprehend

and extract valuable information. For continuous variables, normalization has been performed since data had different scales.

3.7 Data analysis

The analysis of the research was performed with Python language using google colab based on Jupiter notebook.

In response of the 1st objective, Chi-square test of independence was used to identify risk factors and to determine the association between categorical variables and the independent variables, moreover, descriptive statistics were conducted for continuous variables.

In response to the 2nd objective, six supervised learning algorithms for classification were developing to predict the occurrence of diarrhea disease among children under five.

To find out the best model in predicting accurately the occurrence of diarrhea disease among children under five, the evaluation metrics were compared to assess the effectiveness of the model and this responds to the third objective. However, variables that were included in the development of the model were found to be associated or correlated to the target variable.

In response to the 4th objective, feature importance technique was performed to find out most important features contributing in predicting diarrhea disease among children

3.7.1 Supervised learning classifiers

Machine learning uses different statistical, probabilistic, and optimization methods to learn from past data and to detect important patterns from a large dataset. It is comprised of four techniques. There is **supervised machine learning** where data given to the model are labeled and the task is to predict labels for new data, for **unsupervised machine learning** the trained data on the model is unlabeled and the task is to find groupings and structure in the data. **Semi-supervised learning** has both unlabeled and labeled data and for **reinforcement**, the machine is trained to make specific decisions based on the business requirement with the sole motto to maximize efficiency.

Supervised machine learning is suitable to solve two types of problems for instance classification problems and regression problems. This study has considered classification supervised learning since the target variable is labeled and it has two classes (0= no diarrhea and 1= had diarrhea) and its purpose is to build the model by analyzing inputs data and predict the future trend of data, therefore six machine learning models were used such as **naïve Bayes classifier, random forest classifier, gradient boosting classifier, logistic regression, support vector machine, and artificial neural network.**

3.7.1.1 Naïve Bayes Classifier

Naïve Bayes classifier is based on Bayes' theorem where it predicts a category considering available features using probability and it is easy to implement. Bayes theorem shows the probability of an event based on the prior knowledge of conditions related to that event. Even if class' features could correlate with each other, naïve Bayes presumes that a feature in a class is not directly related to any other features.

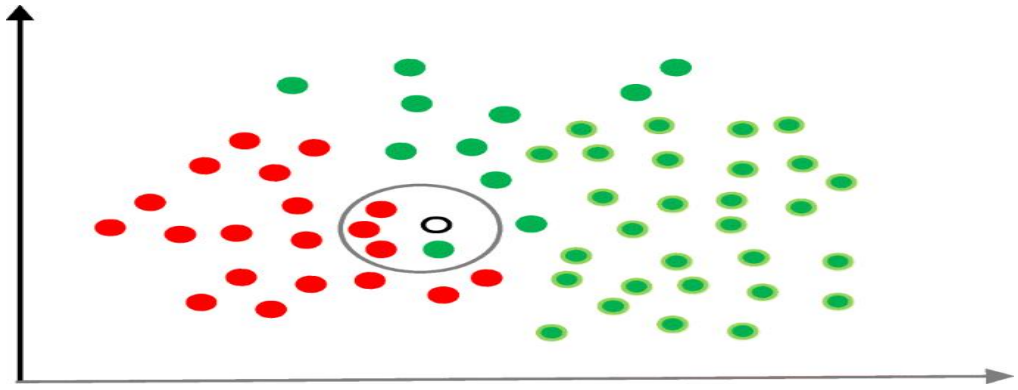


Figure 2: Example of Naïve Bayes classifier

3.7.1.2 Random Forest Classifier

Random Forest classifier is an ensemble of several decision trees, the latter are trained using different subsets of the training dataset, and to classify a new sample, its input vector must be transmitted with each decision tree of the forest and provide an outcome for the classification.

The random forest reduces the variance as it considers different decision trees and results in decreasing the over fitting of the training dataset.

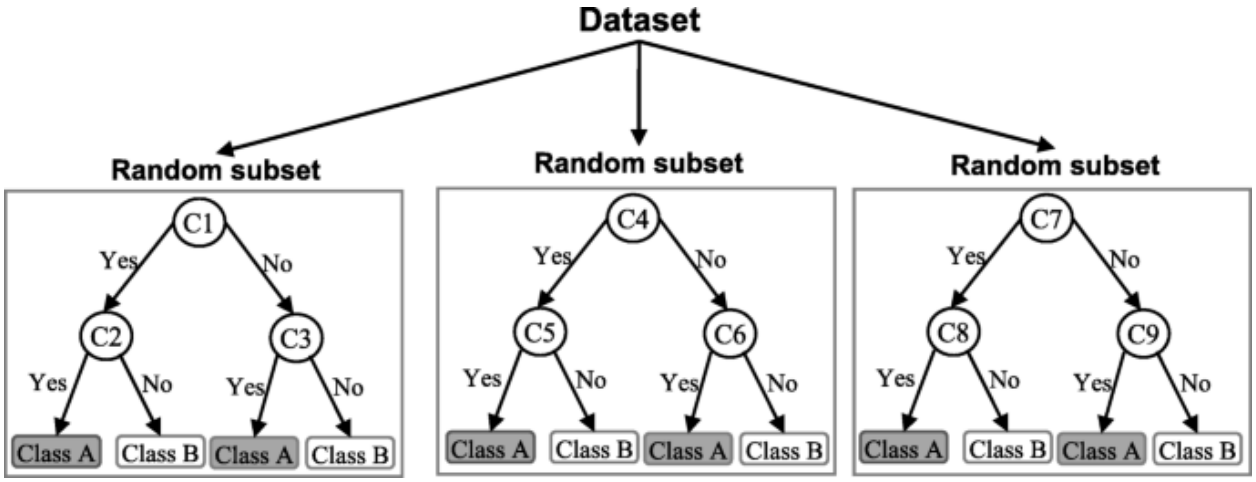


Figure 3: Example of Random Forest Classifier

3.7.1.3 Artificial Neural Network

The artificial neural network is motivated by the performance of the neural network of the human brain. In the human brain, neurons are linked to each other by multiple axon junctions and they facilitate the adjustment, process, and storing the information. An artificial neural network could be characterized by a group of nodes where the output of one node is the input of another node and nodes are classified as a layer. ANN is useful in solving a non-linear relationship in the data.

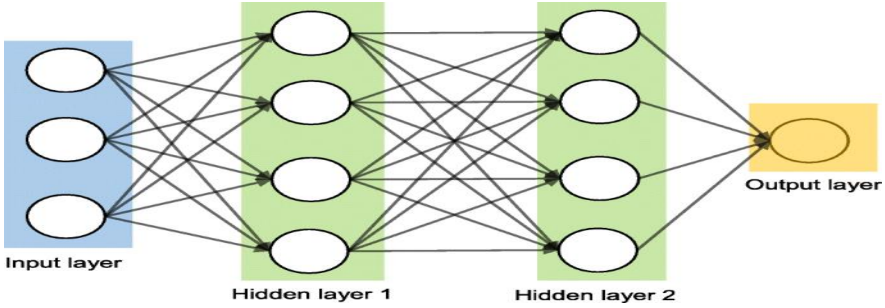


Figure 4: Example of Artificial Neural Network

3.7.1.4 Logistic Regression

Logistic regression is also among the supervised learning specialized in classification, it is a probabilistic model that intends to predict the probability of the dependent variable, moreover, the dependent variable is dichotomous, it has only two categories. As a traditional approach, it aims to find out the influence of one or more independent variables upon the dependent variable.

3.7.1.5 Support Vector Machine

This algorithm can be applied for both regression and classification problems, it represents different categories or classes in a hyperplane in multidimensional space, that decision boundary must be proficient to differentiate the two classes. The support vector machine can have several hyperplanes, but the function selects the best one that divides well the two classes, the main purpose of the algorithm is to find the best hyperplane which maximizes the margin. Support vectors are the closest observations to the hyperplane.

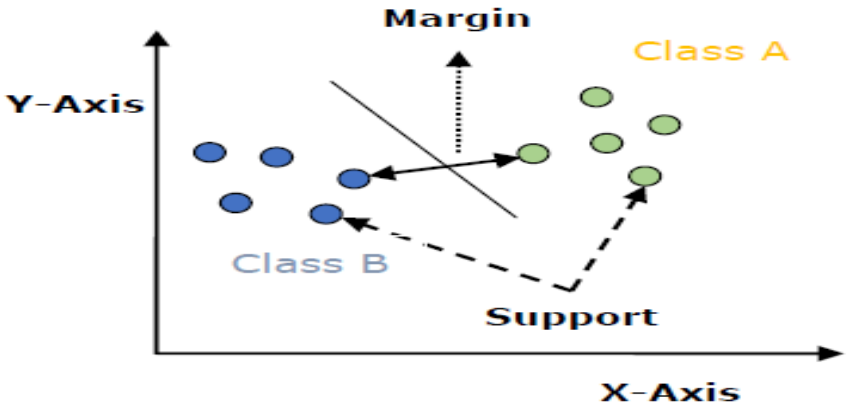


Figure 5: Example of Support Vector Machine

3.7.1.6 Gradient Boosting Classifier

This technique gives an additive predictive model by forming an ensemble of weak predictors, mainly decision trees models; and it minimizes the loss function by choosing a function iteratively this loss function facilitates understanding how accurate a model is to classify the two classes. At each iteration, weight is added to the observations with the worst prediction from the preceding iteration and tries to improve the results.

