UNIVERSITY OF GONDAR

COLLEGE OF INFORMATICS

DEPARTMENT OF INFORMATION SCIENCE

Master of Science Degree in Data Science and Analytics (R)

Predicting diarrhea in under-five age of children using ensemble machine learning Algorithms: In the case of Ethiopia

A Thesis submitted to the Department of Information Science, College of Informatics, University of Gondar, in meeting the preliminary research thesis requirement for partial fulfillment of the award of Master of Science Degree in Data Science and Analytics.

By
Adane Abate

November, 2023

Gondar, Ethiopia

Approval sheet

UNIVERSITY OF GONDAR

COLLEGE OF INFORMATICS

DEPARTMENT OF INFORMATION SCIENCE

MASTER’S PROGRAM IN DATA SCIENCE AND ANALYTICS

Predicting diarrhea in under-five age of children using ensemble machine learning Algorithms: In the case of Ethiopia

By
Adane Abate

Name and Signature of Members of the Examining Board

Name Title Signature Date

1. Tesfahun Melese (PhD) Advisor ____________ ___/___/___/
2. __________________________External Examiner __________ ___/___/___/
3. ___________________________Internal Examiner ___________ ___/___/___/
4. ___________________________Chair Person _____________ ___/___/___/
5. ___________________________Research Coordinator_____________ ___/___/___/

Declaration

I declare that this thesis is my original work and has not been submitted for a degree in any other university's degree program.

____

Date

As a university advisor, I approved the submission of this thesis for examination.

____

Advisor

DEDICATION

I would just like to dedicate this thesis work to all of my family and close friends for their endless love, encouragement, and support, and I am truly thankful to just have a family and friends like them!!!

Acknowledgment

First and foremost, I express my gratitude to the almighty God and the Virgin Mary, mother of Jesus Christ, for providing me with the opportunity, knowledge, and inspiration to complete my thesis and achieve at this stage of my life. I am deeply indebted and very glad to express my heartfelt thanks and appreciation to my Advisor Tesfahun M. (PhD) for his important and constructive comments, criticisms, and professional advice and for doing all these on time from the beginning of the proposal writing to the completion of this thesis.

Finally, I'd like to thank all of my friends who are providing me in achieving this thesis work.

Table of Contents

List of Table

List of Figures

Lists of Acronyms

ANN – Artificial Neural Network

ARI – Acute Respiratory Infection

CatBoost – Categorical Boosting

CSA – Central Statistical Agency

EDHS – Ethiopian Demographic and Health Survey

GBM – Gradient Boosting Machine

IVAC – International Vaccine Access Center

LASSO – Least Absolute Shrinkage and Selection Operator

LDA – Linear Discriminant Analysis

LGBM – Light Gradient Boosting Machine

MDP – Markov Decision Process

QDA – Quadratic discriminant analysis

RF – Random Forest

RL – Reinforcement Learning

SBFS – Sequential Backward Selection

SDG – Sustainable Development Goals

SFFS – Sequential Forward Feature Selection

SMOTE – Synthetic Minority OverTechnique

SVM – Support Vector Machine

U5D – Under-fives Diarrhea

WHO – World Health Organization

XGBoost – Extreme Gradient Boosting

Abstract

According to the World Health Organization (WHO), diarrhea is defined as the passage of three or more loose or watery stools in 24 hours. Diarrheal disease is a major public health problem among under-five children globally. Diarrheal diseases are the leading cause of preventable death, especially among under-five children in developing countries, including Ethiopia.. In Ethiopia, diarrhea is a major public health problem and the second leading cause of clinical presentation at health facilities among children under the age of five, after pneumonia, and the prevalence is higher in rural areas than in urban areas due to a variety of factors, including feeding practices. This study focused on developing a predictive model for diarrheal disease and identifying the best important features for diarrheal disease using ensemble machine learning algorithms. To conduct this study, the researchers used secondary dataset with a total of imbalanced 38873 instances and 31 features from EDHS were used since 2000 to 2016. To balance the dataset, the researchers used SMOTE resampling strategy with binary class label. For building the predictive model a total of 64812 instances with 20 features, and a training and testing dataset split ratio of 80/20 were used.
Performance evaluation is conducted between the chosen algorithms to identify the technique with the highest accuracy. *The overall accuracy of Cat boost, extreme gradient boosting, gradient boosting and Random Forest is 85.21%, 84.81%, 83.95%, and 75.03%, respectively.* Feature importance values are generated to identify their impact on the prediction.
Based on its prediction accuracy, the researcher decided select and use cat boost algorithms for further use in the feature important analysis because it has registered better performance with *85.21%* accuracy. The most determinant risk factors of diarrhea among under five children were identified using feature importance. *Therefore*, Child's weight, Child's age, Region, bord, Types of cooking fuels and Size of child in a birth *are the main important factors.*

Keywords: - Diarrheal disease, Machine learning, Data Analytics, Ensemble Learning

CHAPTER ONE

INTRODUCTION

1.1. Background of the Study

According to the World Health Organization (WHO), diarrhea is defined as the passage of three or more loose or watery stools in 24 hours [1]. In low and middle-income countries, diarrhea is still a leading cause of death and health loss among children under the age of five[1]. Diarrhea is commonly a sign of an infection in the intestinal tract that is caused by different bacteria, viruses, and parasitic entities[2]. Diarrhea morbidity is widespread all over the world, and not only threatens human health but also greatly affects society and the economy[3]. The WHO estimates that 1.5 million children under five years die from diarrheal diseases every year, almost half of them in Africa [3]. Diarrheal diseases are the leading cause of preventable death, especially among under-five children in developing countries, including Ethiopia. Diarrhea is more prevalent in the developing world in large part due to the lack of safe drinking water, sanitation, and hygiene, as well as poorer overall health and nutritional status[4]. There are three main forms of under-five age of diarrhea, each of which can be fatal and necessitates a unique treatment regimen. In an infected person, acute watery diarrhea is associated with significant fluid loss and rapid dehydration. Bloody diarrhea is distinguished by visible blood in the stools which is linked to intestinal damage and nutrient losses in infected people. Persistent diarrhea is an episode of diarrhea with or without blood that lasts at least 14 days[1]. Diarrheal disease is a major public health problem among under-five children globally.

In Ethiopia, diarrhea is a major public health problem and the second leading cause of clinical presentation at health facilities among children under the age of five, after pneumonia, and the prevalence is higher in rural areas than in urban areas due to a variety of factors, including feeding practices[1]. However, most incidents of diarrhea have occurred in mild and acute cases and can lead to severe dehydration, which may consequences in mortality and other health risks as a result of malnutrition[5].

Different researchers used statistical methods and machine learning techniques to analyze and investigate the impact of diarrhea for under-five of age children in the world. In the statistical analysis, different researchers used different statistical approaches like descriptive statistics as well as bivariate and multivariate logistic regression, and cross-sectional study. In this analysis, predicting the future expected outcome of under-five age of child diarrhea is difficult and it needs another technique such as machine learning.

Machine learning is finding out about computational methods for improving overall performance by way of mechanizing the acquisition of knowledge from experience. Machine learning has the purpose of reaching artificial intelligence to tackle solvable issues of a practical nature [6]. Machine learning approaches have been evaluated in the context of systematic reviews of several medical problems including public health and nutrition[7]. Machine learning has been extensively applied in various application domains including medical diagnosis [8].

Identifying the contributing factors of diarrhea is very important for the effective implementation of child health programs and prioritizations[9]. So, different studies in Ethiopia showed that socioeconomic status, monthly income, number of children, methods of complementary feeding, types of water storage equipment, mother's poor hand-washing practices, lack of hand-washing facilities, duration of breastfeeding, and improper waste disposal practices are significant factors for diarrhea occurrence[9][10]. Despite the emphasis given by the Ethiopian Ministry of Health and the respective regional health offices to improve child health, many children are still dying due to easily preventable and treatable diarrheal disease in Ethiopia[9]. Many studies focused on the investigation of factors that affect the public health problem particularly the death of under-five children diarrheal disease using statistical methods and machine learning techniques.

Despite diarrhea among under-five years of age is a huge problem in sub-Saharan Africa, to the best of researchers' knowledge, no study conduct on the predictive model development and the factors associated with it in Ethiopia in one. However, research on predicting under the age of five years of child diarrhea is limited in Ethiopia particularly using machine learning techniques.

The goal of this research is to build a predictive model that can predict whether under-five children are vulnerable to diarrheal disease and also identify the significant risk factors. Conducting national-based studies will help to recognize risk factors of diarrheal diseases that enable the concerned bodies to develop appropriate interventions, which might vary depending on the environmental conditions. So, to investigate the associated risk factors and predict under five children diarrheal disease, there are machine learning techniques used for forecasting the future vulnerability of child diarrheal disease. Because statistical models are designed for inference about the relationships between variables. But machine learning models are designed to make the most accurate predictions possible.

1.2. Statement of the Problem

According to the progress report of the international vaccine access center (IVAC) in 2020, Ethiopia is one of the focus countries for pneumonia and diarrhea progress report of 2020 with 44,692 under-five pneumonia and diarrhea death[11]. Similarly, in the progress report of pneumonia and diarrhea in 2022, 45436 under-five pneumonia and diarrhea death and 13 deaths per 1,000 live births[12]. Diarrhea is a major public health problem and the second leading cause of death among children under-five years globally and accounts for almost one in five child deaths, about 1.5 million each year[13]. In 2019, around 1.3 million people died due to diarrhea, and of the total deaths about 0.54 million were children under under-five years of age worldwide. In 2016, according to the WHO, the under-five mortality rate in low-income countries was 73.1 deaths per 1000 live births, nearly 14 times the average rate in high-income countries (i.e., 5.3 deaths per 1000 live births)[14]. Worldwide, diarrheal disease contributed to 15% of all under-five deaths (approximately 2.5 million deaths each year), making diarrheal disease the second leading cause of death in the youngest members of society[14]. The WHO emphasized safe water, improved sanitation facilities, and hand washing behavior using soap to prevent diarrhea[15].

Internationally, from all causes of child deaths that occur daily, diarrheal diseases account for 15% more than 1600 children deaths under 5 years of age[16]. According to [16] in 2019, in Africa and South Asia, more than four-fifths of all under-five deaths (82%) are caused by diarrheal diseases. Based on the world health organization estimates, diarrhea contributes to more than one in every ten (13%) child deaths in Ethiopia [9]. In 2023, diarrhea accounts for 19% of total child deaths globally and affects about 1.87 million children under the age of 5 years. However, the risk of illness and death from diarrhea varies geographically and even within the same families. The prevalence of diarrheal diseases linked to inadequate water, sanitation, and hygiene is manyfold times higher in Africa than in high-income nations [17]. Diarrheal illnesses are mainly linked with poverty, and hygiene promotion programs can significantly reduce children's diarrhea when interventions are tested in stable communities [17].

In 2022, approximately 1.6 million deaths occur each year globally due to diarrhea with the highest burden occurring in developing countries and economically disadvantaged regions[18]. Globally, diarrhea contributed to 15% of all under-five deaths. Of all child deaths from diarrhea, 78% occur in the African and Southeast Asian regions. In these regions, diarrhea accounts for one in eight deaths among children younger than 5 years per annum[18].

Although the mortality from diarrhea has declined considerably over the past 25 years globally, diarrhea-associated morbidity in sub-Saharan Africa remains unacceptably high[18]. Diarrhea has been associated with reduced growth, impaired cognitive function, reduced vaccine efficacy, and disruption of physical and educational development in children[18].

Ethiopia ranks among the top five countries worldwide with a significant under-five child mortality rate. However, there has been progress in reducing this rate, with an annual reduction of 4.7%. In 2019, the country recorded an average under-five mortality rate of 51 deaths per 1000 live births. Critical factors contributing to under-five mortality in Ethiopia include diseases such as Acute Respiratory Infection (ARI), fever, and diarrhea. These factors necessitate continued efforts and interventions to address the underlying causes and improve child health outcomes in the country[19].Various studies revealed that diarrheal illness in developing countries including Ethiopia is due to many reasons[5]. For instance, in 2017, Hussein [2] aimed to identify the risk factors for the occurrence of children under five years of age in northern Nigeria. The researcher used bivariate and multivariate logistic regression. The findings revealed that maternal education, religion, age, working status, unprotected water source, main floor material, DPT3, and polio3 vaccination were found to be positively associated risk factors for children's diarrhea.

Alemu et'al [14], reviewed the prevalence and determinants of diarrhea among under-five children in Ethiopia. The findings revealed that the pooled prevalence of diarrhea among under-five children in Ethiopia was 22%. The highest prevalence was observed in the Afar region with 27%. Here, lack of maternal education, lack of availability of latrines, urban residence, and lack of maternal hand washing are significantly associated with the diarrheal of children. In 2019, Nwaoha et'al [20] aimed at identifying the most common parasites and potential risk factors for diarrhea among children under-five years of age.

K. Sadiq et'al [21] studied the risk factors for acute diarrhea in children between 0 and 23 months of age in Pakistan. They used univariate and multivariable conditional logistic regression was performed to identify diarrhea-related factors. Factors significantly associated with lower odds of diarrhea in the multivariate analysis included increasing maternal age, breastfeeding, higher paternal education, and belonging to the rich and richest quintiles.

In 2019, [22] P. Leni, aimed to examine factors at the household level that influence the incidence of diarrhea in children under five years old in Indonesia. The researcher used a cross-sectional design with the Chi-Square test and also univariate, bivariate, and multivariate analysis. The findings revealed that multivariate analysis with logic regression shows that the most dominant factors affecting the incidence of diarrhea were toilet facilities, maternal education, and residence.

Hence, studies in [2] [14][20] [21], and [22] were conducted on the risk factors for children's diarrheal disease using statistical methods. In these studies, statistical approaches were used to design the inference about the relationship between variables. However, using statistical approaches, it is difficult to predict the future expected outcome. Then, their findings lack the predictive model for predicting the under-five years of age for children's diarrhea. So, it is important to find out the solution for predicting the future expected outcome of the event. Hence, there is a machine learning technique that is used to forecasting the expected outcome using the result of a statistical approach. In the following paragraphs, the result of the recent studies using machine learning techniques has been addressed.

Uwamahoro [23] developed a predictive model of diarrhea disease among under-five children with machine learning algorithms. The researcher used bivariate analysis to identify the significant risk factors and machine learning for building a predive model. However, this researcher did not identify the significant rules which are important for policymakers to take an intervention in the prevention and control of diarrhea.

Kananura [24] developed machine learning predictive modeling for the identification of predictors of acute respiratory infection and diarrhea. But, the researcher focused on the predictive model development. Similarly, Maniruzzaman et'al [25] built the prediction of children's diarrhea in Bangladesh using machine learning algorithms. L. Ayers [26] also used machine learning approaches for assessing moderate-to-severe diarrhea in children less than 5 years of age, in rural western Kenya.

Although authors [23] [24] [25] and[26] used machine learning techniques for constructing a predictive model for diarrheal disease prediction in different countries, Uwamahoro [23] lacks significant rules, and also authors in [24] [25], and [26] did not identify the significant risk factors that cause for diarrheal disease in the study area.

Then, these gaps motivated us to use machine learning techniques in the investigation of the occurrence of diarrheal disease for under five years age of child. Hence, we aimed to fill gaps that are not addressed by the previous researchers. So, we have contributed to constructing a predictive model using machine learning algorithms and identifying the significant risk factors as well as extracting the significant rules. Therefore, we used ensemble machine-learning techniques to investigate the under five years of age child's diarrhea disease.

So, the findings of this study will have a great role in the prevention and control of diarrhea disease in children. Furthermore, the findings may also contribute to the efforts to monitor progress toward the achievement of the Sustainable Development Goals (SDGs) of 2030, to end preventable deaths of newborns and children under five years of age. No similar method has been used to study diarrhea disease in Ethiopia.

To investigate this study, the researchers formulated the following research questions:

i. Which risk factors are associated significantly with diarrhea diseases among under five age of children?

ii. Which ensemble machine learning technique is suitable to construct a predictive model for under-five age of children diarrhea?

iii. What are the best important features that contribute to predicting the occurrence of diarrhea disease among under-five Children?

iv. What are the important rules that can be generated from the predictive model to guide policy formulation toward reducing and ending child diarrhea in Ethiopia?

1.3. The objective of the study

1.3.1. The general objective of the study

The general objective of this study is to construct a predictive model and identify the significant risk factors for the under-five years of age of child diarrhea using ensemble machine-learning techniques.

1.3.2. The specific objective of the study

To achieve the main objective of this study, here the researchers listed the following specific objectives as follow.

- To identify the significant risk factors for the cause of child diarrheal disease
- To select the best appropriate ensemble machine learning algorithm to build a predictive model for under-five years of age of child diarrhea.
- To find out important features contributing to the best model for predicting diarrhea disease among under-five children in Ethiopia.
- To collect feedback from domain experts about the selected significant factors and generated rules.

1.4. Methodology

1.4.1. Description of the study area

In this study, the researchers investigated the significant risk factors and also predicted the occurrence of diarrheal disease against under-five years of age children in Ethiopia using ensemble machine learning techniques.

1.4.2. Study material

In this study, the researchers used hardware specifications like processor intel® core **i**5, CPU at 2.60GHZ, 8GB RAM, and 1TB hard disk. Similarly, software specifications such as Microsoft Office, Edraw Max, and the Jupiter Notebook.

1.4.3. Definition of variables

Diarrhea is the passage of three or more watery or loose stools per day, and when the mother is considered as increased stool frequency or liquidity [1].

Machine Learning is an AI technique that teaches computers to learn from experience.

1.4.4. Research Design

Experimental research is a scientific method of conducting research in which one or more independent variables are altered and applied to one or more dependent variables to determine their influence on the latter[27]. The use of experimental research designs ensures that the research subjects in each of the experimental conditions are equal in expectation before the administration of the experimental treatment[28]. The ability to conduct consistent, controlled, and repeatable large-scale experiments in all areas of computer science related to parallel, large-scale, or distributed computing and networking are critical to the future and development of computer science[29].

The experimental design is described in statistics as the design of an information-gathering experiment in which a variation is present or not, and it should be executed under the researcher's complete control [27]. The primary goal of experimental research is to develop research with strong causal validity. The highest levels of causal validity are offered by randomized experimental designs.

In this study, the researchers proposed an experimental research design that has a five-step experimental process such as data collection methods, data preprocessing process, model implementation, evaluation model performance, and concludes extracted knowledge from the predictive model. In this study, the experimental research design is implemented[30]. Therefore, according to our objectives, we identified the significant risk factors, built a predictive model using ensemble machine learning algorithms, and also extracted the significant rules which are important for policymakers using the best algorithms.

1.4.5. Description of the study population and sampling method

The Ethiopian Demographic and Health Survey was used as the source of the dataset which has been implemented by a central statistical agency. In this study, children under the age of five years are selected purposively according to the main objective of the study.