3.7.2 Evaluation of the model

To assess the success and the effectiveness of the model, the researcher used the accuracy, recall, precision, F1 score, confusion matrix, and ROC curve.

Confusion Matrix represents the overview of how the model is doing and it tabulates the actual values versus predicted values, where True positive (TP) are the positive cases that the model predicted while they were positive, True negative (TN) are the negative cases that the model predicted while they were negative, False positive (FP) are positive cases that the model predicted yet they were negative, False negative (FN) are negative cases that the model predicted yet they were positive.

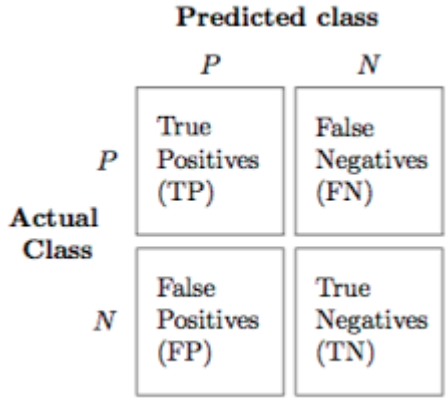


Figure 6: Confusion Matrix

Accuracy of the model is defined as the simplest and easiest metric to interpret; it measures the ratio of correctly predicted observation to the total observations

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

Precision shows how the model result can be reliable if it shows that a point belongs to that class. It is a ratio between the true positive among all positives. It represents the number of positive cases out of all positive cases the model predicts.

$$\boldsymbol{precision} = \frac{\boldsymbol{TP}}{\boldsymbol{TP+FP}}$$

Recall of a class defines how well the model can detect that class. It measures how correctly the model identifies the true positive.

$$\boldsymbol{Recall} = \boldsymbol{Sensitivity} = \boldsymbol{True\ positive\ rate} = \frac{\boldsymbol{TP}}{\boldsymbol{TP} + \boldsymbol{FN}}$$

The F1 score is the harmonic mean of precision and recall, it is used when the precision and recall are both equally important, therefore the F1 score can indicate a good value of precision and a good value of precision simultaneously.

$$\boldsymbol{F1\ Score} = 2 * \frac{\boldsymbol{precision} * \boldsymbol{recall}}{\boldsymbol{precision} + \boldsymbol{recall}}$$

ROC Curve which stands for Receiver Operating Characteristic refers to a graph that illustrates the performance of the classification model; this metric enables binary classification problems to comprehend how well the classifier is doing. It plots the true positive rate against the false-positive rate.

3.7.3 Feature importance Technique

Feature importance is a method that allocates the score on the independent variables considering how they are of use in predicting the dependent variable. This technique provides insight into the model by revealing variables that are most and least important to the model while making predictions and it was calculated under the best model that uncovered its ability to predict children's diarrhea status. Class feature importance was run to identify important variables on each category of the dependent variables and graphs were generated.

CHAPTER FOUR: DATA ANALYSIS

4.1 Introduction

This part portrays the findings of the research where descriptive analysis was performed, chi-square test has been conducted to find out the association between the independent variables and the target variable, moreover, the machine learning algorithms have been run to find the best model to predict the diarrhea disease, finally feature importance technique has been conducted to find out the most important feature in making the prediction.

Below, different charts and tables that display the frequencies and percentage of observations based on input features and the target feature considered in the study were created. This study had 16 independent variables and one dependent variable, out of 16 independent variables thirteen were categorical and 3 were continuous or numeric.

Categorical variables are rotavirus vaccine, duration of breastfeeding, age group, sex, residence, mother's employment, mother's education, wealth index, type of toilet facilities, source of drinking water, main material floor, and toilet facility shared with other households, and disposal of child's stool. Continuous variables are annual precipitation, mean temperature, and the number of children under five under a household.

4.2 Exploratory data analysis

This study considered 7474 children under the age of five as the population size and out of that population 6569 of them had no diarrhea and only 905 (12%) children experienced diarrhea two weeks before the survey. The figure below illustrates children who experienced diarrhea episodes or not two weeks preceding the survey where **class 0** had no diarrhea and **class 1** had diarrhea.

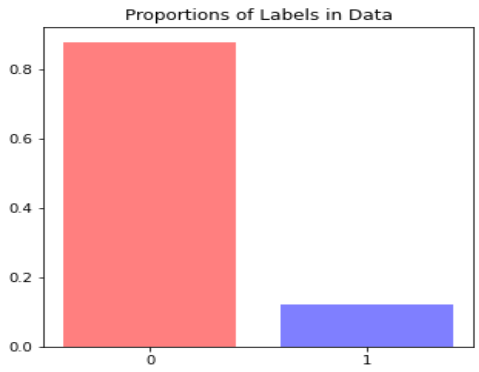


Figure 7: Target variable

4.2.1 Univariate statistics for socio-demographics and meteorological factors

The table below represents descriptive statistics of both categorical and continuous variables. In a population of 7474 under-five children, 50.39% represents male children and 49.61% are female children. For child age, the highest peak was among children between 0 and 11 months (23.87%) and the lowest was among children between 48 months and above with 15.18%. Among 7474 children whose age is under five, 21.88% are urban residents and 78.12% are rural residents. For the wealth index, the highest percentage was 44.76% of households were poor and the lowest percentage of households were in the middle class (19.01%).

The table below shows a summary of frequencies of other categorical variables belonging to socio-economic factors. Among 7474 children under five, 51.36% were still being breastfed and only 0.35% were never breastfed. Out of all women with under-five children, 14.28% had no education, 71.65% had primary education, 11.41% had secondary education and only 2.66% had higher education. For mothers' employment, the study showed that 14.5% were not working and 85.5% were working currently.

For source of drinking water, out of all households with children under five 71.8% used improved sources of drinking water and only 28.2% used unimproved ones. Only 29.76% of households use no improved toilet facilities and 70.24% of them use improved toilets facilities.
Out of all households with under-five children, 78.59% of them don't share toilet facilities with other households while 21.41% of them do share.

Only 14.38% of households use unsanitary disposal for disposing of child stool whereas 85.62% of them used sanitary disposal. For the main material floor, among all the households, 75.39% had earth and sand and only 0.16% uses other materials as the main floor.

Table 2: Frequencies of Categorical Variables

Variables		**Frequency**	**Percent**
Sex	**Male**	3766	50.39
	Female	3708	49.61
	Total	7474	100
Age	**0-11**	1784	23.87
	12-24	1729	23.13
	25-36	1344	17.98
	37-48	1483	19.84
	48+	1134	15.18
	Total	7474	100
Residence	**Urban**	1635	21.88
	Rural	5839	78.12
	Total	7474	100
Wealth Index	**Middle**	1421	19.01
	poor	3345	44.76
	Rich	2708	36.23
	Total	7474	100
Duration of breastfeeding	**Ever breastfed, not currently**	3609	48.29
	never breastfed	26	0.35
	still breastfeeding	3839	51.36
	Total	7474	100
Maternal education	**No education**	1067	14.28
	primary	5355	71.65
	secondary	853	11.41
	higher	199	2.66
	Total	7474	100
Maternal employment	**Not working**	1084	14.5
	working	6390	85.5
	Total	7474	100
Source of drinking water	**unimproved**	2108	28.2
	improved	5366	71.8
	Total	7474	100
Type of toilet facility	**unimproved**	2224	29.76
	improved	5250	70.24

	Total	7474	100
Toilet facility shared with other households	**Not shared**	5824	78.59
	Shared with other HH	1600	21.41
	Total	7474	100
Received 3 doses of rotavirus vaccine	**no**	7043	94.23
	yes	431	5.77
	Total	7474	100
Disposal of child stool	**Unsanitary disposal**	1075	14.38
	Sanitary disposal	6399	85.62
	Total	7474	100
Main material floor	**Earth, sand**	5634	75.39
	dung	60	0.8
	Ceramic tiles	90	1.2
	cement	1678	22.45
	other	12	0.16
	Total	7474	100

Among other factors, continuous variables such as the number of under-five children living in the same household, annual precipitation and mean temperature were considered. Descriptive statistics were performed and the results are displayed in the table below. The study revealed that the mean number of under-five children living in the same household is 2, the mean annual precipitation is 118.5mm and the mean temperature was 19.3 Celsius degrees.

Table 3: Descriptive Statistics for Continuous Variables

Variables	**Mean**	**Standard deviation**
Number of under 5 children living in the same household	2	0.74
Annual precipitation	118.5mm	21.69
Mean temperature	19.3	1.38

4.2.2 Test of association between input features and the target feature

To test the association between explanatory variables and the target variable, the chi-square test has been applied for categorical variables where the null hypothesis was stated as “there is no association” and the alternative hypothesis as “there is an association between a given variable and

the target variable". For a probability value less than 0.05, the null hypothesis was rejected and the inference is made on the alternative hypothesis and vice versa.

The study found that child age group, residence, wealth index, type of toilet facility, main material floor, duration of breastfeeding, rotavirus vaccine and maternal education are associated with the dependent variable. On the other hand, sex of the child, toilet facility shared with other households, water source, maternal employment, and disposal of child stool are not associated with having or not having childhood diarrhea.

Below is the table showing the probability values of different categorical variables.

Table 4: Bivariate Statistics of Categorical Variables

		Prevalence of diarrhea disease		
Variables		**No**	**Yes**	**P value**
Sex	**male**	3290(44.02)	476(6.37)	0.15
	female	3279(43.87)	429(5.74)	
Age group	**0-11**	1547(20.7)	237(3.17)	0.000
	12-24	1384(18.52)	345(4.62)	
	25-36	1187(15.88)	157(2.1)	
	37-48	1367(18.29)	116(1.55)	
	48+	1084(14.5)	50(0.67)	
Residence	**urban**	1467(19.63)	168(2.25)	0.01
	rural	5102(68.26)	737(9.86)	
Wealth Index	**Middle**	1259(16.85)	162(2.17)	0.000
	poor	2848(38.11)	497(6.65)	
	rich	2462(32.94)	246(3.29)	
Duration of breastfeeding	**Ever breastfed but not currently**	3299(44.14)	310(4.15)	0.000
	never breastfed	23(0.31)	3(0.04)	
	still breastfeeding	3247(43.44)	592(7.92)	
Maternal education	**No education**	922(12.34)	145(1.94)	0.000
	primary	4686(62.7)	669(8.95)	
	secondary	767(10.26)	86(1.15)	
	higher	194(2.6)	5(0.07)	
Maternal employment	**Not working**	958(12.82)	126(1.69)	0.59
	working	5611(75.07)	779(85.5)	
Source of drinking water	**unimproved**	1842(24.65)	266(3.56)	0.39
	improved	4727(63.25)	639(8.55)	
Type of toilet facility	**unimproved**	1901(25.43)	323(4.32)	0.000
	improved	4668(62.46)	582(7.79)	
	Not shared	5184(69.36)	690(9.23)	0.0661

Toilet facility shared with other households	**Shared with other HH**	1385(18.53)	215(2.88)	
Rotavirus vaccine	**no**	6205(83.02)	838(11.21)	0.02
	yes	364(4.87)	67(0.9)	
Disposal of child stool	**Unsanitary disposal**	951(12.72)	124(1.66)	0.5331
	Sanitary disposal	5618(75.17)	781(10.45)	
Main material floor	**Earth , sand**	4906(65.64)	728(9.74)	0.0005
	dung	49(0.66)	11(0.15)	
	Ceramic tiles	79(1.06)	11(0.15)	
	cement	1525(20.4)	153(2.05)	
	other	10(0.13)	2(0.03)	

4.3 Model development and results

In this study, different machine learning algorithms have been applied to predict diarrhea disease among children under five. The researcher considered six algorithms for classification problems such as random forest classifier, logistic regression, naïve Bayes classifier, support vector machine, artificial neural network, and gradient boosting classifier.

4.3.1 Random forest classifier

For this classifier, the train accuracy is 86.35 and the test accuracy is 83.07, this means that the model made correct predictions at 83.07% and this classifier is good since the area under the curve is 0.91. According to the confusion matrix, the false-negative rate equals 13.8% and the false positive rate is 19.9%. Furthermore, the precision of 81.17 means that when the model predicts that a child has diarrhea, it is correct at 81.17 % of the time. For the recall, the model correctly identifies 86.11% of children with diarrhea and the F1 score is 83.57.

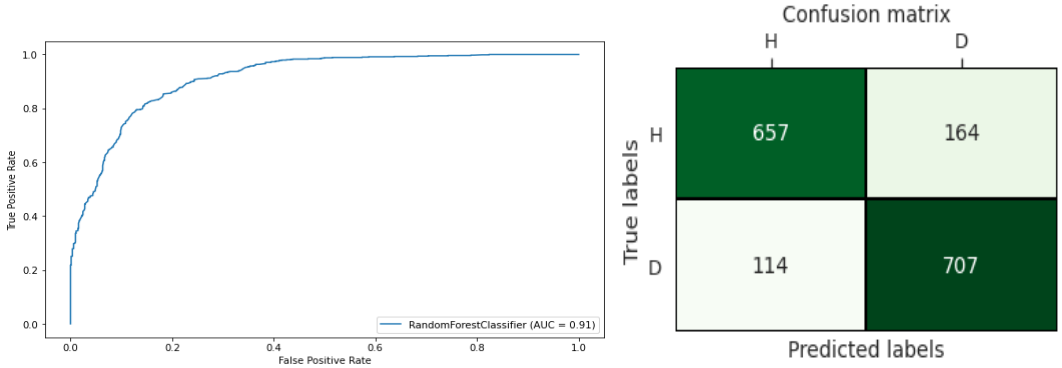


Figure 8: Random Forest Classifier

4.3.2 Logistic regression

For logistic regression, the train accuracy is 61.56 and the test accuracy is 62 which means that the model made correct predictions at 62% of the time and the classifier is moderate as the ROC curve distinguishes the two classes at 66%. For Precision, it is correct 61.08% of the time when the model predicts children with diarrhea and for the recall, the model correctly identifies 66.14 % of children with diarrhea, and the F1 score equals 63.51. Looking at the confusion matrix, the false-negative rate equals 33.8% and the false positive rate is 42.1%.

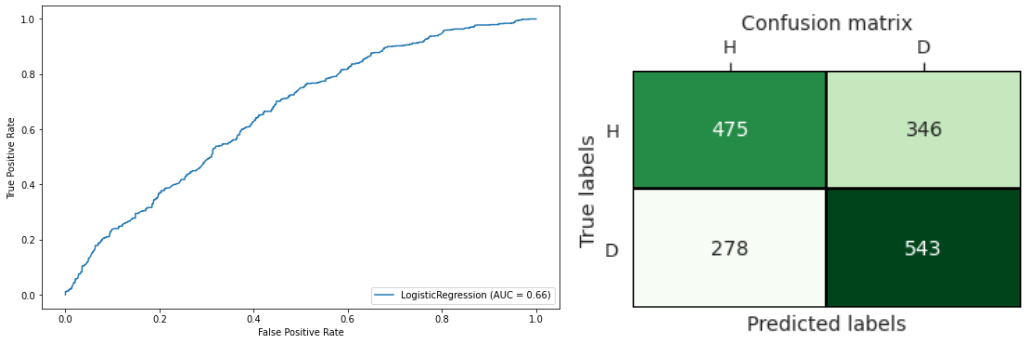


Figure 9: Logistic Regression

4.3.3 Naïve Bayes classifier

According to the naïve Bayes classifier, the train accuracy is 59.4 and the test accuracy is 60.72, this shows that the model made correct predictions at 60.72 % of the time and the classifier is not good since the area under the curve is 0.65. Precision was found to be 57.37, which shows that when the model predicted children with diarrhea, it was correct at 57.37% and for the recall, the model correctly identifies 83.43% of children with diarrhea, and the F1 score was 67.99. Considering the confusion matrix, the false-negative rate equals 16.5% and the false positive rate is 61.9%.