1.4.6. Type of data and data collection method

To conduct this study, the datasets were gathered and understood to have a general understanding of the data. In describing and verifying the data, examining issues such as the format of the data, features' types, feature values, and the quantity of the data in case of the number of rows and several columns. Here, the data source is the Ethiopian demographic and health survey. Then, the data preprocessing tasks such as data cleaning data integration, data transformation, and feature selection were applied.

1.4.7. Method of data analysis

Data analysis is a technique that typically involves multiple activities such as gathering, cleaning, and organizing the data. In data analysis methods and techniques, the researchers used the necessary data analysis technique which is useful for finding insights into data, such as metrics, facts, and figures. Then, the researchers used ensemble machine learning. Additionally, Microsoft Excel and Python programming languages are also used as data analysis tools.

The performance of the predictive model was evaluated using the confusion matrix and we derived accuracy, the area under the receiver operating characteristic curve (ROC) curve, precision, recall, and F-measure from the confusion matrix.

1.5. Significance of the Study

Based on the findings of the study, the following points were identified as significant.

- It helps the child care service provider to work in collaboration with health offices and to screen programs into routine child care services.
- It helps policymakers to formulate urgent rules and regulations to take interventions against diarrheal disease.
- It helps health offices, and NGOs working on child health programs should also target the reduction of child diarrhea.
- It helps healthcare providers to be aware of the importance of diarrhea in their practice, and to identify children at-risk-related diarrhea.

1.6. Scope (limitation and Delimitation)

The scope of this study is limited to building a predictive model and identifying the significant risk factors for the under-five age of child diarrhea in Ethiopia using ensemble machine learning algorithms.

1.7. Ethical Consideration

The researchers conducted this study, based on the consideration of the research ethical issues. And all the Ethical issues were considered in all steps of the research. Any personal information given is kept confidential and only anonymous data is used for this study.

1.8. Thesis Organization

This study report is organized into five chapters. The first chapter briefly discusses the introduction section which includes the background of the study, the statement of the problem, the research questions, and the general and specific objectives of the study, the research methodology, and the significance of the study, and lastly the scope and limitation as well as about the ethical issues. Chapter Two reviews the literature on the analysis of child diarrheal diseases and their significant risk factors. Basic concepts about machine learning techniques and their application as well as related works also included in this chapter. Chapter Three focuses on the overall model development architecture of the research conducted including what the researchers followed to understand the data, collect and analyze the data, associated factors identification, predictive model development, model evaluation, and rule generation. Chapter Four focuses on the experimental setup, results, and as well as discussion. Finally, chapter five provides the conclusion of the research and presents the recommendation for future work.

CHAPTER TWO

2. LITERATURE REVIEW

2.1. Overview

This section of the study includes an overview of children's diarrhea and methods of study for diarrhea, as an overview of the machine learning algorithms including types and their application in real-world problems. Discussion on machine learning model development processes from data collection to model evaluation. Lastly the related works.

2.2. About Diarrhea Diseases

In most countries, diarrhea is defined as three or more loose or watery stools in 24 hours. It may be acute or chronic (persistent)[3]. Diarrhea is the passage of unusually loose or watery stools, at least three times in 24 hours. Diarrhea is more prevalent in the developing world in large part due to the lack of safe drinking water, sanitation, and hygiene, as well as poorer overall health and nutritional status[4].

Diarrheal diseases account for 1 in 9 child deaths worldwide, making diarrhea the second leading cause of death among children under the age of 5. For children with HIV, diarrhea is even more deadly; the death rate for these children is 11 times higher than the rate for children without HIV[31]. Rotavirus is the leading cause of acute diarrhea and causes about 40% of hospitalizations for diarrhea in children under five. Most diarrheal germs are spread from the stool of one person to the mouth of another. These germs are usually spread through contaminated[31]. Diarrheal disease is a significant contributor to child morbidity and mortality, particularly in the developing world. Poor sanitation, unreliable supply of piped water, a lack of personal hygiene and inadequate water supplies are known risk factors for diarrheal disease[32]. Globally, there are nearly 1.7 billion cases of children's diarrheal disease, causing 525 000 deaths each year.1 Children from the developing world are disproportionately affected by diarrhea and experience it [21].

Diarrhea is responsible for the death of more than 90% of under-five children in low and lower-middle-income countries. Regionally, South Asia and sub-Saharan Africa accounted for 88% of deaths in the same age group[33]. In developing countries, it has been estimated that 1.8 million people die annually due to diarrheal diseases and more than 80% of them are children aged under five years[33]. The majority (42%) of deaths due to diarrheal disease were concentrated in Sub-Saharan Africa, including Ethiopia (88 per 1000 live births), where hygiene and sanitation are poor[4]. In Ethiopia, three fourth of the health problems of under-five children are communicable diseases that come from the environment, especially water and sanitation.

2.3. Methods of the Study for Diarrhea

Different researchers used different methods for studying diarrhea and its related issues. Here, we reviewed statistical and machine learning-related methods.

2.3.1. Statistical Approaches

The statistical analysis depends on the objective of the study which is used to do a descriptive analysis of variables[34]. In this analysis, the researchers can access the association between variables and predictive analysis based on multiple regression models and also can use software packages including SAS, SPSS, and STATA [34]. Researchers use a wide range of statistical methods to analyze survey data. They do this using statistical software packages that are designed for research professionals.

In descriptive statistics, measures of central tendency can be used to summarize the performance level of a group of scores, and measures of variability describe the spread of scores among participants [35]. In statistical techniques and analysis, different researchers investigated different domain areas. So, here, the researchers reviewed different statistical papers which focused on the application area of diarrhea as follows.

Authors in [15] investigated the risk factors of diarrhea in children under the age of five in Malawi using a simple and multiple logistic regression model. In the simple logistic regression analysis, sex, age of the child, size of child at birth, region, mother's current age and working status, wealth index, refrigerator, type of cooking fuel, main roof material, bed with mattress, time to get to a water source, type and location of toilet facilities, and hand washing facility with water and soap were associated with diarrhea of under-five at P-values lower than 0.05. The results of multiple logistic regression include sex and age of the child, size of child at birth, region, mother's current age and working status, time to get to a water source, type and location of toilet facilities, and hand washing facility with water and soap were statistically associated with diarrheal disease of under-five children in Malawi.

Researchers in [21] conducted the risk factors for acute diarrhea in children between 0 and 23 months of age in a rural district of Pakistan. Univariate and multivariable conditional logistic regression was performed to identify diarrhea-related factors. Factors significantly associated with lower odds of diarrhea in the multivariate analysis were identified. Finally, younger maternal age, higher paternal education, not breastfeeding, and poverty, which has implications for developing preventive programs and strategies that target populations with a higher risk of diarrhea.

P. Leni in [22] aimed to examine factors at the household level that influence the incidence of diarrhea in children under five years old in Indonesia. They used a cross-sectional design with a chi-square test. Chi-square test results show a relationship between the incidence of diarrheal diseases and drinking water sources p-value = 0.035, toilet facilities p-value = 0.000, maternal education p-value = 0.000, and residence p-value = 0.000. Multivariate analysis with Logic Regression shows that the most dominant factors affecting the incidence of diarrhea were toilet facilities, maternal education, and residence.

Generally, not only these but also many studies were conducted in the application area of diarrhea diseases for under five ages of children in worldwide. However, some of the recent related studies were addressed in section 2.7 below with summaries of studies including recommendations.

2.3.2. Machine Learning Approaches

Machine learning is the branch of computer science that helps computers learn without being explicitly programmed. Machines learn from past examples and historical trends and based on their previous experience a model can be built which can be used for the prediction of new values[36]. Machine learning provides better results for analysis and prediction, especially for large datasets and big data. Machine learning algorithms are used to conduct major analyses such as classification, regression, clustering, and association[37].

Machine learning techniques are used to accomplish tasks using steps such as collection and preparation of data, feature selection, algorithm selection, selection of models and parameters, training the model, and evaluating the performance of the model [38]. Machine learning is a process of making the system learn automatically based on earlier experimental data. The objective of ML is to determine the predictions based on the existing data[39].

The goal of machine learning is to master existing information. Multiple types of research have been conducted on the various ways to help machines learn on their own. Many programmers and mathematicians have addressed various methods and techniques to realize the solution to this problem[40].

2.4. Data Science Life Cycle

Phase 1: Discovery

In this phase, the researchers learn and investigate the problem, develop context and understanding, and learn and assess the data sources needed and available to support the project in terms of people, technology, tools, system, time, and data. Formulate initial hypothesis/research questions and also define objectives to solve a problem and develop an idea of the scope of the data needed, and validate that idea with the domain experts on the project included in this phase [41].

Phase 2: Data Preparation

The second phase of the data analytics lifecycle involves data preparation, which includes the steps to explore, preprocess, and condition data before modeling and analysis[41]. The data preparation phase is generally the most iterative and the one that researchers tend to underestimate most often. it can involve many complex steps to join or merge datasets or otherwise get datasets into a state that enables analysis in further phases[41].

Phase 3: Model Planning

In Phase 3, the researchers assess the structure of the datasets and identify models to apply to the data for clustering, classifying, or finding relationships in the data depending on the goal of the project [41].

Phase 4: Model Building

In phase 4, the researchers need to develop datasets for training, testing, and production purposes. These datasets enable the researcher to develop the analytical model and train the model by training data while holding test data for testing the model. In the model-building phase, an analytical model is developed and fit on the training data and evaluated against the test data[41].

Phase 5: Communicate Results

After executing the model, the researchers need to compare the outcomes of the modeling to the criteria established for success and failure. In Phase 5, the researchers consider how best to articulate the findings and outcomes to the various team members and stakeholders, taking into account caveats, assumptions, and any limitations of the results[41].

Phase 6: Operationalize

In the final phase, the team communicates the benefits of the project more broadly and sets up a pilot project to deploy the work in a controlled way before broadening the work to a full enterprise or ecosystem of users[41].

2.5. Machine Learning Algorithms

2.5.1. Machine Learning

Machine learning is a branch of artificial intelligence that aims at enabling machines to perform their jobs skillfully by using intelligent software[42]. Machine learning allows computer systems to learn directly from examples, data, and experience[43].

Supervised machine learning algorithms have challenges such as the irrelevant input feature presenting training data could give inaccurate results, data preparation and pre-processing is always a challenge, and accuracy suffers when impossible, unlikely, and incomplete values have been inputted as training data[44].

Supervised machine learning algorithms have the advantages to allows the researchers to collect data or produce data output from previous experience, helps you to optimize performance criteria using experience, helps you to solve various types of real-world computation problems, it is clarity of data, and ease of training and it can be very helpful in classification problems[44].

Besides the advantages, it has also disadvantages like it is limited in a variety of sense so that it can’t handle some of the complex tasks in machine learning, it cannot give you unknown information from the training data like unsupervised learning do, it cannot cluster or classify data by discovering its features on its own, unlike unsupervised learning, classifying big data can be a real challenge and training for supervised learning needs a lot of computation time[44].

2.5.2. Types of Machine Learning

Machine Learning algorithms are mainly divided into four categories: supervised learning, unsupervised learning, semi-supervised learning, and reinforcement learning. Though there are three largely recognized categories of machine learning depending on how the system (ML model or agent) is trained such as supervised learning, unsupervised learning, and reinforcement learning[45].

2.5.2.1. Unsupervised Machine learning

In unsupervised learning, models are not supervised using training datasets in the unsupervised learning approach. Instead, models learn from the hidden pattern and unknown information from the datasets. Three types of unsupervised learning include clustering, association, and dimensionality reduction. The clustering technique groups data points together based on their similarities. Dissimilar objects are grouped in distinct groups, whereas similar objects remain in the same group. For instance, for detecting brain tumors from MRI images, clustering algorithms can be utilized. The association also uses rules to discover relationships between features in a dataset. Dimensionality reduction is used to adequately handle complex and massive data so that fewer computing resources and storage will be used by other ML algorithms[45].

2.5.2.2. Reinforcement Learning

Reinforcement learning is an area of ML concerned with how the agents can take actions in an environment that will result in a maximum reward[45]. The problem to be solved in reinforcement learning (RL) is defined as a Markov Decision Process (MDP), i.e., all about sequentially making decisions[46]. Reinforcement learning is a type of machine learning algorithm that enables software agents and machines to automatically evaluate the optimal behavior in a particular context or environment to improve its efficiency, i.e., an environment-driven approach. This type of learning is based on reward or penalty, and its ultimate goal is to use insights obtained from environmental activists to take action to increase the reward or minimize the risk[46].

2.5.2.3. Semi-supervised Learning

Semi-supervised learning can be defined as a hybridization of supervised and unsupervised methods, as it operates on both labeled and unlabeled data[46].

2.5.2.4. Supervised Machine Learning

In supervised learning, the datasets are labeled to train algorithms to classify data or predict outcomes. Supervised learning can be further categorized into regression and classification tasks. Classification algorithms predict the class label of new data points depending on how the model is supervised by training data. Regression algorithms identify correlations between dependent and independent variables to predict the continuous value of the dependent variables[45].

Classification is a supervised learning method in machine learning, referring to a problem of predictive modeling as well, where a class label is predicted for a given example. Classification machine learning is also classified into binary classification, multiclass classification, and multilabel classification [46].

Binary Classification refers to the classification tasks having two class labels such as “true and false” or “yes and no”. Multiclass classification also the classification tasks having more than two class labels. Multi-label classification is an important consideration where an example is associated with several classes or labels. Multilevel classification includes advanced machine learning algorithms that support predicting various mutually non-exclusive classes or labels[46].

A. Decision tree

A decision tree (DT) is a well-known non-parametric supervised learning method. DT learning methods are used for both the classification and regression tasks[46]. The decision tree is similar to a tree structure where the internal nodes represent some test on an attribute, the leaf node shows the label of the class. Decision trees perform the classification of the data by sorting them in a top-down approach starting from the root node to the leaf node, where the leaf node provides the label to the class[40].

The learning algorithm for inducing a decision tree must split the training instance into smaller subsets with a recursive step of the tree-growing process and provide a method for specifying the test condition for different feature types and evaluating the goodness of each test condition. To determine how well a test condition performs, it is needed to compare the degree of impurity of the parent node (before splitting) with the degree of impurity of the child nodes (after splitting). In the decision tree, stopping conditions for tree growing also needed to terminate the tree growing process[47].

Decision tree algorithms have advantages such as being self-explanatory and comprehensible, handling both nominal and numeric input features, being rich enough to represent any discrete value classifier, capable of handling datasets that may have errors and missing values [48]. Decision Tree algorithms are effective and powerful tools that provide human-readable rules of classification [49]. Decision trees represent rules, which can be understood by humans and used in a knowledge system such as a database [50]. A decision tree is a classifier in the form of a tree structure that consists of a Root Node, Decision node, and Leaf node/terminal node. These trees classify instances or examples by starting at the root of the tree and moving through it until a leaf node [50].

The decision tree consists of a root node that has no incoming edges, internal nodes that have exactly one incoming edge and the last one is terminal nodes that have a leaf node no further split into another node [48]. The decision tree is nonparametric that interprets the decision rules easily, it incorporates a range of numeric or categorical data layers easily and there is no need to select uni-modal training data and robust concerning outliers in training data[50].

Even if the decision trees algorithm has such advantages, the decision has several drawbacks, one of which is the need to sort all numerical features to decide where to split a node which becomes costly in terms of running time and memory size, especially when decision trees are trained on large data [49].

Statistical measures

The classification process starts at the root node of the decision tree and recursively progresses until it reaches the leaf node with class labels. The split condition is applied to decide whether the input value should continue towards the left or right subtree using the notion of impurity in each node. To measure the impurity value of a split condition several indices are proposed via the Gini index, information gain, gain ratio, and misclassification rate, and the lowest impurity value is chosen [51].

Entropy

Information gain is based on Entropy. Entropy measures the extent of impurity or randomness in a dataset. The entropy of a variable is the measure of its degree of uncertainty. Entropy is defined as the sum of the probability of each label times the log probability of that same label[51]. For a dataset with one class label, will be 1 and $\log_2(p_i)$ is 0. Hence the Entropy of the homogenous data set is zero. If the entropy is higher the uncertainty/impurity is higher[51].

Information gain

Information gain is the difference between the Entropy of a class and the conditional entropy of the class and the selected feature. It measures the usefulness of a feature **f** in classification i.e., the difference in entropy from before to after the split of set **L** on a feature f. The feature with the highest information gain is the best feature to be selected for a split. Assuming that there are **V** different values for a feature **f, |Lv|** represents the subset of **L** with **f=v**, Information gained after splitting **L** on a feature f is measured as [51].

Gini index

The Gini index determines the purity of a specific class after splitting along a particular feature. The best split increases the purity of the sets resulting from the split[51].

B. Support Vector Machine

A support vector machine is a classification and regression prediction tool that uses machine learning theory to maximize predictive accuracy while automatically avoiding over-fit to the data [53]. The SVM algorithm identifies the hyper plane that separates the data items into two classes while maximizing the marginal distance for both classes and minimizing the classification errors [54]. SVM is mainly used for classification that works on the principle of margin calculation. The margins are drawn between the classes so that the distance between the margin and the classes is maximum and hence, minimizing the classification error[55].

The good performance of SVM in data classification depends on the tuning of several parameters which affect the generalization error[56]. In SVM, selecting different kernel functions is an important aspect of the SVM-based classification, commonly used kernel functions include linear, poly, RBF, and sigmoid [57].

SVM is successful when used for pattern classification problems. Applying the Support Vector approach to a particular practical problem involves resolving many questions based on the problem definition and the design involved with it[53]. The limitation of SVMs lies in the choice of kernel, speed, and size, and lack of transparency in results [58]. Since the dimensions might be very high, SVM might not be able to show the company’s score as a parametric function based on financial ratios or any other functional form. The financial ratios rate is not constant, so each financial ratio's marginal contribution is variable[59].

The primary purpose of the support vector machine is to find a hyperplane in an n-dimensional space that helps in the classification of the data points. The main goal is to search for a hyperplane having the maximum margin. Hyper plane serves as a decision boundary that is used for classifying the data points[40].

C. Logistic Regression

Logistic regression is a common probabilistic-based statistical model used to solve classification and regression issues in machine learning. Logistic regression typically uses a logistic function to estimate the probabilities. It can overfit high-dimensional datasets and works well when the dataset can be separated linearly. The assumption of linearity between the dependent and independent variables is considered a major drawback of Logistic Regression[46].

2.5.2.5. Ensemble Algorithms

A. Gradient Boosting Algorithms

Boosting is an ensemble-learning algorithm that gives different weights for training data distribution for each iteration[60]. Boosting is an ensemble technique in which new models are added to correct mistakes made by existing models. Models are added sequentially until no further refinement can be made. The ensemble technique uses the tree ensemble model which is a set of classification and regression trees [61]. The gradient is used to minimize the loss function, similar to how neural networks use gradient descent to optimize weights. Gradient boosting is a powerful machine-learning technique that achieves state-of-the-art results in a variety of practical tasks[62].