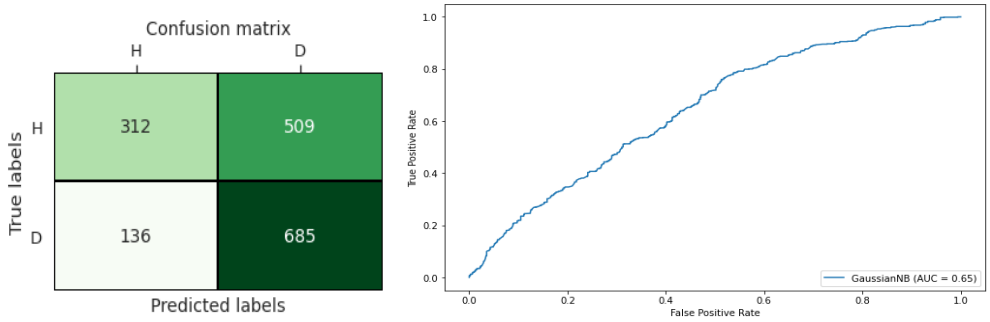


Figure 10: Naive Bayes Classifier

4.3.4 Support vector machine

For this classifier, the train accuracy is 75.17 and the test accuracy is 71.13, therefore the model makes correct predictions at 71.13% of the time. For the precision, if the model predicts children with diarrhea, it would be correct at 68.48% and for the recall, the model can correctly identify 78.32 % of children with diarrhea and the F1 score was 73.07. The ROC curve illustrated the area under the curve which equals 0.77 hence the classifier's ability to distinguish children with diarrhea or not is 77% and when examining the confusion matrix, the false-negative rate is 21.6% and the false positive rate is 36%.

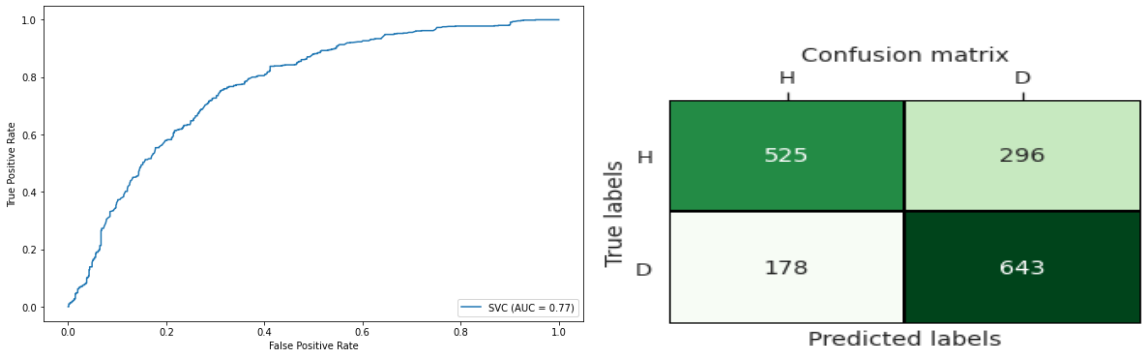


Figure 11: Support Vector Machine

4.3.5 Artificial neural network

Referring to the artificial neural network algorithm, the train accuracy is 76.2 and the test accuracy is 73.3 thus the model made correct predictions at 73.3% of the time. For precision, it is correct at 70.09% of the time when the model predicts children with diarrhea disease, for the recall, the model correctly identifies 81.36% of children with diarrhea, and the F1 score was 75.31. Considering the confusion matrix, the false-negative rate is 18.6% and the false positive rate is 34.7% and the ROC curve revealed that the area under the curve is 0.81, the classifier is good at 81% to separate children with/without diarrhea.

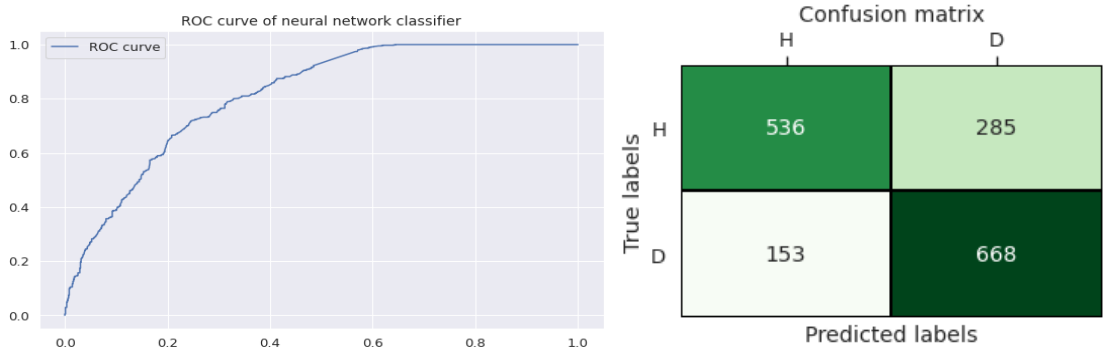


Figure 12:Artificial Neural Network

4.3.6 Gradient boosting classifier

For gradient boosting classifier, the train accuracy is 86.45 and the test accuracy is 86.36, this means that the model predicts correctly at 86.36% of the time. For precision, when the model predicts children with diarrhea, it is correct 82.84 of time, for the recall, the model identifies almost perfectly children with diarrhea at 91.72 and the F1 score was 87.05. Looking at the graph of the confusion matrix, the false-negative rate is 8.2% and the false-positive rate is 19%. The classifier's ability to separate the two classes is 0.95; this gradient boosting classifier is almost perfect.

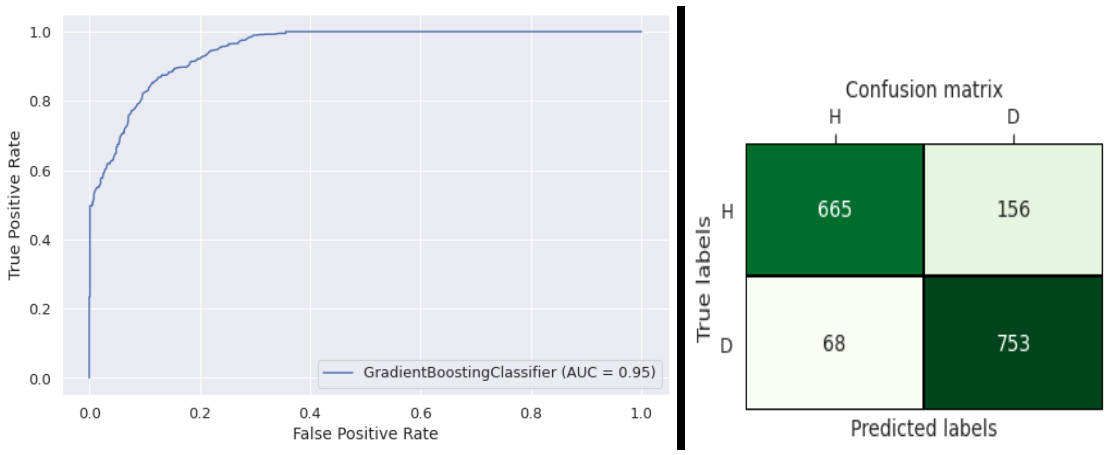
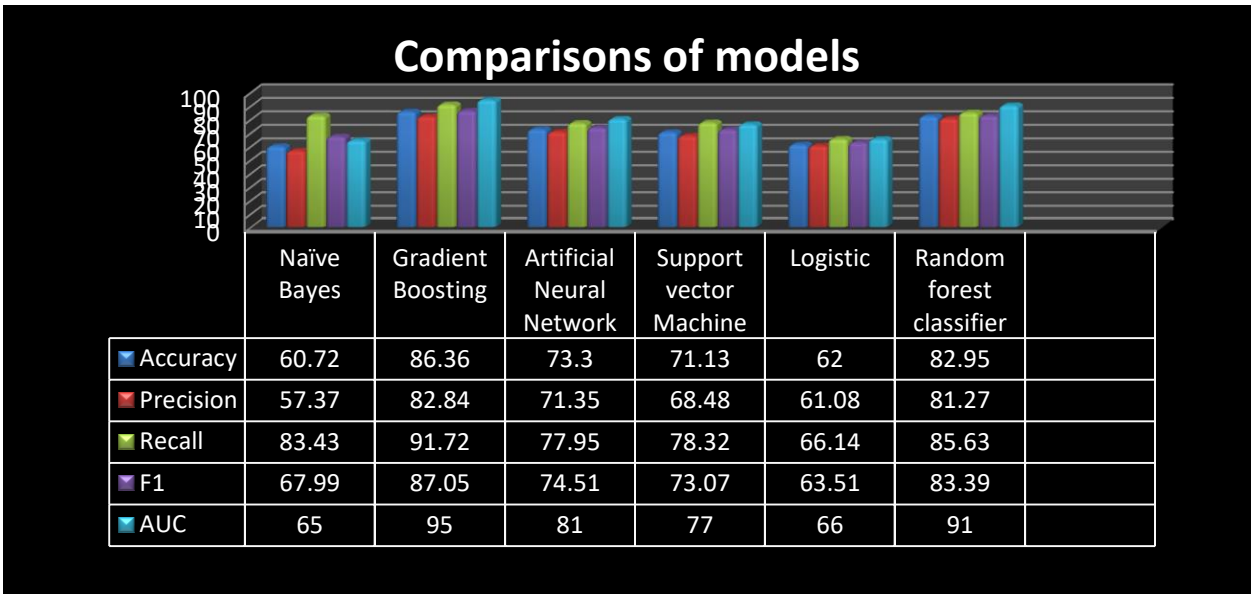


Figure 13: Gradient Boosting Classifier

4.4 Model comparison

The graph below illustrates how evaluation metrics differ on each algorithm.