B. Categorical Boosting

Another machine learning algorithm that is efficient in predicting categorical features is the CatBoost classifier. CatBoost is an implementation of gradient boosting, which uses binary decision trees as base predictors[62].CatBoost uses boosting methodology but implements advances in the algorithm such as permutation-driven ordered boosting and categorical feature support [63]. In contrast to XGBoost, which creates trees layer by layer and subsequently prunes them, CatBoost employs entire binary trees as base predictors. CatBoost, like all gradient-based boosting techniques, has two steps for creating trees. The first step is to choose a tree structure, and the second is to set the leaf value for the fixed tree.

C. Extreme Gradient Boosting

Extreme Gradient Boosting (XGBoost) is a form of gradient boosting that takes more detailed approximations into account when determining the best model[46]. Extreme Gradient Boosting is started from a combination between gradient descent and boosting, called Gradient Boosting Machine (GBM). XGBoost is one of the boosted tree algorithms, which follows the principle of gradient boosting. When compared with other gradient-boosting algorithms, XGBoost makes use of a more regularized model formalization in other to control the over fitting of data, which gives it better performance[64]. XGBoost is a tree ensemble approach in which the trees are added sequentially, and each tree learns from its predecessors, where they aim to minimize the errors of the previous tree. The trees are provided in parallel tree boosting to solve tasks quickly and accurately. XGBoost is created to thrust the extreme computational limits of machines to provide a scalable and efficient library[65].

The XGBoost model integrates several weak learners (decision trees) to develop a strong learner through additive learning[66]. XGBoost follows the gradient-boosting principle by improving the training process and preventing overstating. To this end, XGBoost implements second-order derivatives to minimize the loss function and obtain more accurate trees[67].

D. Light Gradient Boosting

It is a distributed and fast, scalable gradient boosting architecture supported decision tree algorithm in Data Science. LGBM algorithm splits the tree in the order leaf-wise with the simplest fit, other boosting algorithms splits the tree in the order level-wise or depth-wise. a leaf-wise algorithm can reduce loss more than a level-wise algorithm[68].

E. Random Forest

A random forest classifier is well-known as an ensemble classification technique that is used in the field of machine learning and data science in various application areas[46]. The random forest (RF) algorithm is a well-known tree-based ensemble learning method and the bagging-type ensemble. RF differs from other standard trees each node is split using the best among a subset of predictors randomly chosen at that node[64].

Random forests use a large number of unpruned decision trees, which are created by randomizing the split at each node of the decision tree. In a random forest, the features are chosen randomly and the best split among those features is chosen[69]. From a computational standpoint, random forests are appealing because they naturally handle both regression and classification; are relatively fast to train and to predict; depend only on one or two tuning parameters; have a built-in estimate of generalization error; can be used directly for high-dimensional problems; can easily be implemented in parallel[70]. While for statistically, random forest is appealing due to the measure of variable importance, differential class weighting, missing value imputations, and visualization[70].

Random Forests have three main features that gained focus such as accurate prediction results for a variety of applications, the importance of each feature can be measured through model training and the trained model can measure the pairwise proximity between the samples[71]. The random forest is appropriate for high-dimensional data modeling because it can handle missing values and can handle continuous, categorical, and binary data. Besides high prediction accuracy, the random forest is efficient, interpretable, and non-parametric for various types of datasets[72].

The model interpretability and prediction accuracy provided by random forest are very unique among popular machine learning methods. Accurate predictions and better generalizations are achieved due to the utilization of ensemble strategies and random sampling[71]. Random forest is important for many types of applications to provide accurate predictions. It can measure the importance of each feature concerning the training data set [50]. Random forests overcome the problem of over fitting in training data and it reduces the variance and improves the accuracy, they are less sensitive to outlier data, parameters can be set easily and so, this eliminates the need for pruning the trees, identifying the most important features from the training dataset and accuracy are generated automatically, handle the missing values, used for both classification and regression task [71]. Similarly, it has drawbacks as random forests are biased to features with more levels for which the data includes categorical features for the different number of levels. If the data contain groups of correlated features of similar relevance for the output, then smaller groups are favored over larger groups [50].

2.5.3. Machine Learning Model Development Process

2.5.3.1. Data Collection and Understanding

The first phase of any machine learning project is developing an understanding of the business requirements. Identify the data needs and determine whether the data is in proper shape for the machine learning project. The focus should be on data identification, initial collection, requirements, quality identification, insights, and potentially interesting aspects that are worth further investigation.

2.5.3.2. Data Preparation

Once we have appropriately identified the data, we need to shape that data so it can be used to train our model. The focus is on data-centric activities necessary to construct the data set to be used for modeling operations. Data preparation tasks include data collection, cleansing, aggregation, augmentation, labeling, normalization, and transformation as well as any other activities for structured, unstructured, and semi-structured data. Data preparation comprises those techniques concerned with analyzing raw data to yield quality data, mainly including data collecting, data integration, data transformation, data cleaning, data reduction, and data discretization[73][74].

2.5.3.3. Data Integration

Data integration has been extensively studied by the data management community and is a core task in the data pre-processing step of ML pipelines. When the integrated data is used for analysis and model training, responsible data science requires addressing concerns about data quality and bias[75].

2.5.3.4. Data Transformation

Data transformation is the process of converting data from one format to another, typically from the format of a source system into the required format of a destination system. Data transformation is a component of most data integration and data management tasks, such as data wrangling and data warehousing.

2.5.3.5. Feature Selection

Feature selection is the process of identifying and removing irrelevant and redundant features from the training data so that the learning algorithm focuses only on those aspects of the training data useful for analysis and future prediction[76]. Three main models deal with feature selection such as filters, wrappers, and embedded methods[77]. The objective of using feature selection is to improve classification performance in case of speed, learning, and accuracy, and a better understanding of the underlying process that generated the data[78][79].

2.5.3.6. Algorithm Selection

The algorithm selection problem has attracted a great deal of attention, as it endeavors to select and apply the best algorithms for a given task[80]. The algorithm selection problem can be cast as a learning problem: the aim is to learn a model that captures the relationship between the properties of the datasets, or meta-data, and the algorithms, in particular their performance[80].

2.5.3.7. Model Development

In ML model development, the final approach will be the best option when large datasets are available and ultimate performance is the major metric of interest in machine learning model construction[81].

2.5.3.8. Model Evaluation

A machine learning model is validated by evaluating its prediction performance. Evaluation metrics allow for quantifying the performance of the machine learning model[82]. When evaluating machine learning models, researchers need to decide which metrics are important for the business problem they are trying to solve[83]. The regression and classification tasks have different evaluation metrics since the outputs produced by the models are different. Mostly the evaluation metrics are driven by the confusion matrix. In the confusion matrix, a 2x2 table readily summarizes all four possible decisions: true positive (Tp), false negative (Fn), false positive (Fp), and true negative (Tn). The Tp and Tn decisions are correct while the Fn and Fp decisions are incorrect[84].

Table 2. 1 Confusion matrix for the both actual and predicted class

		Predicted Class	
		Negative	Positive
Actual Class	Negative	Tn	Fp
	Positive	Fn	TP

Tp is a true positive that is an outcome where the model *correctly* predicts the positive class. Tn is a true negative that is an outcome where the model *correctly* predicts the negative class. Fp is a false positive that is an outcome where the model *incorrectly* predicts the positive class. Fn is a false negative that is an outcome where the model *incorrectly* predicts the negative class.

Sensitivity /Recall is defined as the ratio of correctly classified majority class values (true positives) divided by the sum of correctly classified majority class values (true positives) and incorrectly classified minority class values (false positives). it should be high[51]. Sensitivity reflects the ability of the observer to correctly classify the target and is calculated as

$$\textit{\textbf{Sensitivity /recall}} = \frac{\boldsymbol{Tp}}{\boldsymbol{Tp+Fn}} = \frac{\boldsymbol{Tp}}{\boldsymbol{p}} \quad ------\text{Equation 2. 1.} Sensitivity/recall$$

***Specificity** reflects the ability of the observer to correctly classify the target and is calculated as* $\boldsymbol{specificity} = \frac{\boldsymbol{Tn}}{\boldsymbol{Fp+Tn}} ----- \;\; ----$ *Equation 2. 2. Specificity*

Precision is defined as the ratio of correctly classified majority class values (true positives) divided by the sum of correctly classified majority class values (true positives) and incorrectly classified majority class values (false positives). it should be high[51].

Precision is defined as the ratio of correctly classified majority class values (true positives) divided by the sum of correctly classified majority class values (true positives) and incorrectly classified majority class values (false positives).

$$\boldsymbol{Precision} = \frac{\boldsymbol{Tp}}{\boldsymbol{Tp+Fp}} \quad ------------ \textit{Equation 2. 3 precision}$$

The **false-positive rate** (FP) is the proportion of negatives cases that were incorrectly classified as positive, as calculated using the equation:

$$\boldsymbol{FP} = \frac{\boldsymbol{Fp}}{\boldsymbol{Tn+Fp}} \quad ---------\text{Equation 2. 4. False-positive rate}$$

The true negative rate (Tn) is defined as the proportion of negatives cases that were classified correctly, as calculated using the equation:

$$\boldsymbol{Tn} = \frac{\boldsymbol{Tn}}{\boldsymbol{Tn+Fp}} \quad ----------------\text{Equation 2. 5. True negative rate}$$

Accuracy is the number of correctly predicted data points out of all the data points. The accuracy is the proportion of the total number of correct predictions. It is determined using the equation:

$$\textbf{Accuracy} = \frac{\textbf{Tp+Tn}}{\textbf{Tp+Tn+Fn+Fp}} \quad --------------\text{Equation 2. 6 Accuracy}$$

Model accuracy in machine learning is the measurement used to determine which model is best at identifying relationships and patterns between features in a dataset based on the input. The better a model can generalize to ‘unseen’ data, the better predictions and insights it can produce, which in turn deliver more business value[85].

Receiver operating characteristics (ROC) curve

ROC analysis and the resulting ROC curve is a method that captures the relationship between sensitivity and specificity as well as the range of decision thresholds that every observer has no matter what their level of expertise and experience[84]. For evaluating the predictive model using ROC metrics, the true positive rate and false positive rate have a great impact on it. Then, to expect a better predictive model, the true positive rate should be increased and the false positive rate should be decreased, so these metrics indicated that increasing the true positive rate and decreasing the false positive rate.

Cross-validation is a data resampling method to assess the generalization ability of predictive models and to prevent overfitting and used to estimate the true prediction error of models and to tune model parameters[86]. The cross-validation estimate of a given performance metric is simply the mean of the performance in each fold of the cross-validation procedure [87].

To minimize the over fitting and under fitting of the performance, cross-validation plays a great role. In this case, the researchers used stratified 10-fold cross-validation was applied. This indicated that the total dataset is split into ten separate phases and if the nine phases are used for training the predictive model, the remaining phase is also used for testing the model and iterating it.

2.6. Application of Machine Learning

A major application field of machine learning is intelligent decision-making by data-driven predictive analytics. The basis of predictive analytics is capturing and exploiting relationships between explanatory variables and predicted variables from previous events to predict the unknown outcome[46].

Machine Learning furnishes the ability of insights on automatically recognizing patterns and determining the prediction models for the structured and unstructured data even in the absence of explicit programming instructions[39]. Machine learning approaches have been evaluated in the context of systematic reviews of several medical problems including drug class efficacy assessment, genetic associations, public health, cost-effectiveness analyses, toxicology, treatment effectiveness, and nutrition[7]. Machine learning has been extensively applied in various application domains including medical diagnosis, credit risk analysis, customer profiling, market segmentation, targeted marketing, retail management, and fraud detection[8].

The applications of machine learning cannot be underestimated in the current scenario, all sectors have applied the concept of machine learning and artificial intelligence[88]. Machine learning has been used in different application areas in information retrieval[89], court decision prediction [90], student performance prediction[91], in covid-19 disease prediction[92], and insights on machine learning and applications in the sectors like medical, banking, IoT, and another day to day uses[88]. Applying ML in the areas of disease diagnosis, treatment, patient monitoring, drug discovery, and epidemiology, among others, allows for foreseeing the potential and impact of ML in the design and implementation of new and better solutions in the mentioned areas[92].

Machine learning allows for better consideration of a wider array of potential risk factors by employing algorithms to parse data, running through multiple iterations of variables in the dataset to learn the optimum model for explaining outcomes of focus[93].

2.7. Related Works

In this study, we reviewed related papers which were conducted using statistical approaches and machine-learning techniques about child diarrheal diseases. So, here we put the recent papers.

J. Moon et'al [15] explored the risk factors for diarrhea in children under five in Malawi. The study utilized variables with a P-value below 0.05 in the simple logistic analysis, which were then included in the multiple logistic regression model. The findings revealed that demographic characteristics such as the child's sex and age, the child's size at birth, region, mother's age, and working status were significantly associated with the risk of diarrhea. However, a key gap in this research is the absence of a predictive model to accurately forecast the occurrence of diarrhea in children.

K. Sadiq *et'al* [21] studied the risk factors for acute diarrhea in children between 0 and 23 months of age. Univariate and multivariable conditional logistic regression was performed to identify diarrhea-related factors. Factors significantly associated with lower odds of diarrhea in the multivariate analysis included increasing maternal age, breastfeeding, higher paternal education, and belonging to the rich and richest quintiles. Even if they aimed to identify risk factors, they lack the predictive model and the significant rules.

P. Leni [22] aimed to examine factors at the household level that influence the incidence of diarrhea in children under five years. Multivariate analysis with logic regression was used as an approach and the most dominant factors affecting the incidence of diarrhea were toilet facilities, maternal education, and residence.

T. Kapwata et'al [32] studied diarrheal disease concerning possible household risk factors in South African villages. They used adjusted odds ratio (AOR) and confidence interval for identifying the associated risk factors. After adjusting for confounders, the occurrence of diarrhea was statistically significantly associated with sourcing water from an indoor and storing cooked food in non-refrigerated conditions. These researchers lack predictive model development and significance rules.

Ayuk *et'al i*n [94], aimed to determine the prevalence of diarrhea and its associated determinants after community vaccination. They used multivariate analysis and identified the factors using confidence intervals. Multivariable analysis showed that mothers'/caregivers' age, level of education, sex, child's age (≤ 12 months), mother's/caregiver's knowledge of diarrhea, toilet facility, source of drinking water for the child, time trekked to fetch drinking water and Rota vaccine were independently associated with diarrhea.

H. Hussein et'al [2] aimed to identify the risk factors for the occurrence of childhood diarrhea among children aged between 0-5 years in northern Nigeria regions. Bivariate and multivariate logistic regression was computed to assess independent factors of children's diarrhea. The results of this study showed that maternal education, religion, age, working status, unprotected water source, main floor material, DPT3, and polio3 vaccination were found to be positively associated risk factors for children's diarrhea after adjusting for other variables. However, the predictive model and the significant rules were not considered in their findings.

F. Nwaoha et'al [20] studied the prevalence of diarrhoea, and associated risk factors, in children aged 0-5 years. The researchers used confidence intervals to identify the associated risk factors for diarrhea diseases. This paper also lacks predictive model development and significant rules.

D. Mosisa,et'al [95] studied the determinants of diarrheal diseases among under-five children. Descriptive, bivariate, and multivariate binary logistic regression analyses were done by using SPSS. Sociodemographic determinants such as being a child of 12–23 months of age and mothers' history of diarrheal diseases were significantly associated with diarrheal diseases among under-five children. Environmental and behavioral factors such as lack of a hand-washing facility near a latrine a lack of hand-washing practice at critical times, improper domestic solid waste disposal, and not being vaccinated against rotavirus were found important determinants of diarrheal diseases among under-five children.

C.Town et'al [96] investigated diarrhoea among children aged under five years and risk factors in informal settlements. A cross-sectional study was conducted in 707 households in six informal settlements(IS) and two formal settlements (FS). Most are households used public taps (74.4%) and shared toilets (93.0%), while FS households used piped water on premises (89.6%) and private toilets (98.3%). The overall under_five diarrhea (U5D) prevalence was 15.3% and was higher in FS. Water treatment, good hand-washing practices, and Hepatitis A vaccination had significant preventing effects on U5D.

K. Ghosh et'al [97] studied the prevalence of diarrhea among under-five children in India and its contextual determinants. They applied spatial analysis software i.e. ArcGIS 10.8 and GeoDa 1.18 including Moran's index and logistics regression to show the spatial prevalence and auto-correlation of diarrhoea among neighbourhood districts and their contextual determinants. The findings revealed that the analysis of socio-economic determinants show the prevalence of diarrhea among under 5 children is higher in rural areas, among children not staying in Pacca house, living with unimproved sanitation facilities, belonging to an underprivileged community (highest among OBC followed by SC/STs), children of younger mothers (<25 years) and "poor' households considering the wealth index.

Bekele and Merdassa [98] conducted a study to determine determinants of diarrhea among under-five-year-old children in the health extension model and non-model families. The community-based comparative cross-sectional study was conducted and a multi-stage sampling technique was used. Bivariate analysis was done to select candidate variables at $p \leq 0.2$. Determinants of children's diarrhea were determined by a multivariable logistic regression model at a p-value less than 0.05. The findings revealed that the two-week prevalence of diarrhea among under-five children in model and non-model families was 7.8% (95% CI=4.5-11.1%) and 27.8% (95% CI 22.3-33.3%), respectively.

T. Solomon et'al [99] studied diarrheal morbidity and predisposing factors among children under 5 years of age in rural East Ethiopia using logistic regression. The identified risk factors were maternal diarrhea, handwashing after contact with child feces, the use of a dipper to draw water from containers, and the presence of a refuse disposal facility.

H. Lanyero *et'al* [100] aimed to determine the prevalence of antibiotic use in managing diarrhea in children under 5 years of age in rural communities using a cross-sectional approach. The findings revealed that the determinants of antibiotics included children living in peri-urban settings (AOR: 3.41, CI: 1.65–7.08, P = 0.001), getting treatment from a health facility (AOR: 1.76, CI: 1.06–2.93, P = 0.029), and having diarrhea with ARIs (AOR: 3.09, CI: 1.49–6.42, P =0.003).

N. Gessesse and A. Tarekegn [101] studied the Prevalence and associated factors of diarrhea among under-five children using a comparative cross-sectional study. They used binary logistic regression for assessing the independent factors associated with the dependent variable also the adjusted odds ratio at a confidence level of 95% and a p-value of less than or equal to 0.05 were used for determining the association factors. Their findings revealed that Shallow water and maternal diarrhea were determinants of childhood diarrhea. Place of birth and maternal diarrhea in non-model households.

M. Hartman *et'al* [102] conducted to identify risk factors for mortality of under-five the age of the child using multivariable logistic regression. Risk factors for death in the multivariable analysis included younger age (for <6 months compared with older ages, female sex, presenting with persistent diarrhea, no vomiting, severe dehydration, and being negative for rotavirus on an enzyme-linked immunosorbent assay test.

Generally, in those aforementioned studies, researchers used statistical approaches to conduct the risk factors and related issues of under-five children's diarrhea diseases. However, they lack the predictive model and significant rules. Besides this statistical approach, there is another approach for conducting children's diarrhea diseases which is called the machine learning approach. So, here, we also reviewed the recent and related work which are studied using machine learning approaches as follows.

Uwamahoro [23] developed a predictive model of diarrhea disease among under-five children with machine learning algorithms in Rwanda. With bivariate analysis, residence, age group, wealth index, type of toilet facility, main material floor, duration of breastfeeding, rotavirus vaccine, and maternal education are associated with children's 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, households with earth and sand as the 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. However, this researcher did not identify the significant rules which are important for policymakers.

Kananura [24] developed machine learning predictive modeling for the identification of predictors of acute respiratory infection and diarrhea in Uganda’s rural and urban settings. The predictors were grouped into four categories: child characteristics, maternal characteristics, household characteristics, and immunization. The results highlight the need for a holistic approach with multisectoral emphasis in addressing the occurrence of ARI and diarrhea among children. The findings revealed that the gradient-boosted algorithm was the best-selected model for the identification of the predictors of ARI (Accuracy: 100% -rural and 100%-urban) and diarrhea (Accuracy: 70%-rural and 100%-urban).

Maniruzzaman et’al [25] built the prediction of children's diarrhea in Bangladesh using a machine learning approach. Logistic regression (LR) is used to determine the high-risk factors of diarrhea. Then four ML-based approach namely naïve Bayes (NB), linear discriminant analysis (LDA), quadratic discriminant analysis (QDA), and support vector machine (SVM) was applied to predict the child's diarrhea status, and accuracy, sensitivity, and specificity are used to evaluate the performance of these classifiers. The findings indicated that SVM with radial basis kernel yielded 65.61% accuracy, 66.27% sensitivity, and 52.28% specificity which are comparatively better than others.

L. Ayers [26] used machine learning approaches for assessing moderate-to-severe diarrhea in children less than 5 years of age, in rural western Kenya. The researcher aimed to examine machine learning statistical methods to address weaknesses associated with using traditional logistic regression models. Least Absolute Shrinkage and Selection Operator (LASSO), use of classification trees, and random forest. A principal finding in all three studies was that machine learning methodological approaches are useful and feasible to implement in epidemiological studies. The results from all three machine learning approaches were supported by comparable logistic regression results indicating their usefulness as epidemiological tools.

Generally, in the case of Ethiopia, predicting diarrhea in children under five years old using ensemble machine-learning algorithms is crucial. Several machine learning studies conducted in different countries, such as Rwanda, Uganda, and Bangladesh, have developed predictive models using various machine learning algorithms. However, due to cultural differences, economic, religious, and social characteristics of countries, these models may not be applicable in Ethiopia. Therefore, further research is needed to develop and evaluate an ensemble machine learning-based approach that considers the unique characteristics of Ethiopia to predict diarrhea status among under-five children. This approach can provide more accurate predictions and significant rules that can aid policymakers and health practitioners in addressing this critical public health issue.

Table 2. 2 Summary of Literature Review

Authors (year)	Title	Methodologies	Finding	Gaps
J. Moon, et al (2019)	Explored the risk factors for diarrhea in children under five in Malawi	Using multiple logistic regression model	Child's sex and age, the child's size at birth, region, mother's age, and working status were significantly associated with the risk of diarrhea.	The researcher has selected the risk factors without considering their importance with diarrhea and they have not selected factors that may cause diarrhea

K. Sadiq et al.(2022)	Assessed the risk factors for acute diarrhea in children between 0 and 23 months of age in Pakistan	Univariate and multivariable logistic regression	increasing maternal age, breastfeeding, higher paternal education are the main factors of diarrhea in children U 5		The researchers used small scale data set with limited method of discovering factors that cause to diarrhea, and they have used statistical tools only, so did not discover hidden patterns of the data.
H. Hussen et al.(2017)	Identify the risk factors for the occurrence of childhood diarrhea among children aged 0-5 years in northern Nigeria.	Bivariate and multivariate logistic regression	Maternal education, religion, age, working status, unprotected water source, main floor material, DPT3 and polio3 vaccination were risk factors for childhood diarrhea.		However, the predictive model and the significant rules were not considered in their findings.
Bekele et al. (2021)	Determine determinants of diarrhea among U5 -year-old children in the health extension model and non-model families	Bivariate analysis and multivariable logistic regression	Water source, place of childbirth, child vaccination against Rotavirus, and vitamin A supplementation were independently associated with the occurrence of diarrhea in under-five children		factors of diarrhea are not enough and it needs adding more risk factors to improve predictive performance
Kananura et al(2021)	Predictive modelling for identification of predictors of ARI and diarrhea in Uganda.	Using decision tree and random forest and logistic regression	GBM	Accuracy: 94%	The model performance is poor since they have been used without calibrating the hyper-parameter and the ml algorithm selection is not good enough to fit the data.
			Lasso	Accuracy: 75%	
			LR	Accuracy : 75%	
Uwamahoro et al. (2021)	Prediction of diarrhea disease among under-five children with machine learning algorithms in Rwanda	RF, LR, NB, SVM, NN, GB	RF	83.07%	The study lacks generalization of machine learning algorithms in predicting diseases like diarrhea, and have not put a clue for policy makers how to use the models
			LR	62%	
			NB	60.72 %	
			SVM	71.13%	
			ANN	76.2%	
			GB	86.45%	
Maniruzzaman (2020)	Prediction of children's diarrhea in Bangladesh using a machine learning approach.	LDA, NB, SVM	NB	50.85%	They have used only single classifiers and they have not seen the other machine learning algorithms and their model performance is not good for classifying either a child has diarrhea or not
			LDA	51.8%	
			SVM	65.61% The researcher conclude that SVM is the best classifier for	

				predicting childhood diarrhea	

Summary

Based on a thorough review of statistical and machine learning studies on diarrhea in under five childeren in Ethiopia, it is evident that there is a significant need for accurate prediction models for under-five age children in Ethiopia. This is especially important given that diarrhea is a major cause of morbidity and mortality in this age group. By developing a predictive model, we can identify important risk factors for diarrhea and generate important rules that aid policymakers, health practitioners, and government bodies in giving much-needed attention to this issue. Moreover, the development of such a model can contribute to fulfilling the Ethiopian sustainable goals of reducing the child mortality rate by 2030 as outlined by the United Nations World Organization. Previous research has utilized various machine learning algorithms such as decision trees , logistic regression, Naïve bayes and artificial neural network. However, an ensemble approach could potentially improve accuracy and reduce bias. By combining multiple algorithms, the model can better handle complex and diverse datasets, leading to more reliable predictions. Furthermore, incorporating additional factors recommended by the World Health Organization can improve the algorithm's performance.

Therefore, developing an ensemble machine learning algorithm for predicting diarrhea in under-five age children in Ethiopia has the potential to greatly improve public health outcomes in the region. It can aid in the early detection and treatment of diarrhea by identifying important risk factors and generating significant rules. This study can contribute to the existing body of knowledge on predicting diarrhea in under-five age children in Ethiopia and provide novel insights into improving public health outcomes in the region.

CHAPTER THREE

3. RESEARCH METHOLODGY

3.1 Proposed Model Architecture

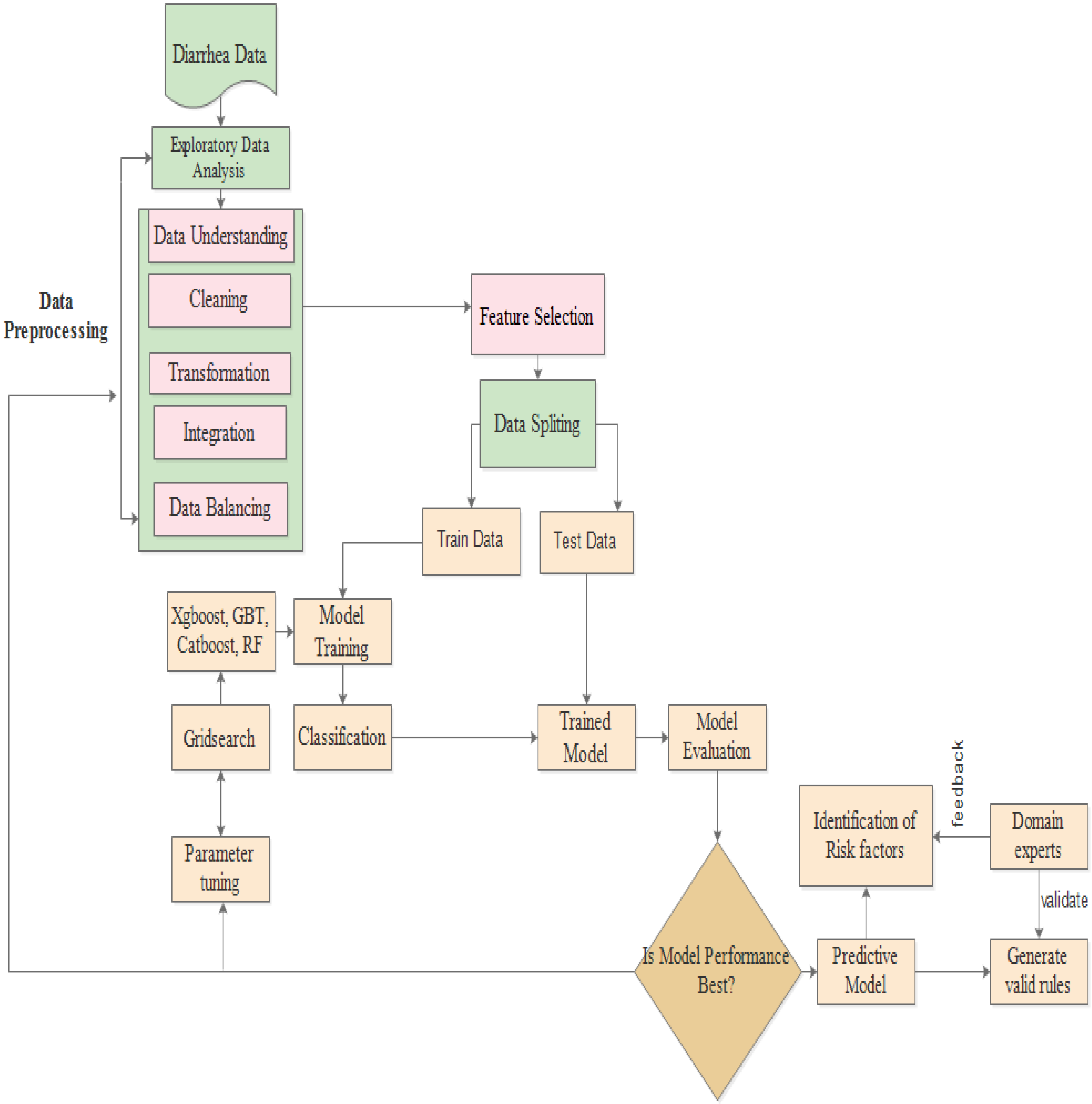


Figure 3. 1 Proposed model architecture for diarrhea against under five age predictive model

3.1.1 EDHS Diarrhea dataset

We obtained the data from the Ethiopian Demographic and Health Survey (EDHS) conducted from 2000 to 2016. The EDHS data was collected by the Ethiopian Central Statistical Agency (CSA) using a multistage sampling design. The dataset includes information on various aspects, such as dwelling characteristics, demographic details of household members, reproductive history of women, and child health data and also vaccination coverage information was obtained.

3.1.2 Data Preprocessing

In order to gain insights and effectively utilize the data, it is crucial to study its associations, patterns, and abnormalities. This involves data preprocessing, which is the initial step towards preparing the data for modeling purposes [127]. To develop an accurate prediction model while saving time and optimizing memory usage, the researchers need to explore different data preprocessing steps. These steps include data cleaning, data transformation and feature selection.

3.1.2.1 Data Understanding

The data understanding phase plays a crucial role in comprehending the dataset's features, their applicability to the research, and creating a target dataset with relevant variables for the discovery process. It is essential to understand the existing data as real-world data is often unclean and unsuitable for direct application in machine learning processes. By utilizing the visualization techniques, the researchers gained the deeper understanding of the datasets in this study, enabling us to describe and select the appropriate features for the target dataset.

3.1.2.2 Data Cleaning

Data cleaning is the process of identifying and correcting inaccurate records from a dataset [103]. Handling data that contain missing, redundant, and inconsistent values is crucial for the data cleaning process and data wrangling in general.

3.1.2.3 Data Integration

In this study, the researchers integrated four EDHS datasets from 2000 to 2016.

3.1.2.4 Data Transformation

Data transformation is the process of converting continuous data into discrete data, which is useful for better data representation, data volume reduction, better data visualization, and expressing data at various levels of granularity for data analysis.

3.1.2.5 Data Balancing

Due to the presence of a skewed distribution of data among the classes in data space and one class is a higher number than the other is commonly known as imbalanced data[104]. When dealing with imbalanced datasets, classification models tend to favor the majority classes, leading to misclassification of minority class instances and subsequently poor predictive accuracy. To address this issue, various re-sampling techniques, such as over-sampling and under-sampling, can be employed at the data level. In this study, we have utilized SMOTE over sampling methods.

3.1.3 Feature Selection

The researchers employed a different feature selection method, as a dimensionality reduction technique[105] aims to choose a small subset of the relevant features from the original ones by removing irrelevant, redundant, or noisy features using the wrapper method. It will give us the features that are most influential on the target variable which is diarrhea. The feature selection usually leads to improved learning performance, such as increased learning accuracy, reduced computational cost, and improved model interpretability.

3.1.4 Train and Test splitting

Once the data is well-prepared and the output class is balanced to mitigate issues like over fitting or under fitting, the dataset is divided into two parts: training and testing sets. The training set is utilized to train the selected machine learning model and enable it to learn the dataset's properties accurately[106]. We also used the test dataset for evaluating the performance of the developed model.

3.1.5 Model evaluation

After developing the predictive model using a training set, a model’s performance is evaluated, which is an important part of any data science research. The goal of model evaluation is to estimate a model's generalization accuracy on future data[107]. The model's performance is assessed using performance evaluation metrics namely confusion matrix, accuracy, ROC, cross-validation, precision, recall, and confusion matrix, and also using domain experts’ feedback. Model selection is the operation for finding the best model for a given set of data.

3.1.6 Hyperparameter tuning

Hyperparameter tuning[108] is one of the basic methods in the machine learning algorithm. ML algorithms require user-defined inputs to make a balance of correctness and generalization. Although there are many Hyperparameter optimization/tuning algorithms available now, the researchers used grid search that will build and evaluate a model carefully for each combined effect of Hyperparameter specified in a grid based on the data set nature and size.

3.1.7 Risk Factor Identification

Identifying risk factors is used to determine and understand the basic factors that improve the ML models that can efficiently identify the socio-demographic risk factors of diarrhea [109]. To identify determinant risk factors, the researchers used the feature_importances package. The importance of all the features used in building predictive models is calculated using a feature importance experiment conducted.

3.1.8 Generate Rules

Using the developed predictive model, rules were generated which is important to describe the underlying model's hypothesis. Rule generation algorithm is useful for experts to verify and confirm the importance of the developed predictive model in the area of diarrhea disease predication.

3.2 Model Development

To answer our research questions which are defined in section 1.2, we conducted successive empirical experiments. In this section, the researchers described the characteristics of the dataset, basic preprocessing techniques used for this study, and the evaluation techniques which have been applied to evaluate the performance of the predictive model.

3.2.1 Data Source and Descriptions

This study was conducted using secondary dataset from 2000 to 2016 collected in the five-year interval by EDHS, the most recent nationally representative dataset on under five child diarrhea disease. Until this survey, Ethiopia had nine regional states and two administrative cities. The survey was a population-based cross-sectional study designed to provide population, human rights, and health indicator estimates at national and regional levels, as well as urban and rural residents. The EDHS data were collected using five questionnaires (household, women, child immunization and health facility). We only included under five children who also had complete information about diarrhea. The original dataset contains 31 columns including the class label and 38873 records with a total size of 4.71 MB.

3.2.2 Dataset Description

In Table 3.1 we have described the detailed description of features, data types, and feature values of socio demographic and economic factors of both wife and her partner/husband.

Table 3. 1 List of features and their description

No.	Symbol	Feature Descriptions	Data Types	Values
1.	V012	Current age of mother	Contioniuse	15-49
2.	V024	Region	Nominal	1,2,3,..
3.	v025	Place of residents	Nominal	Urban, rural
4.	v113	Source of drinikng water	Nominal	Piped water,tube well, protected, uprotected, other
5.	v116	Types of toilet Faciliies	Nominal	No facility,Flush related ,Ventliated Improved, other
6.	v130	Religion	Nominal	Orthodox,catholi, muslim, protestnt, other
7.	v136	Family size	Discrtes	1-14
8.	v152	Age of household head	Contionuse	15-49
9.	v502	Martial status	Nominal	Never marreid, Currntly marreid, formlerly
10.	v106	Highest edcuation level	Nominal	No , Primary Secondary, higher educat
11.	v013	Age in five groups	Discrtes	15-49
12.	v714	Mothor working status	Nominal	Yes, no

13.	v501	Husband's desire for child	nominal	Yes,no
14.	v404	Currently breast feeding	Nominal	Yes, no
15.	v405	Currently amenorrheic	nominal	Yes, no
16.	v218	Number of living child	Discretes	1-12
17.	v213	Current pregnant	Nominal	Yes, no
18.	v160	Toilet facility with other household heads	Nominal	Yes, no
19.	v157	Frequency of reading newspapers	Discrtes	Not at all,less than once a week, at least once a week, almost every day
20.	v158	Frequency of listening to radio	Discrtes	Not at all,less than once a week, at least once a week, almost every day
21.	v159	Frequency of watching television	Discrtes	Not at all,less than once a week, at least once a week, almost every day
22.	v161	Types of cookoing fuels	Nominal	Electricity, IPG, Natural gas, biogas, kerosence, wood, other
23.	v705	Houshold head occupations	Nominal	Did not work,proffesional, crtical,sales, farmer, other
24.	bord	Birth order Number	Contiouse	1-18
25.	m15	Place of child delivery	Nominal	Home, private Health center, government health center, other
26.	h10	Ever had vaccination	Nominal	Yes, no
27.	v701	Education level of household head	Nominal	No education, primary, secondary, higher

28.	m18	Size of child in a birth	Continouse	Very large, lager, average, smaller,very small
29.	hw1	child's age in months	discrete	0-59
30.	hw2	child's weight	discrete	6-99
31.	H11	Have you Diarrhea diseses	Nomonal	Yes, No

3.2.3 Data preprocessing

Data preprocessing is an important phase of the data analysis activity which involves the construction of the final data set which the data is used into the modeling tool from the preliminary raw data[110]. Real-world data is inclined to be extremely disposed to noisy, missing, and inconsistent due to their typically huge size and their likely origin from multiple, heterogeneous sources. Preparing data is required to get the best results that are done within ML algorithms on data. To clean the disposed or the original dataset, the dataset must be passed through the following data pre-processing methods and techniques.