Therefore, the best model was predicted by gradient boosting classifier since it has a high area under the curve, together with high precision and recall, furthermore, this algorithm demonstrated to be an almost perfect classifier with the ability to separate correctly children with diarrhea and children with no diarrhea at 95% and with the lowest false-negative rate of 8.2%.



	Naïve Bayes	Gradient Boosting	Artificial Neural Network	Support vector Machine	Logistic	Random forest classifier	
Accuracy	60.72	86.36	73.3	71.13	62	82.95	
Precision	57.37	82.84	71.35	68.48	61.08	81.27	
Recall	83.43	91.72	77.95	78.32	66.14	85.63	
F1	67.99	87.05	74.51	73.07	63.51	83.39	
AUC	65	95	81	77	66	91	

Figure 14: Comparison of models

4.5 Feature importance

Here, feature importance was computed on the gradient boosting classifier since it was the robust model. This section is comprised of selecting features that are judged to contribute importantly to building a predictive model of childhood diarrhea. Below the figure shows the important feature where the top five important features for predicting childhood diarrhea are high annual precipitation, children whose age group is 12 to 24 months, children from poor families and households with unimproved toilet facilities, and who material floor is earth and sand.

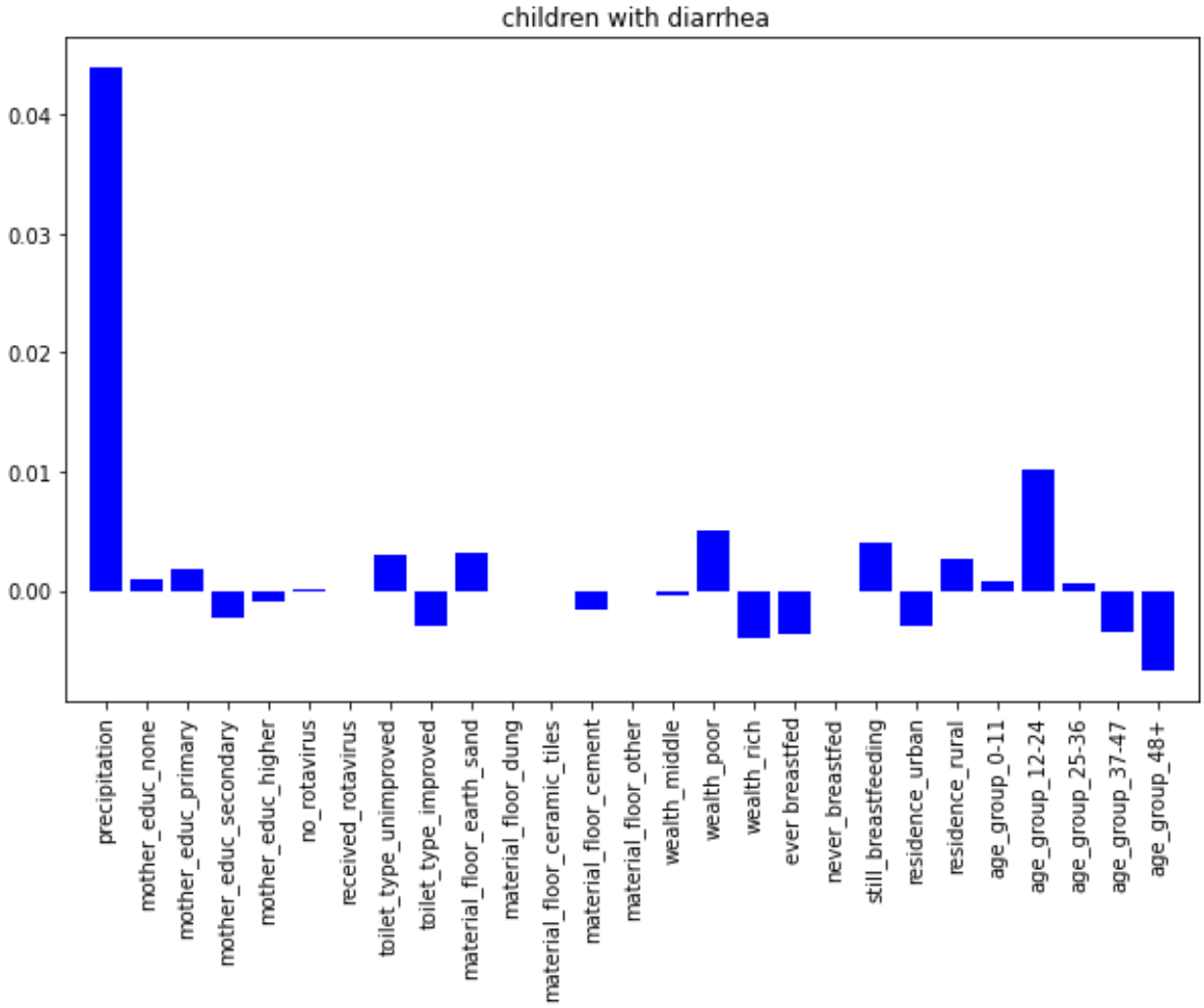


Figure 15: Feature Importance

CHAPTER FIVE: DISCUSSION OF FINDINGS

This chapter highlights the discussion of the research findings, the conclusion, and recommendations regarding different entities.

5.1 Discussion

The research purposed to build a predictive model for the occurrence of diarrhea disease among children fewer than five years. Machine learning techniques designed for classification problems were applied to develop that model and six algorithms were used such as random forest classifier, logistic regression, naïve Bayes classifier, support vector machine, artificial neural network, and gradient boosting classifier.

To evaluate the performance of the models, some evaluations metrics for instance accuracy, precision, recall, ROC curve, and f1 score have been adopted to assess the ability of the models to predict diarrhea disease among children under five. Accuracy is the easiest metric used by different researchers, but it performs poorly in case of class imbalance and the recall measures how well the model identifies the positive class. Therefore, this discussion focused also on the recall metric since the researcher aimed to minimize the false negative.

For this research, six classifiers were compared considering their evaluation metrics and the best model was assessed in terms of accuracy, recall, and area under curve (AUC) and the study revealed that gradient boosting classifier outperformed other models with high accuracy; high recall, and high AUC and this classifier is an additive model which improves weak learners and minimizes the loss.

The gradient boosting classifier is accurate at 86.36% for making correct predictions and it has the ability of 91.72% to correctly identify children with diarrhea disease. Moreover, the classifier came out as an almost perfect classifier since the area under the curve was 0.95, and this illustrates the ability of the classifier to distinguish children with diarrhea and children with no diarrhea at 95%. When looking at its confusion matrix, the gradient boosting model will predict correctly positive

cases of children with diarrhea and only 8 % of children will be misclassified to have no diarrhea yet it is not the case, this will not be an issue because the main sign of this disease is watery stools, so it will be identified easily without running any test.

The model performed well considering other algorithms used by other researchers since (Abdullah Zahirzda, 2021) found in a cross-sectional study undertaken on children in Afghanistan that random forest was the best model with an accuracy of 81.48%, his model identified correctly 82% of children with diarrhea and the area under the curve was 89.8%. (Md. Maniruzzaman, 2020) conducted a similar study in Bangladesh and uncovered that the support vector outperformed other classifiers with 65.61% of accuracy and 66.27% of recall.

Moreover, referring to the figure 8, the gradient boosting classifier revealed important factors that contributed to the prediction of diarrhea disease among children aged less than five through feature importance methodology, those factors are high annual precipitation, children with 12 to 24 months, household from poor class, a household with earth and sand as main material floor and households with unimproved toilets.

Precipitation has been found to have a quite remarkable influence in predicting diarrhea disease among children and this study conforms with (Xinyu Fang, 2020) who explained that precipitation influenced in predicting the incidence of infectious diarrhea by using a random forest model, and this might due to how the incidence of diarrhea vary with climate factors where high precipitation can flush enteric pathogens from waste in canals used for drinking water source since the source is polluted and this lead to significant exposure and the occurrence of diarrhea.