A. Data Cleaning

During collecting or entering data, transforming or extracting data, exploring or analyzing data, and submitting the draft report for peer review the data may have some errors, outliers, and some missing values. Dirty data can confuse the mining procedure, resulting in unreliable output. Therefore; data cleaning is demanded to handle missing values, identify and remove outliers, and remove unwanted data.

- **Handling inconsistency**

For data cleaning, we have changed all the values with inconsistency like unknown characters, null, NAN, NA, did not know Not stated, N/A, nan, None and none" , as a null value and filled the null values with the imputation techniques according to its data types either using the mean or mode.

- **Handling missing Values**

In the field of data-driven research, handling the missing values of data using different technique is very important. For instances, the researchers handled the missing values using either by deleting tuple, filling in the missing value using the attribute mean and mode. Before handling missing values, we should understand why and where data is missing and know the characteristics of missingness because it contributes to the success or failure of the analytical process. The number of missing values that appear in our dataset is given in table 3.2 below.

Mean or mode imputation is consists of filling in the missing value for a given feature by the mean or mode of all known values of that feature[111]. The mean imputation method of missing value treatment involves replacing a missing value with the overall sample mean for numerical features only. After identifying the types and values of features the researchers have used the mean imputation techniques like features of hw1 and hw2. We also used mode imputations for imputing missing values of categorical features like v013, v159, v160, v701, h10, v705, v714, m15 and m18 are imputed by mode of the data.

Table 3. 2 Missing values in frequency and percentage

Features Names	**Missing value in a frequency**	**Missing values in percentages**	**Features Names**	**Missing value in a frequency**	**Missing values in percentages**
v013	1	0.0025%	v705	753	1.93%
v159	39	0.10%	v714	23	0.059%
v160	22803	58.66%	m15	1	0.0025%
v701	1072	2.70%	m18	159	0.409%
h10	12979	33.38%	hw1	4944	12.71%
			hw2	6461	16.72%

- **Handling outliers in the dataset**

Outliers are also referred to as abnormalities, discordant, deviants, and anomalies, whereas noise can be defined as mislabeled examples or class noise or errors in the values of attributes[112]. These incorrect attribute values may be due to data encoding problems, incorrect data collection, and irregularity in naming convention or technology limitation. In this study, the researchers have identified and detected noise or outlier values from the data with normalization methods and with the support of domain experts. For instance, as indicated in figure 3.2, outliers in child's weight were handled using interquartile range with median.

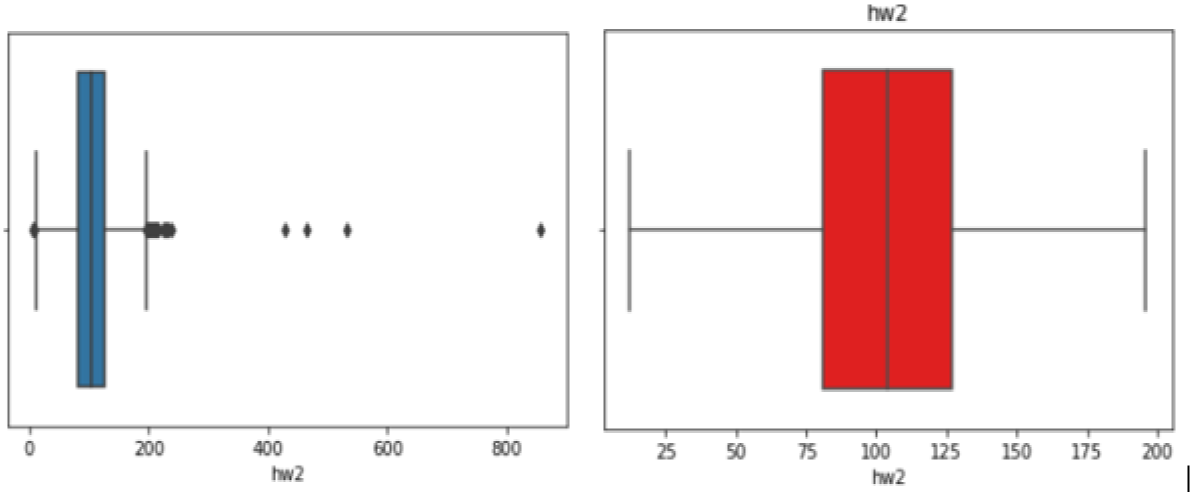


Figure 3. 2 Child's Weight in Kilogram

B. Data Transformation

Data transformation is the process that transforms the data in alternative forms for the machine learning process[113]. By replacing a large number of continuous attribute values with a small number of interval labels, the original data is reduced and simplified data. To make the dataset appropriate for this study the data discretization and data normalization technique is applied on continuous attributes to minimize distinct values of attributes. The technique of transforming continuous data attribute values into a finite set of intervals with minimal information loss is known as data discretization[114]. The goal of attribute discretization is to identify compact data representations in the form of categories that are appropriate for the learning task and reduce training while maintaining as much information as possible from the continuous original attribute. For example, features such as maternal age, husband age, total number of children, number of under-five age of children, child age and family size have been discretized.

Table 3. 3 data discretization

Feature names	Range	Interval	Discretized values
Maternal age	15-49	[15-19], [20-24], [25-39], [30-34], [35-39], [40-44], [45-49]	0,1,2,3,4,5,6
husband age	15-90	[15-24], [25-34], [35-44], [45-54], [55-64], [65-74], [75-84], [85-93]	0,1,2,3,4,5,6,7
total no_children	0-17	[0-4], [5-9], [10-15], [16-17]	0,1,2,3,4

number of u5 children	0-5	[0-2], [3-5]	0,1
Child age	0-59	[0-9], [10-19], [20-29], [30-39], [40-49], [50-9]	0,1,2,3,4,5
Family size	2-18	[0-4],5-9], [10-15], [16-18]	0,1,2,3

C. Handling Redundant Values

Data redundancy exists due to the occurrence of duplicate data values for data instances that leads to the issue of shortage of storage. The most common way of handling duplicity is by finding chunks of similar values and removing the duplicates from the chunks[115]. Originally, we have a total of 38873 instances with 31 features including the target class label. In this study, after integrating four datasets into one dataset, we got 59 redundant data. After that, all the duplicated instances were dropped in the data set. Then, we have a total of 38814 instances.

D. Target Class of the Study

This study is focused on predicting the status of diarrheal disease against under five children in Ethiopia. Then, here we have target class of diarrhea having values of yes and no. This indicated that the class label has binary class values which shows that the presence or absence of diarrheal disease with under five ages of child.

E. Class Imbalance

Due to the presence of a skewed distribution of data among the classes in data space and one class is a higher number than the other is commonly known as imbalanced data. In this ML algorithms achieve pretty high accuracy just by predicting the majority class, but fail to capture the minority class[116]. Various solutions have been presented to overcome the challenges associated with a class imbalance on the data and algorithmic levels[162]. To address this issue, we used the SMOTE[118] oversampling approach is quite effective in improving the classification accuracy of the minority instance. Figure 3.3 A illustrated that the class label before applying the data balancing technique. In this figure, the class label was having a value of 32406 of No diarrheas and 6408 yes diarrhea. But, figure 3.3 B illustrated the class label after applying the data resampling technique called SMOTE over resampling techniques that is increase the minority class label to the majority class label. So, the class label is increased to 64812 instances.

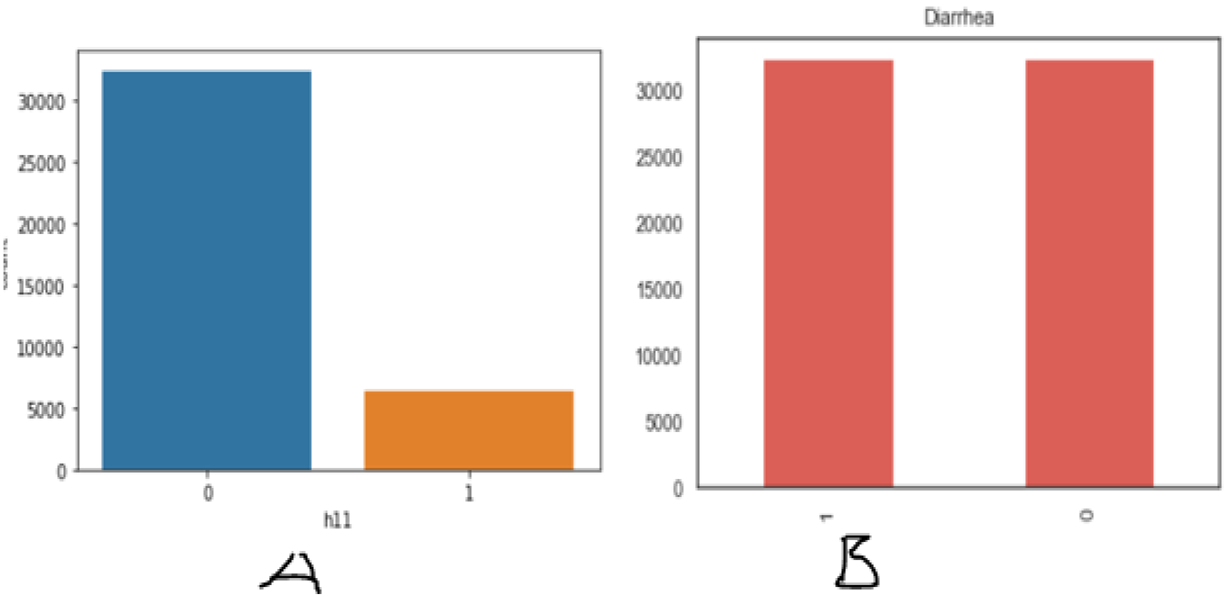


Figure 3. 3 Data Balancing

3.2.4 Feature Selection

Wrapper Method

In this study, the sequential forward and backward feature selection technique with random forest classifiers as an accuracy metric is used. Here to select relevant features with both techniques we used a greedy search strategy, by importing "pandas" libraries, and "SequentialFeatureSelector" and "RandomForestClassifier".

Sequential Forward Feature Selection (SFFS) is implemented using the SequentialFeatureSelector class of mlxtend library, by setting the forward parameter to True. It takes two parameters as input estimator and n_features_to_select. We have used the RF algorithm as the estimators to select an optimal number of features. The n_features is a user-defined parameter to specify the number of features to select ultimately. It has been set to 20 integer values using accuracy score as the best performance metric to select the best subset of features. Sequential Backward Selection (SBFS) follows the same idea with sequential forward but works in the opposite direction. Here instead of starting with no feature and greedily adding features, we start with all the features and greedily remove features from the set. The direction parameter controls whether forward or backward SFS is used.

Using the wrapper feature selections, we executed two experiments using SFFS and SBFS. Then after, the step backward method registered better performance as compared to forward method using 20 selected relevant features. Due to this, the researchers decided to use step backward feature selection with 20 selected features for further predictive model development of diarrheal disease for under five children.

Figure 3. 4 Wrapper Feature Methods

Step Forward	Step Backward
v012	v013
v024	v024
v025	v025
v113	v113
v130	v130
v136	v152
v152	v158
v157	v160
v158	v161
v159	v213
v160	v218
v161	v404
v213	v405
v218	v501
v404	v502
v405	v701
v701	v714
v714	bord
m18	m18
hw2	hw2

3.2.5 Dataset Splitting

When ML algorithms are employed to generate predictions on data, the train-test split technique is used to measure their performance[119]. In this study, we split the balanced 64812 total instances of the dataset with 20 selected features into a train and test size split proportion such as a 0.2 test split size proportion with stratified random sub-sampling. This splitting technique is based on the assumptions and standards of the train test splitting principle. Indeed 80/20 is the most recommended one, which should give us good prediction accuracy. Out of 64812 instances, 51849 instances are for training datasets that are used for building a predictive model and 12963 instances are used for testing datasets are used for the performance of the developed predictive model.

3.2.6 Predictive Model Development for Diarrhea against Under Five Child

To develop a prediction model for diarrhea against children data preparation, model selection, and parameter tuning for the algorithm are needed. The researchers need to prepare datasets for training, testing, and also setting the tuning parameters for each algorithm.

3.3.1. Hyperparameter tuning

While training the machine learning models using the default value of their parameters may not yield the most optimal model, we configured to their best values based on the tunning parameters for each algorithm. For example, parameters that choose the depth and number of leaves in XGBoost affect the model's accuracy, and choosing optimal values for these parameters will improve the accuracy of the model. There are several ways to perform hyper-parameter tuning[120]. Here, we used Grid search for selecting tuning parameters for an ensemble machine learning algorithms such as random forest, Gradient Boosting, Cat boosting and Extreme gradient boosting algorithms. Gird search is a method of hyper-parameter tuning that builds and evaluates a model methodologically for each combination of algorithm parameters supplied in a grid. Table 3.4 illustrated that the tunning parameter which are identified by grid search hyperparameters tunning for ensemble machine learning algorithms.

Table 3. 4 Tunning parameters for Ensemble Algorithm using GridSearchCV

Random Forest		Gradient Boost		Cat Boost		XGBoost	
Max depth	15	n-estimators		Iterations		n_estimators	
Max leaf node	450	learning_rate	1	depth		learning_rate	
n-estimators	500	Max depth	19	12 leaf reg		Num_leaves	
Criteria	gini			Learning rate		Max depth	

3.3.2. Model evaluation

The evaluation metric plays a critical role in achieving the optimal classifier during the classification training[121]. Many generative classifiers employ accuracy as a measure to discriminate against the optimal solution during the classification training[121]. The performance of the model is evaluated using performance evaluation metrics such as confusion matrix, accuracy, ROC, cross-validation, precision, recall, and confusion matrix, and also the risk factors and relevant rules validated by domain experts.

Confusion Matrix

In the confusion matrix, a 2x2 table readily summarizes all four possible decisions: true positive (Tp), false negative (Fn), false positive (Fp), and true negative (Tn). The Tp and Tn decisions are correct while the Fn and Fp decisions are incorrect[84].

Table 3. 5 confusion matrix for both actual and predicted class

		Predicted Class	
		Negative	Positive
Actual Class	Negative	TN	FP
	Positive	FN	TP

TP is a true positive that is an outcome where the model correctly predicts the positive class. TN is a true negative that is an outcome where the model correctly predicts the negative class. FP is a false positive that is an outcome where the model incorrectly predicts the positive class. FN is a false negative that is an outcome where the model incorrectly predicts the negative class.

Accuracy

Model accuracy in machine learning is the measurement used to determine which model is best at identifying relationships and patterns between features in a dataset based on the input. The better a model can generalize to 'unseen' data, the better predictions and insights it can produce, which in turn deliver more business value[85]. The accuracy of a machine learning classification algorithm is one way to measure how often the algorithm classifies a data point correctly. Accuracy is the number of correctly predicted data points out of all the data points. The accuracy is the proportion of the total number of correct predictions. It is determined using the equation: $\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP+TN+FN+FP}}$

Equation 3. 6 Accuracy

Sensitivity /Recall is defined as the ratio of correctly classified majority class values (true positives) divided by the sum of correctly classified majority class values (true positives) and incorrectly classified minority class values (false positives). it should be high [51]. Sensitivity reflects the ability of the observer to correctly classify the target and is calculated as

$$Sensitivity\ /recall\ = \frac{Tp}{Tp+Fn} = \frac{Tp}{p}$$

Equation 3. 2. $Sensitivity/recall$

Specificity reflects the ability of the observer to correctly classify the target and is calculated as $specificity\ = \frac{Tn}{Fp+Tn}$

Equation 3.3. specificity

The false-positive rate (FP) is the proportion of negatives cases that were incorrectly classified as positive, as calculated using the equation:

$$FP\ = \frac{Fp}{Tn + Fp}$$

Equation 3. 4. False-positive rate

Precision is defined as the ratio of correctly classified majority class values (true positives) divided by the sum of correctly classified majority class values (true positives) and incorrectly classified majority class values (false positives). it should be high [51].

$$Precision\ = \frac{Tp}{Tp + Fp}$$

Equation 3. 5 precisions

The true negative rate (Tn) is defined as the proportion of negatives cases that were classified correctly, as calculated using the equation:

$$Tn = \frac{Tn}{Tn + Fp}$$

Equation 3. 67. True negative rate

$$Error\ rate\ = 1 - accuracy = \frac{Fp+Fn}{Tp+Tn+Fn+Fp}$$

Equation 3. 78 Error Rate

Receiver operating characteristics (ROC) curve

ROC analysis and the resulting ROC curve is a method that captures the relationship between sensitivity and specificity as well as the range of decision thresholds that every observer has no matter what their level of expertise and experience[84].

Cross-Validation

Cross-validation is a data resampling method to assess the generalization ability of predictive models and to prevent over fitting and used to estimate the true prediction error of models and to tune model parameters[86]. The cross-validation estimate of a given performance metric is simply the mean of the performance in each fold of the cross-validation procedure [87].

In this study, evaluation metrics such as accuracy, cross-validation, ROC, precision, and recall as well as the confusion matrix are used that evaluate the performance of the prediction model.

5.1. Requirement Specification

To construct a predictive model for diarrhea against under five children using ensemble ML algorithms, hardware, and software specifications are necessary. The hardware system consists of an Intel(R) Core(TM) i5- CPU 2.60GHz, 8GB RAM laptop running windows 10 (64-bit). The software systems we used Microsoft office word 2016 for writing documentation, Microsoft office excels 2016 for understanding the datasets manually, Edraw max to draw figures, and we used Python 3.8 and its associated third-party libraries for building the models, processing the data, and visualizing the results, etc., and Microsoft office presentation 2016 for thesis presentation.

CHAPTER FOUR

4.1. EXPERIMENTAL SETUP RESULT AND DISCUSSION

4.2. Experimental Setup

Dataset

Initially, the original dataset consists of 38873 instances and 31 features from 2000 to 2016 EDHS Ethiopia. However, the dataset contained missing values, duplicated instances and it is imbalanced. Among these features, features such as v013, v159, v160, v701, h10, h705, v714, m15, m18, hw1 and hw2 had missing values. We have also 59 duplicated instances. After drooped the duplicated instances we had 38814 instances. But this dataset has class imbalance and we applied SMOTE and we got a total of 64,816 instances which are ready for further predictive model. Despite having a total of 31 features, not all of them were equally important for the development of the predictive model of diarrheal disease among under five ages of children in Ethiopia.

Tools and package

To construct a predictive model for diarrhea among under five children using ensemble ML algorithms, hardware, and software specifications are necessary. For instance, Laptop computer with 4RAM, 500HD, 2.60 GHZ as well as software tools such as Microsoft office, Edraw max and anaconda distribution with Jupiter Notebook. In Jupiter Notebook, we have used different python modules, packages and also libraries. For example, packages like sklearn, mlxtend and also libraries such as matplotlib, seaborn, pandas, train-test-split, XGBoost, CatBoost, etc.

4.3. Result

4.3.1. Significant risk Factors

Hence, researchers implemented wrapper feature selection methods to identify the significant features which are essential for developing the predictive model. According the result of wrapper feature selection methods (step forward and backward), features selected by the step backward feature selection method with random forest classifiers registered the highest performance the predictive model.

To identify relevant features using wrapper methods, which involve a sequential greedy search strategy, we imported "pandas" libraries such as "SequentialFeatureSelector" and "RandomForestClassifier". For the "SequentialFeatureSelector", we configured parameter sets including (k_features=20, forward="True" for the forward method, and forward="False" for the backward selection method, scoring="accuracy", cross-validation cv=5). Additionally, we set the necessary tuning parameters for the RandomForestClassifier using "GridSearchCV", such as max_depth=10, max_leaf_nodes=21, criterion="entropy", n_estimators = 200 and random state = 43).

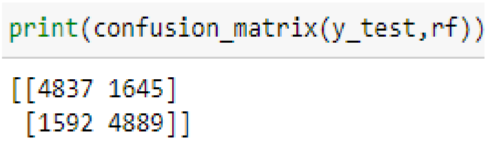
Therefore, the significant factors for building the predictive model for diarrheal diseases among under five children in Ethiopia have been identified using step backward feature selection method as Age group of mother, region, Place of residents, Source of drinking water, religion, Age of household head, Frequency of listening to radio, toilet facility with other household heads, Types of cooking fuels, Current pregnant , Number of living child, Currently breast feeding, Currently amenorrheic , Husband's desire for child, Marital status ,Education level of household head, Mother working status, Birth order Number, Size of child in a birth and , child's weight.

4.3.2. Predictive Model Development

Once the relevant features were identified, the researchers proceeded to build a predictive model for diarrheal disease prediction using ensemble algorithms including random forest, Cat boosting, Gradient Boosting, and Extreme Gradient Boosting ensemble algorithms. Table 4.1 showcases the predictive model developed using selected 20 features for diarrheal disease prediction. The experimental result of the predictive model with confusion matrix for ensemble algorithms.

Confusion matrix for:

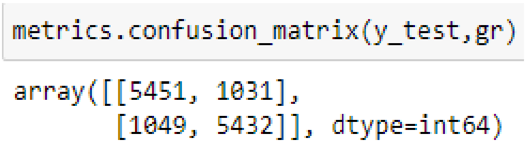
Random Forest:



```
print(confusion_matrix(y_test,rf))
```

```
[[4837 1645]
 [1592 4889]]
```

Gradient Boosting:

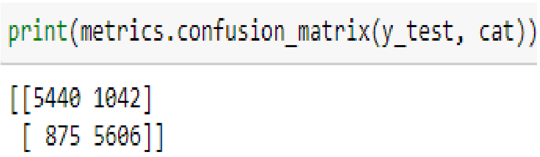


```
metrics.confusion_matrix(y_test,gr)
```

```
array([[5451, 1031],
       [1049, 5432]], dtype=int64)
```

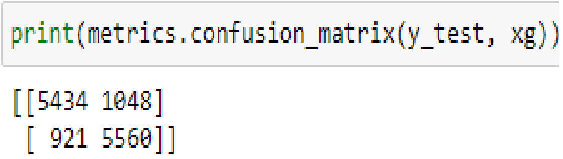
CAT Boost

XGBoost



```
print(metrics.confusion_matrix(y_test, cat))
```

```
[[5440 1042]
 [ 875 5606]]
```



```
print(metrics.confusion_matrix(y_test, xg))
```

```
[[5434 1048]
 [ 921 5560]]
```

Table 4. 1 The Experimental Results for Predictive Model of Diarrheal diseases

Evaluation metrics	**Ensemble Algorithms**			
	RF	**GATB**	**CATB**	**XGBT**
Accuracy	**75.03**	**83.95**	**85.21**	**84.81**
Precision	74.82	84.04	85.28	84.82
Recall	75.43	83.81	85.21	84.81
F1 Score	75.12	83.93	85.21	84.81

The researchers have conducted a four experiments using the selected ensemble machine learning algorithms for developing the predictive model of diarrheal disease among under five children in Ethiopia. The performance of the predictive models was evaluated using various metrics, as presented in Table 4.1. When compared to the predictive model performances of algorithms listed in Table 4.1, CatBoost exhibited the highest accuracy followed by XGBoost.

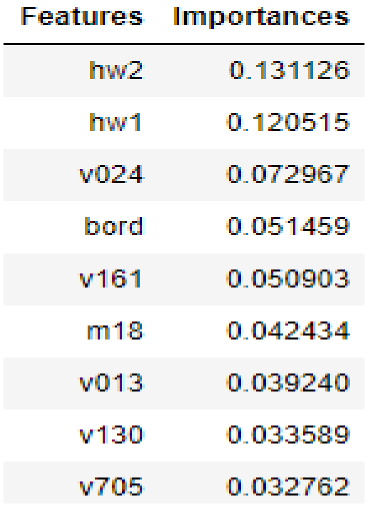
Based on the comprehensive analysis of model performance, the ensemble algorithm that demonstrated the highest accuracy is selected as the optimal choice for predicting whether children have diarrhea or not.

Table 4.2. Comparison of Algorithm Performance based on Accuracy

Classification Algorithms	Correctly classified		Incorrectly classified	
	Number	Accuracy	Number	Error rate
RF	9726	**75.03%**	3237	**24.97%**
GBT	10883	**83.95%**	2080	**16.05%**
CATB	11046	85.21%	1917	14.79%
XGBT	10994	**84.81%**	1969	15.19%

4.3.3. Important Features for Diarrheal Disease Predication

During the development of the predictive model, a total of 20 features were initially selected using step backward feature selection technique. However, considering the importance of prioritizing certain features for policymakers to focus on interventions, addressing all 20 features simultaneously becomes challenging. To overcome this, we employed the feature importance score technique to rank the associated risk factors based on their significance (refer to Figure 4.3 below). Out of the 20 features, features hw2, hw1, v024, bord, v161, m18, v013, v130, and v705 were found to have high influences on diarrhea concerning diseases.



Features	Importances
hw2	0.131126
hw1	0.120515
v024	0.072967
bord	0.051459
v161	0.050903
m18	0.042434
v013	0.039240
v130	0.033589
v705	0.032762

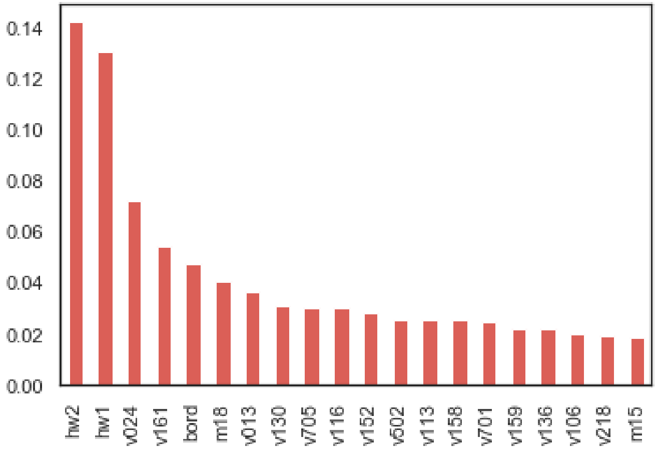


Figure 4.3 Risk factors

4.4. Discussion

As previously mentioned, a study was conducted involving a total of 64,814 instances. During the study, 31 features were considered, but their importance varied. Additionally, the ensemble algorithms utilized in developing the predictive model demonstrated diverse performances. By doing empirical experiments, we come up with the following three basic findings. First: on the identification of the significant risk factors for developing the predictive model. Second: on the selection of the best appropriate ensemble machine learning algorithm and then identification of the best important using feature importance method.

RQ1: Which risk factors are associated significantly with diarrhea diseases among under five ages of children?

After preprocessed the data, we need to select the associated significant risk factors for building the predictive model. Hence, we executed two experiment using wrapper feature selection methods such as step forward and backward methods with random forest as estimator. From these experiments, features selected by the step backward method registered better performance as compared to the step forward method. In this method of feature selection, features such as Age group of mother, region, Place of residents, Source of drinking water, religion, Age of household head, Frequency of listening to radio, toilet facility with other household heads, Types of cooking fuels, Current pregnant , Number of living child, Currently breast feeding, Currently amenorrhea , Husband's desire for child, Marital status ,Education level of household head, Mother working status, Birth order Number, Size of child in a birth, and child's weight are selected for model development.

RQ2: Which ensemble machine learning technique is suitable to construct a predictive model for under-five age of children diarrhea?

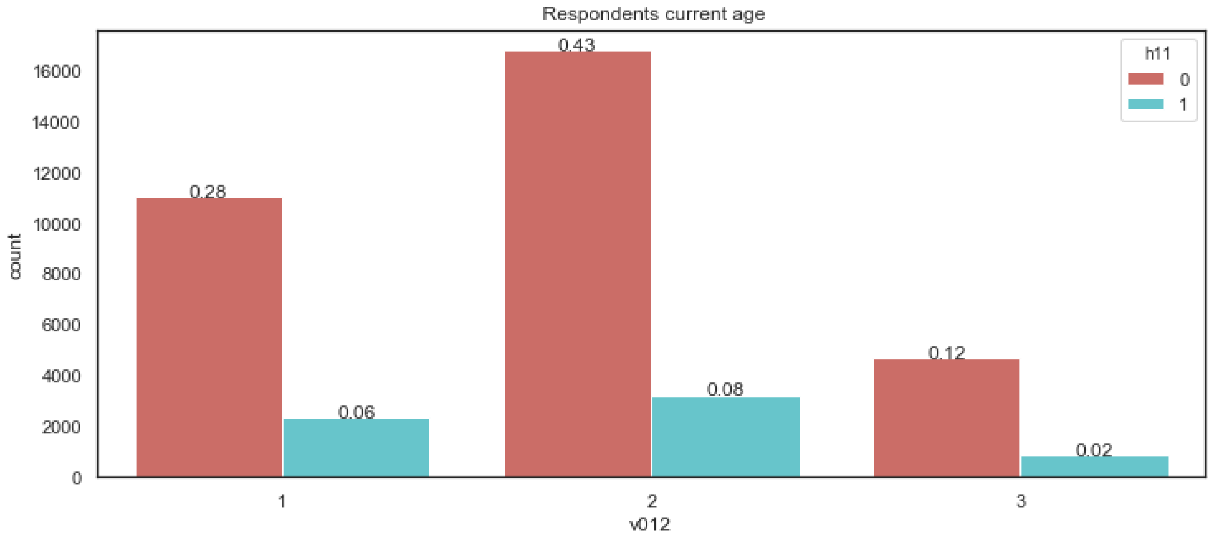
To answer this research question and come up with the best performed predictive model a set of four ensemble machine learning algorithms such as random forest, gradient boosting, Cat boosting and extreme gradient boosting ensemble algorithms were trained. These machine learning algorithms trained properly with tuning of their hyper parameters to optimize their predictive performance and with the help of grid search as indicated in table 3.4. For the above-mentioned algorithms, accuracy, precision, recall, f1 score and ROC are used the evaluation metrics. According to the experimental result of table 4.1, eXtreme gradient boosting (XGBoost) algorithm is outperformed in terms of accuracy, precision, recall, f1 score and ROC as compared other listed ensemble algorithms. Hence, CatBoost exhibited the highest accuracy followed by XGBoost.

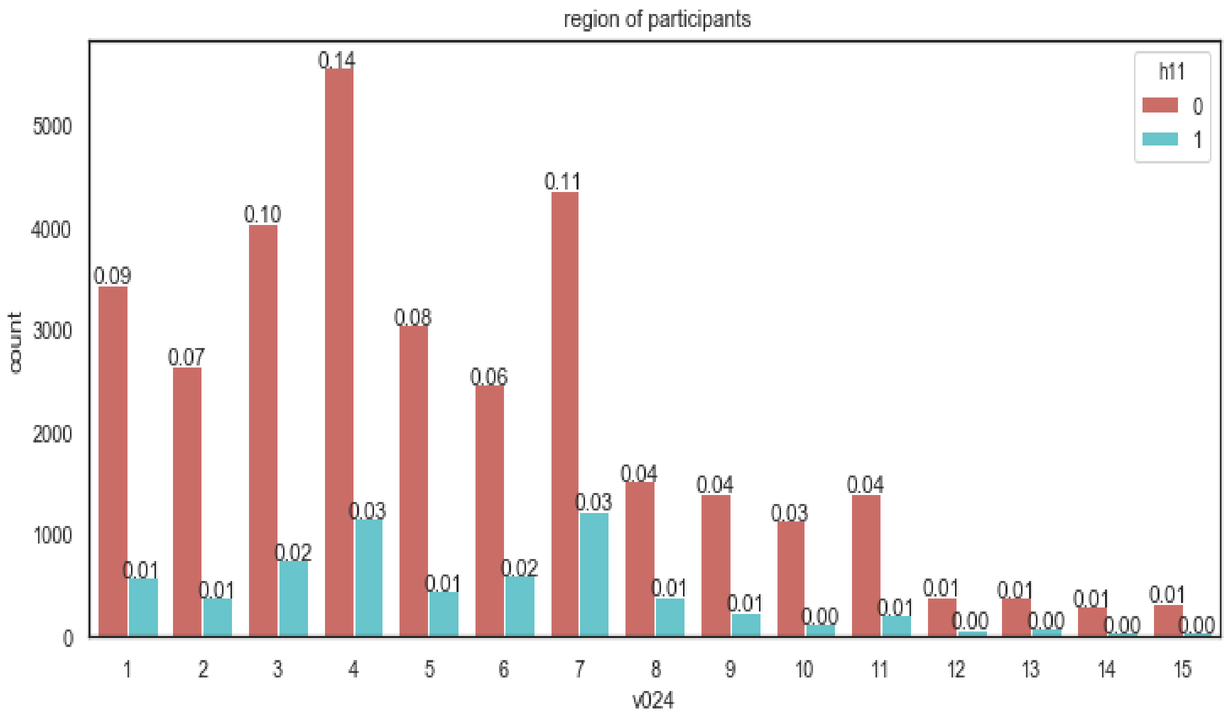
RQ3: What are the best important features that contribute to predicting the occurrence of diarrhea disease among under-five Children?

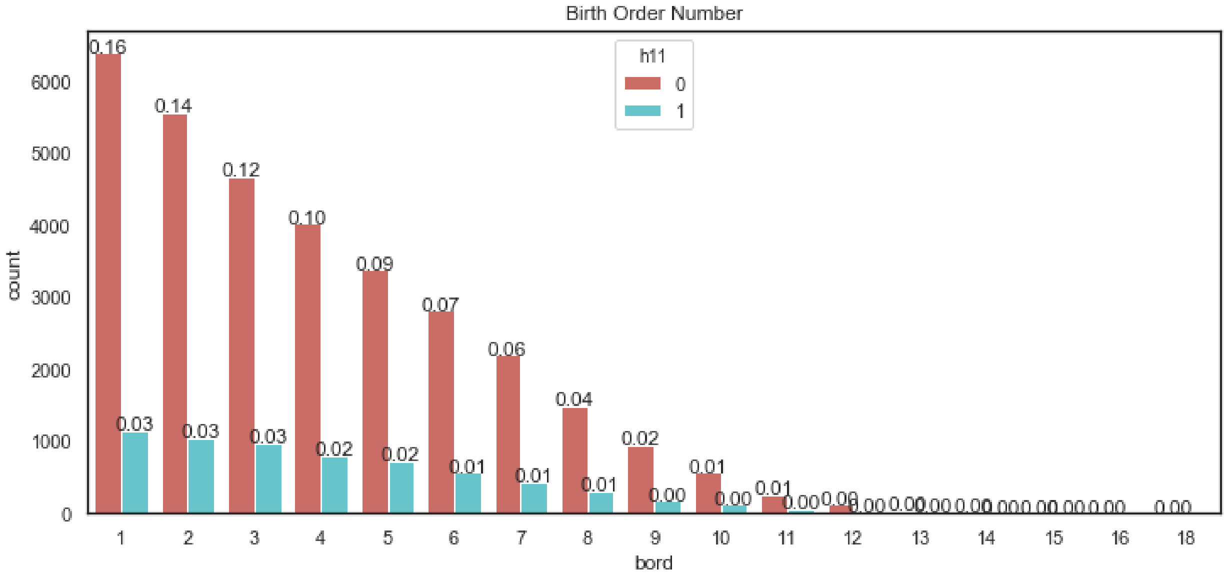
The second question (RQ2) to determine what are the best important features that contribute to predicting the occurrence of diarrhea disease among under-five Children? We have selected 20 features in the data set using sequential backward feature selection for built a predictive ML model with the selected model using these 20 features.

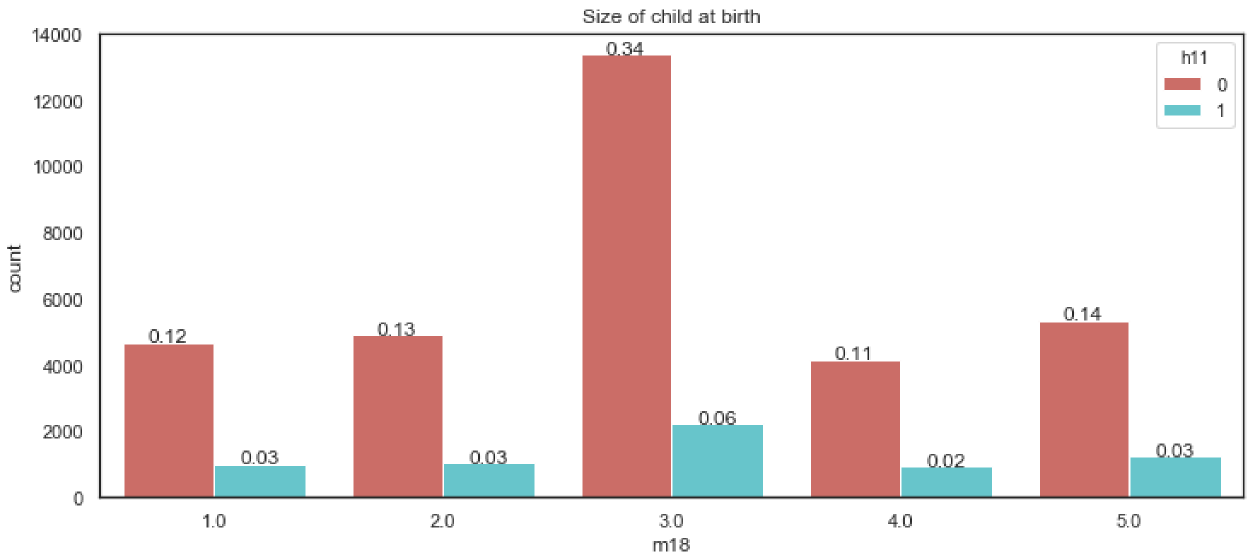
The importance of all the features in the data set is calculated using a feature importance experiment conducted through the feature_importances package from sklearn python to identify the most relevant risk factors. In doing this we have selected important features after building the modeling process by using the algorithm that achieves the highest performance in terms of accuracy score which is the XGBoost ensemble method. The selected important determinant risk factors that mostly affect the predictive model performance are: hw2, hw1, v024, bord, v161, m18, v013, v130, and v705 were found to have high influences on diarrhea concerning diseases.