Children aged 12 to 24 months contributed highly as well in predicting childhood diarrhea, this is in line with the findings of (Getachew Yismaw Workie, 2019) because at that age children are not being breastfed exclusively, they started walking and they are exposed to the household environment, to unclean food and water and sometimes they are left to play without any supervision.

Belonging to poor households influenced the model in predicting diarrhea status among children and this is in line with (Wondwoson Woldu, 2016) where the occurrence of the disease was more

likely in poor families than to their counterparts, and the explanation for that is that people from that class have a poor living condition, as a result, they might not have adequate sanitation and clean water.

Earthen and sand floor contributes in predicting children with diarrhea disease and this is due to the fact such floor contains microorganisms that can cause diarrhea. This is similar to the finding of (Jean Nsabimana, 2017) who uncovered that the earthen floor is associated with diarrhea disease. Unimproved toilet showed an influence in predicting diarrhea among children, which is consistent with the study conducted in Malawi by (Juyoung Moon, 2019) where children with unimproved toilet facilities have a high chance of suffering from diarrhea disease and this is explained by the fact that the defecation will be disposed unsafely and this will attract flies that can contaminate food/water and freely spread the disease to people.

CHAPTER SIX: CONCLUSION AND RECOMMENDATIONS

6.1 Conclusion

The research was successful since all the objectives were met, the main objective was to build a model that predicts the occurrence of diarrhea disease among under-five children with machine learning techniques considering the socio-demographic and meteorological variables from RDHS 2014-2015, this was achieved with gradient boosting classifier with 86.36% of accuracy and the recall of 91.72% identifying children with diarrhea disease.

The model identified vulnerable groups of children susceptible to suffer from the diarrhea disease where children aged 12 to 24 months, children from a poor household, children from households with earth and sand as the main material floor, children with unimproved toilet facility and children from the region with high annual precipitation.

Therefore, this model could be operationalized and used by community health workers and at health centers level during routine immunization and growth monitoring session where they can use that system to detect the likelihood of a child to get diarrhea, from there they can provide advice and measures for mothers to prevent the disease earlier. Providing efficient measures will hinder severe diarrhea and dehydration earlier, and such interventions will mitigate the number of admissions in hospitals due to diarrhea, hence it will lessen the morbidity of diarrhea and the mortality of fewer than five children indirectly.

6.2 Recommendations

Per the findings of this study, the researcher has highlighted some recommendations towards different entities.

6.2.1 Recommendations to health organizations

As the model identified the most contributing factors to predict the occurrence of diarrhea among children with age less than five years, health institutions and other policymakers should use that

information to implement essential interventions which improve the quality of life of children and reduce the morbidity of the disease.

Climate factors such as precipitation had the greatest influence to predict diarrheal disease in this research, for that reason government and other stakeholders should consider the season variation to improve policies to reduce diarrhea disease.

Health organizations could use this model at health centers and at community health workers level to detect earlier the likelihood of diarrhea among children considering socio-demographics and meteorological factors and this would lessen the morbidity from diarrhea.

6.2.2 Recommendations to scientific and other researchers

This study highlighted the relevance of supervised machine learning techniques in predicting diarrhea disease, consequently, researchers should adopt other machine learning techniques and deep learning techniques to predict diarrhea and other diseases and develop a vigorous system that could be beneficial for different organizations and mitigate other health issues.

BIBLIOGRAPHY

Abdullah Zahirzda, G. C. (2021). A Data Mining Model for Predicting Diarrhea in Afghan Children.

Aldo A. M. Lima, D. B. (2019, February 8). Etiology and severity of diarrheal diseases in infants at the semiarid region of Brazil: A case control study. *PLOS Neglected Tropical diseases*.

Archana B Patel, M. M. (2011). what zinc supplementation does and does not achieve in diarrhea prevention: a systematic review and meta analysis. *BMC Infectious diseases*, 1471-2334/11/122.

B.Azhagusundari, A. S. (2013). Feature Selection based on Information Gain. *International Journal of Innovative Technology and Exploring Engineering (IJITEE)* , 2278-3075.

Boby Reiner, S. I. (2018). Variation in childhood diarrheal morbidity and mortality in Africa 2000-2015. *Journal of Medecine*.

Brian A Maponga, D. C. (2013). Risk factors for contracting watery diarrhoea in. *BMC INfectious diseases*, 13:567.

Christopher Troeger, M. I. (2018). Rotavirus Vaccination and the Global Burden of Rotavirus. Diarrhea Among Children Younger Than 5 Years. *JAMA Pediatrics*, 958-965.

Cynthia Boschi-Pinto, C. F. (2006). Disease and Mortality in Sub-Saharan Africa. 2nd edition. *National Center for Biotechnology Information*.

Getachew Yismaw Workie, T. Y. (2019). Environmental factors affecting childhood diarrheal disease among under-five children in Jamma district, South Wello zone, Northeast Ethiopia. *BMC Infectious Disease*, 19:804.

Hanifah Rohmah, T. H. (2015). Role of Exclusive Breastfeeding in Preventing Diarrhea. *Althea Medical Journal*.

health, M. o. (2018). *Fourth Health Sector Strategic Plan July 2018-June 2024.* Kigali.

Health, M. o. (2020). *Rwanda Health Sector Performance Report 2019-20.* KIGALI.

Hema Sekhar Reddy Rajula, G. V. (2020). Comparison of Conventional Statistical Methods with Machine Learning in Medicine: Diagnosis,Drug Development, and Treatment. *Medicina*, 56, 455.

Ibrahim Khalil, P. R. (2017). How to measure the true impact of diarrheal diseases. *DEFEATDD.*

Jean Nsabimana, C. M. (2017). Factors Contributing to Diarrheal Diseases among Children Less than Five years in Nyarugenge District, Rwanda. *Journal of Tropical Diseases*.

Juyoung Moon, J. W. (2019). Risk factors of diarrhea of children under five in Malawi: based on Malawi Demographic and Health Survey 2015–2016. *Journal of Global Health Science*.

Ladislas NSHIMIYIMANA, P. M. (2019). Diarrhoeal Diseases in Children under ve years. *Research square*.

Lei Tian, X. Z. (2016). Characteristics of bacterial pathogens associated with acute diarrhea in children under 5 years of age: a hospital-based cross-sectional study. *BMC Infectious diseases*, 16:253.

Liliana Carvajal-Vélez, A. A. (2016). Diarrhea management in children under five in sub-Saharan Africa: does the source of care matter? A Countdown analysis. *BMC Public Health*, 16:830.

Malachie Tuyizere, T. N. (2019). Factors Associated with Childhood Diarrhea in Rwanda: A Secondary Data Analysis of the Rwanda Demographic and Health Survey 2014-15. *Rwanda Journal of Medicine and Health Sciences* , 230-234.

Md. Maniruzzaman, M. S. (2020). Prediction of Childhood Diarrhea in Bangladesh using Machine Learning Approach. *Insights of Biomedical Research*, vol 4,111-116.

Nan-nan HUANG, H. Z.-q.-j.-b. (2021). The Short-term Effects of Temperature on Infectious Diarrhea among Children under 5 Years Old in Jiangsu, China: A Time-series Study (2015–2019). *Current Medical Science*, 211-218.

NISR. (2014-2015). *Rwanda Demographic Health Survey.* Kigali: National Institute of Statistics of Rwanda.

NISR, M. o. (2020). *Rwanda Demographic Health Survey 2019-20, Key indicators report.* Kigali.

Ruixue Li, Y. L. (2020). Diarrhea in Under Five Year-old Children in Nepal: A Spatiotemporal Analysis Based on Demographic and Health Survey Data. *International Journal of Environmental Research and Public Health*.

Rwanda, N. I. (2018). *Statistical YearBook 2017.*

Shyam Sundar Budhathok, M. B. (2016). Eco-social and behavioural determinants of diarrhoea in under-five children of Nepal: a framework analysis of the existing literature. *Tropical Medecine and Health*.

Sisay Shine, S. M. (2020). Prevalence and associated factors of diarrhea among under-five children in Debre Berhan town, Ethiopia 2018: a cross sectional study. *BMC Infectious diseases*, 20:174.

Sokhna Thiam, A. N. (2017). Prevalence of diarrhoea and risk factors among children under five years old in Mbour, Senegal: a cross-sectional study. *Infectious Diseases of Poverty*.

Starr, D. (2017). Acute Diarrhoea in children. *Patient*.

Thomas Sinmegn Mihrete, G. A. (2014). Determinants of childhood diarrhea among underfive children in Benishangul Gumuz Regional State,North West Ethiopia. *BMC Pediatrics*.

UNICEF, W. (2013). *Endeing preventable child deaths from pneumonia and diarrhea by 2025: The integrated Global Action Plan for pneumonia and diarrhea(GAPPD).* World Health Organization.