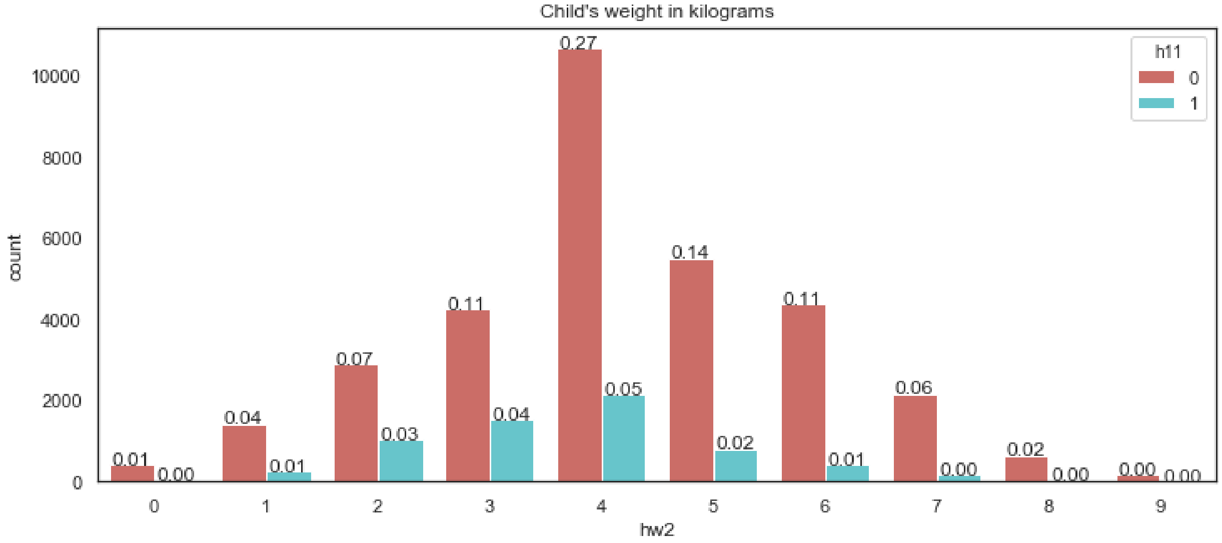
To understand the underlying structure of diarrheal disease specifically in the context of Ethiopia with the above selected best important features, a descriptive statistical technique was employed to analyze the relevant features. This approach aimed to reveal patterns, trends, and relationships among the variables associated with diarrheal disease in the Ethiopian context. So here, we described the relationship between each feature that contributes to the diarrhea diseases.

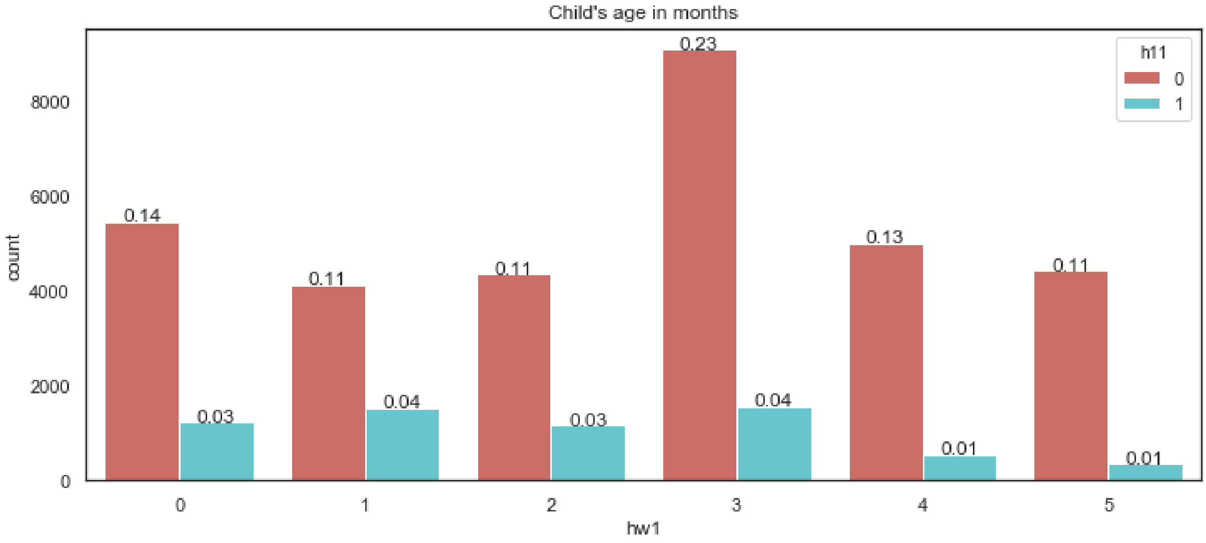


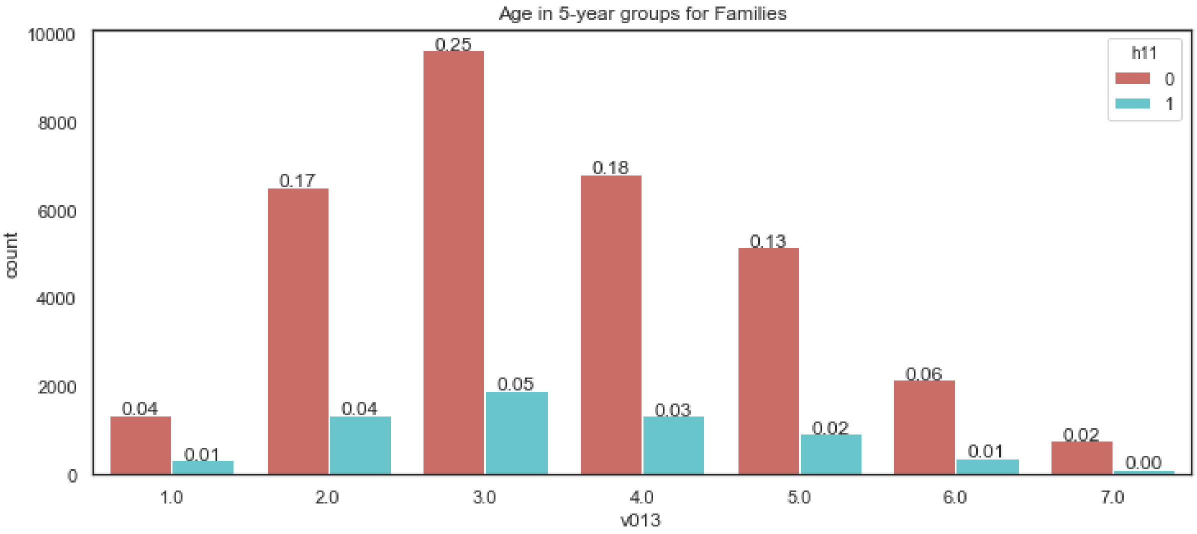

region of participants
h11
0
1
count
5000
4000
3000
2000
1000
0
v024
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
0.09
0.01
0.07
0.01
0.10
0.02
0.14
0.03
0.08
0.01
0.06
0.02
0.11
0.03
0.04
0.01
0.04
0.01
0.03
0.00
0.04
0.01
0.01
0.00
0.01
0.00
0.01
0.00
0.01
0.00

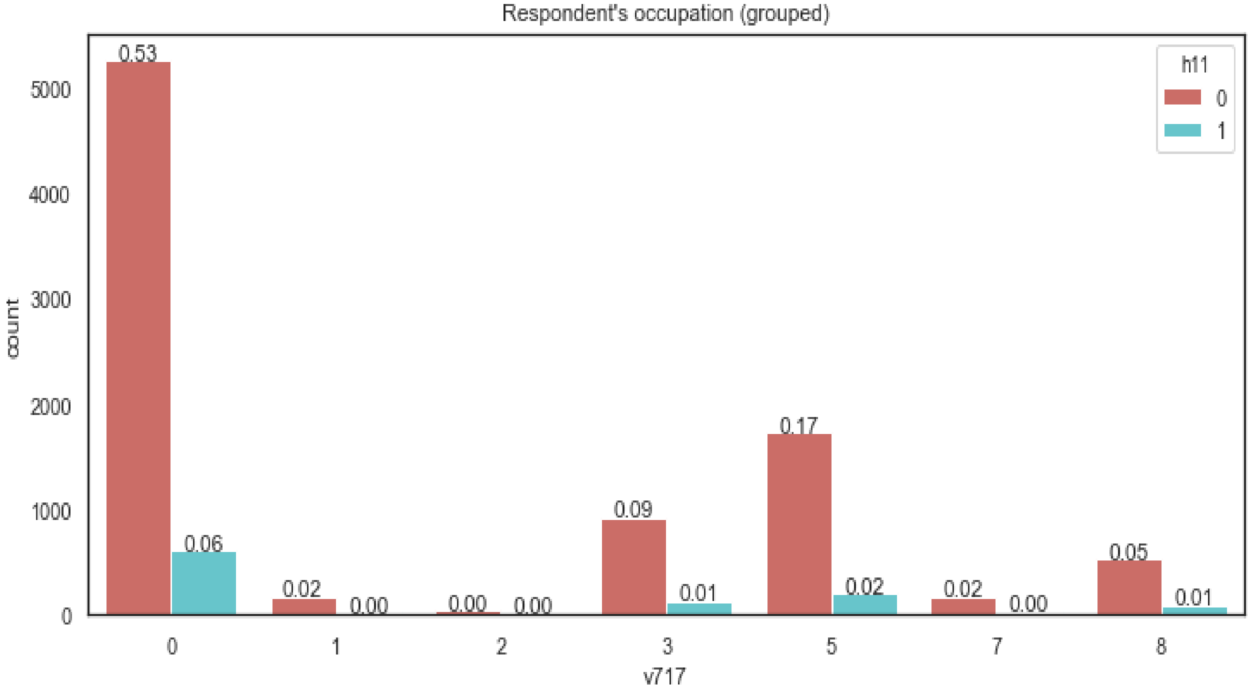

Birth Order Number
h11
0
1
count
6000
5000
4000
3000
2000
1000
0
bord
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
18
0.16
0.03
0.14
0.03
0.12
0.03
0.10
0.02
0.09
0.02
0.07
0.01
0.06
0.01
0.04
0.01
0.02
0.00
0.01
0.00
0.01
0.00
0.00
0.00
0.00
0.00
0.00
0.00
0.00
0.00
0.00
0.00
0.00
0.00

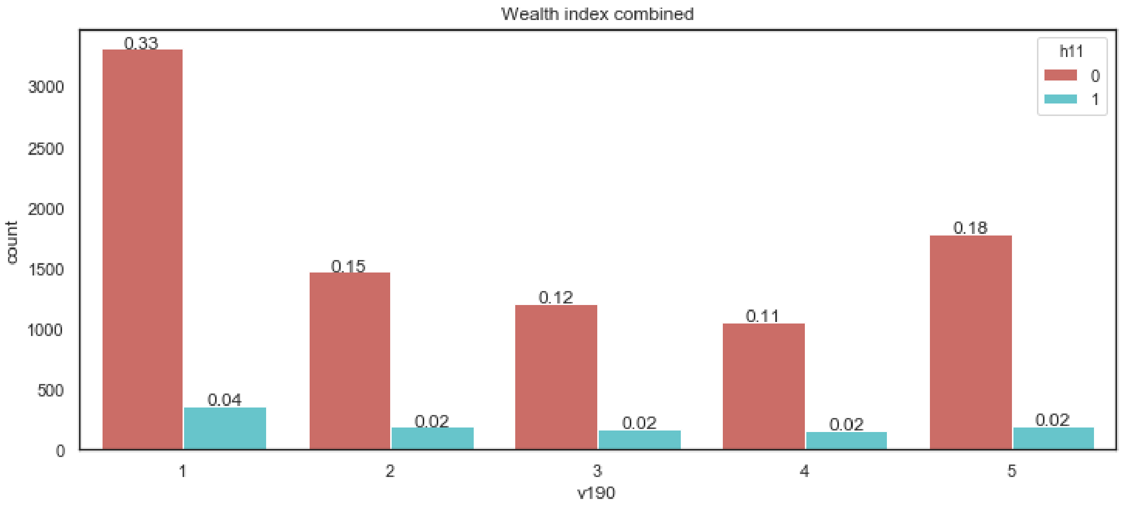

Size of child at birth
count
14000
12000
10000
8000
6000
4000
2000
0
h11
0
1
0.12
0.03
0.13
0.03
0.34
0.06
0.11
0.02
0.14
0.03
1.0
2.0
3.0
4.0
5.0
m18


Child's weight in kilograms
count
10000
8000
6000
4000
2000
0
h11
0
1
0.01
0.00
0.04
0.01
0.07
0.03
0.11
0.04
0.27
0.05
0.14
0.02
0.11
0.01
0.06
0.00
0.02
0.00
0.00
0.00
0
1
2
3
4
5
6
7
8
9
hw2


Child's age in months
count
8000
6000
4000
2000
0
h11
0
1
0.14
0.03
0.11
0.04
0.11
0.03
0.23
0.04
0.13
0.01
0.11
0.01
0
1
2
3
4
5
hw1


Age in 5-year groups for Families
count
0
2000
4000
6000
8000
10000
v013
1.0
2.0
3.0
4.0
5.0
6.0
7.0
h11
0
1
0.04
0.01
0.17
0.04
0.25
0.05
0.18
0.03
0.13
0.02
0.06
0.01
0.02
0.00


Respondent's occupation (grouped)
count
0
1000
2000
3000
4000
5000
v717
0
1
2
3
5
7
8
h11
0
1
0.53
0.06
0.02
0.00
0.00
0.00
0.09
0.01
0.17
0.02
0.02
0.00
0.05
0.01



Domain Expertise Evaluation of the Findings

To assess the effectiveness of the significant risk factors and the best important features in enhancing the predictive model performance, domain experts were involved in the validation process. Three experts from the University of Gondar, CMHS, and Department of Pediatrics were invited to evaluate the best important features and their associated risk factors, considering their significance.

During the evaluation, three domain experts (2 male and 1 female from the University of Gondar Pediatric department) were involved for overall suggestions about the relevance of associated risk factors and generated sample rules. At the end of their evaluation, they were asked to assign values (Poor=1, Fair=2, Good=3, very good=4, and Excellent=5) based on the evaluation criteria set in Table 4.3 below. As the domain experts' response is summarized in Table 4.3. Below, the values indicated the numbers of evaluators who evaluate the system as Poor, Fair, Good, Very Good, and Excellent concerning evaluation criteria.

Table 4. 2 Domain expert's evaluation

No	**Criteria evaluation**	**Poor**	**Fair**	**Good**	**Very good**	**Excellent**	**Remark**
		1	2	3	4	5	
1.	The relevance of significant risk factors for diarrheal disease follow-up	0	0	1	2	0	3.66

2.	The relevance of the developed predictive model for diarrheal diseases	0	0	0	2	1	4.33
3.	The relevance of the selected best important features for intervention of policy in diarrheal diseases	0	0	0	2	1	4.33
4.	Appropriateness of generated rules for policy intervention	0	0	2	0	1	3.66
5.	Interpretability and importance of generated rules by domain experts	0	0	1	2	0	3.66
Total percentage of Values in each Metrics							**3.92**
		Average					**78.56**

As we can understand from Table 4.3 here above, for the first criteria, 1, and 2 of the total evaluators give good, and very good, respectively, responses to the relevance of risk factors for diarrheal disease follow-up. This means that 33.33% and 66.66% of the evaluators are good and very good respectively. For the second criteria, the evaluators give 2 and 1 of very good and excellent respectively. This means that 66.66%% and 33.33% of the evaluators are very good, and excellent, respectively. The third criteria are the relevance of the selected best important features for intervention of policy in diarrheal diseases. Here, the evaluator's give 2 and 1 of good and excellent respectively. This means that 66.66%% and 33.33% of the evaluators are very good, and excellent, respectively. In the fourth criteria, the appropriateness of generated rules for policy intervention. In this criterion, 66.66 % of evaluators gives good result and 33.33 % of evaluator gave excellent values. Lastly, the interpretability of extracted rules by domain experts were considered and 1 evaluator gave good value and 2 evaluators gave very good values. This means that 33.33%% and 66.66%% of the evaluators are good and very good respectively.

Finally, the researchers understands that the study was effective in identifying the relevance of risk factors for diarrhea against under five children. Then, the researchers understood that the significant risk factors, predictive model, selected best important features have been important for the policy interventions and also achieved 76.58% of the domain expert's acceptance.

CHAPTER FIVE

CONCLUSION AND RECOMMENDATION

5.1. Conclusion

According to the World Health Organization (WHO), diarrhea is defined as the passage of three or more loose or watery stools in 24 hour. In low and middle-income countries, diarrhea is still a leading cause of death and health loss among children under the age of five. Diarrhea is commonly a sign of an infection in the intestinal tract that is caused by different bacteria, viruses, and parasitic entities. According to the progress report of the international vaccine access center (IVAC) in 2020, Ethiopia is one of the focus countries for pneumonia and diarrhea progress report of 2020 with 44,692 under-five pneumonia and diarrhea death. Although the mortality from diarrhea has declined considerably over the past 25 years globally, diarrhea-associated morbidity in sub-Saharan Africa remains unacceptably high. Diarrhea has been associated with reduced growth, impaired cognitive function, reduced vaccine efficacy, and disruption of physical and educational development in children. Ethiopia ranks among the top five countries worldwide with a significant under-five child mortality rate. However, there has been progress in reducing this rate, with an annual reduction of 4.7%. In 2019, the country recorded an average under-five mortality rate of 51 deaths per 1000 live births. Critical factors contributing to under-five mortality in Ethiopia include diseases such as Acute Respiratory Infection (ARI), fever, and diarrhea.

To tackle this problem, the researcher conducted this study. The study aimed to develop a predictive model for diarrhea in under-five age of children using ensemble machine learning Algorithms in Ethiopia. The data source for this research is the Ethiopia Demographic and Health Survey collected from the 2000 to 2016 EDHS. To handle data imbalance, the researcher applied an oversampling technique called Synthetic Minority Technique (SMOTE) and after correcting data imbalances data set size is increasing from 32406 instances to 64812 instances with 30 features. The researcher has been used 20 best features from 30 features of the dataset using sequential forward selection. Here, the researcher decided to use these features to develop a predictive model using ensemble machine learning algorithms namely XGBoost, CatBoost Gradient boost, and Random Forest.