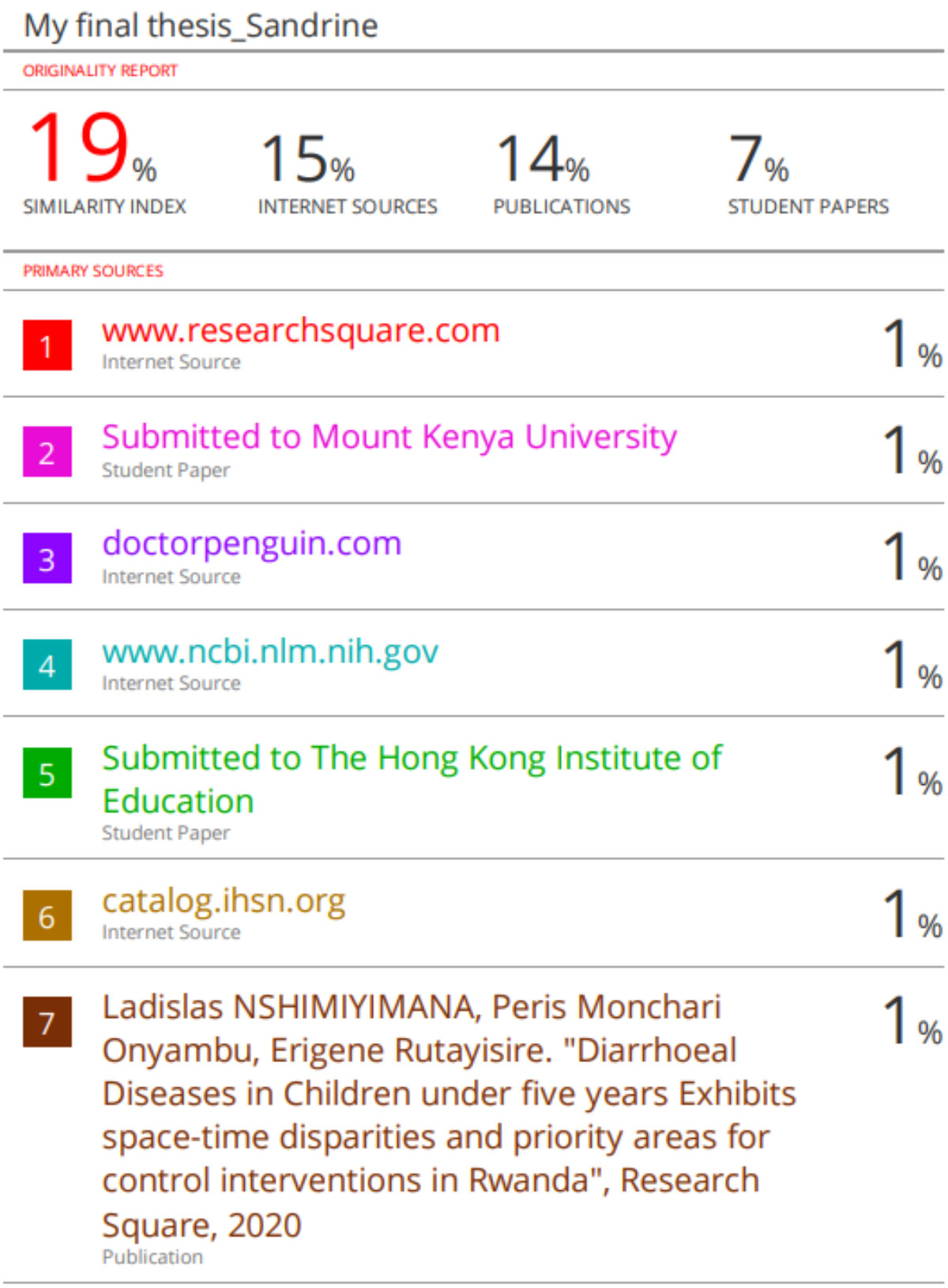
W.-P. SCHMIDT, B. G. (2009). A simulation model for diarrhoea and other common recurrent infections: a tool for exploring epidemiological methods. *Epidemiology and Infection*, 644–653.

WHO. (2017, May 2nd). *Diarrhoeal diseases.* Retrieved from World health Organization: https://www.who.int/news-room/fact-sheets/detail/diarrhoeal-disease

Wondwoson Woldu, B. D. (2016). Socioeconomic factors associated with diarrheal diseases among under-five children of the nomadic population in northeast Ethiopia. *Tropical Medicine and Health*.

Xinyu Fang, W. L. (2020). Forecasting incidence of infectious diarrhea. *BMC Infectious disease*, 20:222.

APPENDIX



My final thesis_Sandrine

ORIGINALITY REPORT

19%	15%	14%	7%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	www.researchsquare.com Internet Source	1%
2	Submitted to Mount Kenya University Student Paper	1%
3	doctorpenguin.com Internet Source	1%
4	www.ncbi.nlm.nih.gov Internet Source	1%
5	Submitted to The Hong Kong Institute of Education Student Paper	1%
6	catalog.ihsn.org Internet Source	1%
7	Ladislas NSHIMIYIMANA, Peris Monchari Onyambu, Erigene Rutayisire. "Diarrhoeal Diseases in Children under five years Exhibits space-time disparities and priority areas for control interventions in Rwanda", Research Square, 2020 Publication	1%

8	Submitted to University of Rwanda Student Paper	<1%
9	afahobckpstorageaccount.blob.core.windows.net Internet Source	<1%
10	dr.ur.ac.rw Internet Source	<1%
11	Submitted to CSU Northridge Student Paper	<1%
12	journals.plos.org Internet Source	<1%
13	Jean Bosco NDIKUBWIMANA, Frederic NGENDAHIMANA, Angelique DUKUNDE, Evariste GATABAZI, Faustin HAKIZIMANA. "Risk Factors Associated With Under-Five Diarrhea and Their Effect on Under-Five Mortality in Rwanda: Secondary Data Analysis of 2014–2015 Rwanda Demographic and Health Survey (Rdhs).", Research Square Platform LLC, 2021 Publication	<1%
14	Nan-nan Huang, Hao Zheng, Bin Li, Gao-qiang Fei, Zhen Ding, Jia-jia Wang, Xiao-bo Li. "The Short-term Effects of Temperature on Infectious Diarrhea among Children under 5 Years Old in Jiangsu, China: A Time-series	<1%

	Study (2015–2019)", Current Medical Science, 2021 Publication	
15	www.longdom.org Internet Source	<1 %
16	Submitted to University of Ghana Student Paper	<1 %
17	tropmedhealth.biomedcentral.com Internet Source	<1 %
18	cdr.lib.unc.edu Internet Source	<1 %
19	Submitted to London School of Hygiene and Tropical Medicine Student Paper	<1 %
20	www.neliti.com Internet Source	<1 %
21	atm.amegroups.com Internet Source	<1 %
22	escholarship.org Internet Source	<1 %
23	www.nejm.org Internet Source	<1 %
24	Chris Guure, Ernest Tei Maya, Samuel Dery, Baaba da-Costa Vrom, Refah M. Alotaibi, Hoda Ragab Rezk, Alfred Yawson. "Factors	<1 %

	influencing unmet need for family planning among Ghanaian married/union women: a multinomial mixed effects logistic regression modelling approach", Archives of Public Health, 2019 Publication	
25	downloads.hindawi.com Internet Source	<1 %
26	www.dovepress.com Internet Source	<1 %
27	Gulam Muhammed Al Kibria, Vanessa Burrowes, Allysha Choudhury, Atia Sharmeen, Swagata Ghosh, Arif Mahmud, Angela KC. "Determinants of early neonatal mortality in Afghanistan: an analysis of the Demographic and Health Survey 2015", Globalization and Health, 2018 Publication	<1 %
28	ur.ac.rw Internet Source	<1 %
29	www.analyticsvidhya.com Internet Source	<1 %
30	Vecino-Ortiz, A.I.. "Determinants of demand for antenatal care in Colombia", Health policy, 200805 Publication	<1 %
	bmcpediatr.biomedcentral.com	

31	Internet Source	<1 %
32	byjus.com Internet Source	<1 %
33	dspace.jaist.ac.jp Internet Source	<1 %
34	Brian A Maponga, Daniel Chirundu, Notion T Gombe, Mufuta Tshimanga, Gerald Shambira, Lucia Takundwa. "Risk factors for contracting watery diarrhoea in Kadoma City, Zimbabwe, 2011: a case control study", BMC Infectious Diseases, 2013 Publication	<1 %
35	Submitted to University of Johannsburg Student Paper	<1 %
36	Submitted to Baylor University Student Paper	<1 %
37	Submitted to Kenyatta University Student Paper	<1 %
38	statistics.caricom.org Internet Source	<1 %
39	Getachew Yismaw Workie, Temesgen Yihunie Akalu, Adhanom Gebreegziabher Baraki. "Environmental factors affecting childhood diarrheal disease among under-five children in Jamma district, South Wello zone,	<1 %

	Northeast Ethiopia", BMC Infectious Diseases, 2019 Publication	
40	Submitted to Higher Education Commission Pakistan Student Paper	<1 %
41	Submitted to Monash University Student Paper	<1 %
42	Rahul Bawankule, Abhishek Singh, Kaushalendra Kumar, Sarang Pedgaonkar. "Disposal of children's stools and its association with childhood diarrhea in India", BMC Public Health, 2017 Publication	<1 %
43	Submitted to University of Lancaster Student Paper	<1 %
44	Setegn Muche Fenta, Teshager Zerihun Nigussie. "Factors associated with childhood diarrheal in Ethiopia; a multilevel analysis", Archives of Public Health, 2021 Publication	<1 %
45	Submitted to University of Birmingham Student Paper	<1 %
46	helda.helsinki.fi Internet Source	<1 %
47	link.springer.com Internet Source	<1 %

48	Submitted to Coventry University Student Paper	<1 %
49	Hema Sekhar Reddy Rajula, Giuseppe Verlato, Mirko Manchia, Nadia Antonucci, Vassilios Fanos. "Comparison of Conventional Statistical Methods with Machine Learning in Medicine: Diagnosis, Drug Development, and Treatment", Medicina, 2020 Publication	<1 %
50	Submitted to Laureate Higher Education Group Student Paper	<1 %
51	Setegn Muche Fenta, Teshager Zerihun Nigussie. "Individual- and Community-Level Risk Factors Associated with Childhood Diarrhea in Ethiopia: A Multilevel Analysis of 2016 Ethiopia Demographic and Health Survey", International Journal of Pediatrics, 2021 Publication	<1 %
52	Submitted to South Dakota Board of Regents Student Paper	<1 %
53	Zemichael Gizaw, Wondwoson Woldu, Bikes Destaw Bitew. "Child feeding practices and diarrheal disease among children less than two years of age of the nomadic people in Hadaleala District, Afar Region, Northeast	<1 %

	Ethiopia", International Breastfeeding Journal, 2017 Publication	
54	files.eric.ed.gov Internet Source	<1 %
55	jscholarship.library.jhu.edu Internet Source	<1 %
56	open.library.ubc.ca Internet Source	<1 %
57	www.ru.ac.bd Internet Source	<1 %
58	docshare.tips Internet Source	<1 %
59	dokumen.pub Internet Source	<1 %
60	erepository.uonbi.ac.ke:8080 Internet Source	<1 %
61	www.duo.uio.no Internet Source	<1 %
62	Submitted to Stockholm University Student Paper	<1 %
63	hdl.handle.net Internet Source	<1 %
64	scholarworks.uark.edu Internet Source	<1 %

65	techscience.com Internet Source	<1 %
66	www.researchgate.net Internet Source	<1 %
67	www.ukessays.com Internet Source	<1 %
68	1library.net Internet Source	<1 %
69	Getachew Kabew Mekonnen, Bizatu Mengiste Alemu, Worku Mulat, Geremew Sahilu, Helmut Kloos. "Risk factors for acute childhood diarrhea: A cross-sectional study comparing refugee camps and host communities in Gambella Region, Ethiopia", Travel Medicine and Infectious Disease, 2019 Publication	<1 %
70	academicworks.cuny.edu Internet Source	<1 %
71	iariw.org Internet Source	<1 %
72	microdata.statistics.gov.rw Internet Source	<1 %
73	pt.scribd.com Internet Source	<1 %
	www.cours-gratuit.com	

74	Internet Source	<1 %
75	www.wcfia.harvard.edu Internet Source	<1 %
76	Bukayaw wudie, Addisu Melese, Daniel Mekonnen. "Campylobacter Jejuni and Its Antimicrobial Susceptibility Pattern Among Under- Five Children with Gastroenteritis in Northwest Ethiopia", Research Square Platform LLC, 2021 Publication	<1 %
77	Submitted to De Montfort University Student Paper	<1 %
78	Jay Saha, Pradip Chouhan. "Do malnutrition, pre-existing morbidities, and poor household environmental conditions aggravate susceptibility to Coronavirus disease (COVID-19)? A study on under-five children in India", Children and Youth Services Review, 2021 Publication	<1 %
79	Linjian Lei, Shengli Sun, Yue Zhang, Huikai Liu, Wenjun Xu. "PSIC-Net: Pixel-Wise Segmentation and Image-Wise Classification Network for Surface Defects", Machines, 2021 Publication	<1 %
80	Oliva Bazirete, Manassé Nzayirambaho, Aline Umubyeyi, Marie Chantal Uwimana, Evans	<1 %

Marilyn. "Influencing Factors for Prevention of Postpartum Hemorrhage and Early Detection of Women at Risk in The Northern Province of Rwanda: Beneficiary and Health Worker Perspectives", Research Square, 2020
Publication

---

81 Rabab B. Alkutbe, Abdulrahman Alruban, Hmidan Alturki, Anas Sattar, Hazzaa Al-Hazzaa, Gail Rees. "Fat mass prediction equations and reference ranges for Saudi Arabian Children aged 8–12 years using machine technique method", PeerJ, 2021 <1 %
Publication

---

82 Restu Windi, Ferry Efendi, Arina Qona'ah, Qorinah Estiningtyas Sakilah Adnani, Kadar Ramadhan, Wedad M. Almutairi. "Determinants of Acute Respiratory Infection Among Children Under-Five Years in Indonesia", Journal of Pediatric Nursing, 2021 <1 %
Publication

---

83 Sofia Anwar, Aisha Iftikhar, Aisha Asif, Zahira Batool. "Households Socio-Economic Determinants of Childhood Diarrhoea Morbidity in Selected South Asian Countries", Review of Economics and Development Studies, 2015 <1 %
Publication

---

84 archpublichealth.biomedcentral.com
Internet Source

		<1 %
85	dhsprogram.com Internet Source	<1 %
86	liboasis.buse.ac.zw:8080 Internet Source	<1 %
87	mrs.org Internet Source	<1 %
88	openaccesspub.org Internet Source	<1 %
89	protocolexchange.researchsquare.com Internet Source	<1 %
90	pubs.sciepub.com Internet Source	<1 %
91	scholarworks.waldenu.edu Internet Source	<1 %
92	www.childinfo.org Internet Source	<1 %
93	www.edureka.co Internet Source	<1 %
94	www.mdpi.com Internet Source	<1 %
95	www.rhsupplies.org Internet Source	<1 %

96	www.scirp.org Internet Source	<1 %
97	Eric Butera, Assumpta Mukabutera, Etienne Nsereko, Cyprien Munyanshongore et al. "Prevalence and risk factors of intestinal parasites among children under two years of age in a rural area of Rutsiro district, Rwanda – a cross-sectional study", Pan African Medical Journal, 2019 Publication	<1 %
98	Sakib Shahriar, A. R. Al-Ali, Ahmed H. Osman, Salam Dhou, Mais Nijim. "Machine Learning Approaches for EV Charging Behavior: A Review", IEEE Access, 2020 Publication	<1 %
99	"Advances in Computing Systems and Applications", Springer Science and Business Media LLC, 2021 Publication	<1 %
100	Aldo A. M. Lima, Domingos B. Oliveira, Josiane S. Quetz, Alexandre Havt et al. "Etiology and severity of diarrheal diseases in infants at the semiarid region of Brazil: A case-control study", PLOS Neglected Tropical Diseases, 2019 Publication	<1 %

101	Lei Tian, Xuhui Zhu, Zhongju Chen, Weiyong Liu, Song Li, Weiting Yu, Wenqian Zhang, Xu Xiang, Ziyong Sun. "Characteristics of bacterial pathogens associated with acute diarrhea in children under 5 years of age: a hospital-based cross-sectional study", BMC Infectious Diseases, 2016 Publication	<1 %
102	Melkamu Molla Ferede. "Socio-demographic, environmental and behavioural factors of diarrhoea among under-five children in Rural Ethiopia: further analysis of the 2016 Ethiopian demographic and health survey", Research Square, 2019 Publication	<1 %
103	Thomas Sinmegn Mihrete, Getahun Asres Alemie, Alemayehu Shimeka Teferra. "Determinants of childhood diarrhea among underfive children in Benishangul Gumuz Regional State, North West Ethiopia", BMC Pediatrics, 2014 Publication	<1 %

Exclude quotes	On	Exclude matches	Off
Exclude bibliography	On		