The performances of the models are evaluated using both objective and subjective based evaluation metrics such as the standard metrics of accuracy, F1 score, precision, recall, and by domain experts' respectively. In this study, ensemble machine learning algorithm is identified using the accuracy of predictive model performance. Then, the best algorithm that predicts the status of diarrhea in under five children is constructed by CatBoost having 85.21% of accuracy, and 76.8% with subjective evaluation metrics. By doing practical experiments this study has come up with three basic findings with the following conclusions. The first one is, significant risk factor identification, the second one is we experimentally examine to identify the best performing algorithm among the four algorithms which are used in this study. From the results of the experiment, CatBoost is identified as the suitable ensemble machine learning algorithm that can be used to predict the status of diarrhea in under five children the case of this study. The third one is, we practically showed that important factors have enhance the performance of predictive models for the diarrhea U5 prediction approach. Important factors were extracted from using CatBoost and these important factors can be used by policymakers to formulate evidence-based policies and interventions towards preventing and controlling diarrhea in U5 children in Ethiopia. Some of the selected important factors are Child's weight, child's age, Region, bord, Types of cooking fuels and Size of child in a birth, etc.

5.2. Recommendation

References

[1] Y. Feleke, A. Legesse, and M. Abebe, “Prevalence of Diarrhea , Feeding Practice , and Associated Factors among Children under Five Years in Bereh District ,” vol. 2022, 2022.

[2] H. Hussein, “Prevalence of Diarrhea and Associated Risk Factors in Children Under Five Years of Age in Northern Nigeria : A Secondary Data Analysis of Nigeria Demographic and Health Survey 2013 .,” no. May, 2017.

[3] D. G. Feleke, E. S. Chanie, F. T. Admasu, S. Bahir, A. T. Amare, and H. K. Abate, “Two-week prevalence of acute diarrhea and associated factors among under five years’ children in Simada Woreda, South Gondar Zone, Northwest Ethiopia, 2021: a multi-central community based cross-sectional study,” *Pan Afr. Med. J.*, vol. 42, 2022, doi: 10.11604/pamj.2022.42.12.32599.

[4] K. Alemayehu, L. Oljira, M. Demena, A. Birhanu, and D. Workineh, “Prevalence and Determinants of Diarrheal Diseases among Under-Five Children in Horo Guduru Wollega Zone, Oromia Region, Western Ethiopia: A Community-Based Cross-Sectional Study,” *Can. J. Infect. Dis. Med. Microbiol.*, vol. 2021, 2021, doi: 10.1155/2021/5547742.

[5] A. Getachew and J. Azanaw, “Diarrhea Prevalence and Associated Factors among Under-Five Children in the Periphery Area of Azezo Sub-city , Gondar , Northwest Ethiopia : A community based cross-sectional study,” pp. 1–21, 2022.

[6] M. C. Medeiros, “Forecasting with Machine Learning Methods,” *Adv. Stud. Theor. Appl. Econom.*, vol. 53, pp. 111–149, 2022, doi: 10.1007/978-3-031-15149-1_4.

[7] A. Bannach-Brown *et al.*, “Machine learning algorithms for systematic review: Reducing workload in a preclinical review of animal studies and reducing human screening error,” *Syst. Rev.*, vol. 8, no. 1, 2019, doi: 10.1186/s13643-019-0942-7.

[8] G. Tzanis, I. Katakis, and I. Vlahavas, “Modern Applications of Machine Learning,” no. May 2014, 2006.

[9] P. Health, “Moderate to Severe Diarrhea and Associated Factors Among Under-Five Children in Wonago District , South Ethiopia : A Cross-Sectional Study,” pp. 437–443, 2020.

[10] A. B. Dagnew *et al.*, “Prevalence of diarrhea and associated factors among under-five children in Bahir Dar city, Northwest Ethiopia, 2016: A cross-sectional study,” *BMC Infect.*

Dis., vol. 19, no. 1, pp. 3–9, 2019, doi: 10.1186/s12879-019-4030-3.

[11] "Pneumonia Progress Report 2020," 2020.

[12] Thuy Linh Nguyen, "Pneumonia & Diarrhea Progress Report," *IVAC Johns Hopkins Univ.*, 2015, [Online]. Available: http://www.jhsph.edu/research/centers-and-institutes/ivac/resources/IVAC-2015-Pneumonia-Diarrhea-Progress-Report.pdf

[13] M. Alemayehu, T. Alemu, and A. Astatkie, "Prevalence and Determinants of Diarrhea among Under-Five Children in Benna Tsemay District , South Omo Zone , Southern Ethiopia : A Community-Based Cross-Sectional Study in Pastoralist and Agropastoralist Context," vol. 2020, pp. 1–11, 2020.

[14] A. A. Alemu, M. S. Bitew, K. A. Gelaw, L. B. Zeleke, and G. M. Kassa, "Prevalence and determinants of uterine rupture in Ethiopia: a systematic review and meta-analysis," *Sci. Rep.*, vol. 10, no. 1, pp. 1–20, 2020, doi: 10.1038/s41598-020-74477-z.

[15] J. Moon, J. W. Choi, J. Oh, and K. Kim, "Risk factors of diarrhea of children under five in Malawi: based on Malawi Demographic and Health Survey 2015–2016," *J. Glob. Heal. Sci.*, vol. 1, no. 2, pp. 1–13, 2019, doi: 10.35500/jghs.2019.1.e45.

[16] B. Melese, W. Paulos, F. H. Astawesegn, and T. B. Gelgelu, "Prevalence of diarrheal diseases and associated factors among under-five children in Dale District, Sidama zone, Southern Ethiopia: A cross-sectional study," *BMC Public Health*, vol. 19, no. 1, pp. 1–10, 2019, doi: 10.1186/s12889-019-7579-2.

[17] G. Kabew, B. Mengistie, G. Sahilu, and H. Kloos, "Impact of hygiene promotion intervention on acute childhood diarrhea : evidence from a cluster-randomized trial in refugee communities in Gambella Region , Ethiopia," vol. 00, no. 0, pp. 1–15, 2023, doi: 10.2166/washdev.2023.111.

[18] D. Wolde, G. A. Tilahun, K. S. Kotiso, and G. Medhin, "The Burden of Diarrheal Diseases and Its Associated Factors among Under-Five Children in Welkite Town : A Community Based Cross-Sectional Study," vol. 67, no. October, pp. 1–9, 2022, doi: 10.3389/ijph.2022.1604960.

[19] D. Chilot *et al.*, "Geographical variation of common childhood illness and its associated factors among under - five children in Ethiopia : spatial and multilevel analysis," *Sci. Rep.*, pp. 1–11, 2023, doi: 10.1038/s41598-023-27728-8.

[20] A. F. Nwaoha, C. C. Ohaeri, and E. C. Amaechi, "Prevalence of diarrhoea, and associated

risk factors, in children aged 0-5 years, at two hospitals in Umuahia, Abia, Nigeria," *UNED Res. J.*, vol. 9, no. 1, pp. 7–14, 2017, doi: 10.22458/urj.v9i1.1672.

[21] K. Sadiq *et al.*, "Risk factors for acute diarrhoea in children between 0 and 23 months of age in a peri-urban district of Pakistan: a matched case–control study," *Int. Health*, pp. 1–7, 2022, doi: 10.1093/inthealth/ihac022.

[22] P. Leni, "Household Risk Factors on the Event of Diarrhea Disease," *J. Ilmu Kesehat. Masy.*, vol. 10, no. 1, pp. 50–58, 2019, doi: 10.26553/jikm.2019.10.1.50-58.

[23] "AFRICAN CENTER OF EXCELLENCE IN DATA SCIENCE," 2022.

[24] R. M. Kananura, "Machine learning predictive modelling for identification of predictors of acute respiratory infection and diarrhoea in Uganda's rural and urban settings," *PLOS Glob. Public Heal.*, vol. 2, no. 5, p. e0000430, 2022, doi: 10.1371/journal.pgph.0000430.

[25] Maniruzzaman, I. Shaykhul, A. Menhazul, Amanullah, and H. Sadiq, "Prediction of Childhood Diarrhea in Bangladesh using Machine Learning Approach," *Insights Biomed. Res.*, vol. 4, no. 1, pp. 111–116, 2020, doi: 10.36959/584/456.

[26] T. L. Ayers, "Machine learning approaches for assessing moderate-to-severe diarrhea in children < 5 years of age, rural western Kenya 2008-2012," 2016, [Online]. Available: https://scholarworks.gsu.edu/sph_diss/10

[27] D. Gile, "Experimental research," *Res. Transl. Interpret.*, no. January, pp. 220–228, 2015, doi: 10.30574/wjarr.2022.16.3.1152.

[28] O. Mitchell, "Experimental Research Design," *Encycl. Crime Punishm.*, pp. 1–6, 2015, doi: 10.1002/9781118519639.wbecpx113.

[29] D. G. Feitelson, "Experimental computer science," *Commun. ACM*, vol. 50, no. 11, pp. 24–26, 2007, doi: 10.1145/1297797.1297817.

[30] T. T. Huynh-Cam, L. S. Chen, and K. V. Huynh, "Learning Performance of International Students and Students with Disabilities: Early Prediction and Feature Selection through Educational Data Mining," *Big Data Cogn. Comput.*, vol. 6, no. 3, 2022, doi: 10.3390/bdcc6030094.

[31] H. Services, "Diarrhea : Common Illness , Global Killer," pp. 1–4.

[32] T. Kapwata, A. Mathee, W. J. Le Roux, and C. Y. Wright, "Diarrhoeal disease in relation to possible household risk factors in South African villages," *Int. J. Environ. Res. Public Health*, vol. 15, no. 8, 2018, doi: 10.3390/ijerph15081665.

[33] G. Debalkie, D. Id, Y. Y. Id, W. Aleminew, and Y. A. Id, “Diarrhea and associated factors among under five children in sub-Saharan Africa : Evidence from demographic and health surveys of 34 sub-Saharan countries,” vol. 24, pp. 1–13, 2021, doi: 10.1371/journal.pone.0257522.

[34] K. J. Begum and A. Ahmed, “The importance of statistical tools in research work,” *Int. J. Sci. Innov. Math. Res.*, vol. 3, no. 12, pp. 50–58, 2015, [Online]. Available: https://www.arcjournals.org/pdfs/ijsimr/v3-i12/10.pdf

[35] W. J. Lammers and E. Babbie, “Experimental Design : Statistical Analysis of Data,” *Fundam. Behav. Res.*, pp. 1–38, 2005.

[36] S. Angra and S. Ahuja, “Machine learning and its applications: A review,” *Proc. 2017 Int. Conf. Big Data Anal. Comput. Intell. ICBDACI 2017*, no. January, pp. 57–60, 2017, doi: 10.1109/ICBDACI.2017.8070809.

[37] B. Chandramohan, “Prediction and Prevention of Domestic Violence From Social Big Data Using Machine Learning Approach,” *Int. J. Pure Appl. Math.*, vol. 120, no. 6, pp. 3549–3561, 2018.

[38] J. Alzubi, A. Nayyar, and A. Kumar, “Machine Learning from Theory to Algorithms: An Overview,” *J. Phys. Conf. Ser.*, vol. 1142, no. 1, 2018, doi: 10.1088/1742-6596/1142/1/012012.

[39] R. P. Ram Kumar, S. Polepaka, S. F. Lazarus, and D. V. Krishna, “An insight on machine learning algorithms and its applications,” *Int. J. Innov. Technol. Explor. Eng.*, vol. 8, no. 11 Special issue 2, pp. 432–436, 2019, doi: 10.35940/ijitee.K1069.09811S219.

[40] N. Dhanda, S. S. Datta, and M. Dhanda, “Machine learning algorithms,” *Res. Anthol. Mach. Learn. Tech. Methods, Appl.*, vol. 11, no. 9, pp. 849–869, 2022, doi: 10.4018/978-1-6684-6291-1.ch044.

[41] EMC Education Services, “Data Science & Big Data Analytics,” *Data Sci. Big Data Anal.*, 2015, doi: 10.1002/9781119183686.

[42] M. Mohammed, M. B. Khan, and E. B. M. Bashie, *Machine learning: Algorithms and applications*, no. July. 2016. doi: 10.1201/9781315371658.

[43] H. Taherdoost, “Machine Learning Algorithms,” *Encycl. Data Sci. Mach. Learn.*, no. June, pp. 938–960, 2023, doi: 10.4018/978-1-7998-9220-5.ch054.

[44] L. Patel and K. A. Gaurav, “Introduction to Machine Learning and Its Application,” vol. 5,

no. 1, pp. 262–290, 2020, doi: 10.4018/978-1-7998-2718-4.ch014.

[45] M. N. Islam, S. N. Mustafina, T. Mahmud, and N. I. Khan, “Machine learning to predict pregnancy outcomes: a systematic review, synthesizing framework and future research agenda,” *BMC Pregnancy Childbirth*, vol. 22, no. 1, pp. 1–19, 2022, doi: 10.1186/s12884-022-04594-2.

[46] I. H. Sarker, “Machine Learning: Algorithms, Real-World Applications and Research Directions,” *SN Comput. Sci.*, vol. 2, no. 3, pp. 1–21, 2021, doi: 10.1007/s42979-021-00592-x.

[47] B. Concepts, D. Trees, and M. Evaluation, “Classification : Basic Concepts , Decision Trees , and”.

[48] L. Rokach and O. Maimon, “Chapter 9,” no. January, 2005, doi: 10.1007/0-387-25465-X.

[49] Y. Ben-Haim and E. Tom-Tov, “A streaming parallel decision tree algorithm,” *J. Mach. Learn. Res.*, vol. 11, pp. 849–872, 2010.

[50] P. T. R, “A Comparative Study on Decision Tree and Random Forest Using R Tool,” *Ijarcce*, vol. 4, no. 1, pp. 196–199, 2015, doi: 10.17148/ijarcce.2015.4142.

[51] S. Tangirala, “Evaluating the impact of GINI index and information gain on classification using decision tree classifier algorithm,” *Int. J. Adv. Comput. Sci. Appl.*, no. 2, pp. 612–619, 2020, doi: 10.14569/ijacsa.2020.0110277.

[52] A. G. Karegowda, A. S. Manjunath, and M. A. Jayaram, “Comparative Study of Attribute Selection Using Gain Ratio and Correlation Based Feature Selection,” *Int. J. Inf. Technol. Knowl. Manag.*, vol. 2, no. 2, pp. 271–277, 2010.

[53] V. Jakkula, “Tutorial on Support Vector Machine (SVM)”.

[54] S. Uddin, A. Khan, E. Hossain, and M. A. Moni, “Comparing different supervised machine learning algorithms for disease prediction,” vol. 8, pp. 1–16, 2019.

[55] A. Dey and A. S. Learning, “Machine Learning Algorithms : A Review,” vol. 7, no. 3, pp. 1174–1179, 2016.

[56] D. K. Srivastava and L. Bhambhu, “Data classification using support vector machine,” *J. Theor. Appl. Inf. Technol.*, vol. 12, no. 1, pp. 1–7, 2010.

[57] A. Pradhan, “SUPPORT VECTOR MACHINE-A Survey,” no. September 2012, 2017.

[58] H. Bhavsar and M. H. Panchal, “A Review on Support Vector Machine for Data Classification,” *Int. J. Adv. Res. Comput. Eng. Technol.*, vol. 1, no. 10, pp. 2278–1323,

2012.

[59] S. Karamizadeh, S. M. Abdullah, M. Halimi, J. Shayan, and M. J. Rajabi, “Advantage and drawback of support vector machine functionality,” *I4CT 2014 - 1st Int. Conf. Comput. Commun. Control Technol. Proc.*, no. I4ct, pp. 63–65, 2014, doi: 10.1109/I4CT.2014.6914146.

[60] I. Hanif, “Implementing Extreme Gradient Boosting (XGBoost) Classifier to Improve Customer Churn Prediction,” 2020, doi: 10.4108/eai.2-8-2019.2290338.

[61] C. Series, “Extreme gradient boosting (XGBoost) method in making forecasting application and analysis of USD exchange rates against rupiah Extreme gradient boosting (XGBoost) method in making forecasting application and analysis of USD exchange rates against rupi,” pp. 0–11, 2021, doi: 10.1088/1742-6596/1722/1/012016.

[62] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, “Catboost: Unbiased boosting with categorical features,” *Adv. Neural Inf. Process. Syst.*, vol. 2018-Decem, no. Section 4, pp. 6638–6648, 2018.

[63] J. Tanha, Y. Abdi, N. Samadi, N. Razzaghi, and M. Asadpour, “Boosting methods for multi-class imbalanced data classification: an experimental review,” *J. Big Data*, vol. 7, no. 1, 2020, doi: 10.1186/s40537-020-00349-y.

[64] A. A. Ibrahim, R. L. Ridwan, M. M. Muhammed, R. O. Abdulaziz, and G. A. Saheed, “Comparison of the CatBoost Classifier with other Machine Learning Methods,” *Int. J. Adv. Comput. Sci. Appl.*, vol. 11, no. 11, pp. 738–748, 2020, doi: 10.14569/IJACSA.2020.0111190.

[65] L. K. Smirani, H. A. Yamani, L. J. Menzli, and J. A. Boulahia, “Using Ensemble Learning Algorithms to Predict Student Failure and Enabling Customized Educational Paths,” *Sci. Program.*, vol. 2022, 2022, doi: 10.1155/2022/3805235.

[66] M. A. Ganaie, M. Hu, M. Tanveer*, and P. N. Suganthan*, “Ensemble deep learning: A review,” 2021.

[67] C. Qin, Y. Zhang, F. Bao, C. Zhang, P. Liu, and P. Liu, “XGBoost optimized by adaptive particle swarm optimization for credit scoring,” *Math. Probl. Eng.*, vol. 2021, 2021, doi: 10.1155/2021/6655510.

[68] P. Syam, S. Chand, and G. Divya, “A Light Gradient Boosting Machine Regression Model for Prediction of Agriculture Insurance Cost over Linear Regression,” pp. 200–208, 2022,

doi: 10.3233/APC220027.

[69] I. Polaka and I. E. Tom, “Decision Tree Classifiers in Bioinformatics,” no. May 2014, 2010, doi: 10.2478/v10143-010-0052-4.

[70] A. Cutler, D. R. Cutler, and J. R. Stevens, “Random Forests,” no. January, 2011, doi: 10.1007/978-1-4419-9326-7.

[71] J. Ali, R. Khan, N. Ahmad, and I. Maqsood, “Random Forests and Decision Trees,” *Int. J. Comput. Sci. Issues*, vol. 9, no. 5, pp. 272–278, 2012.

[72] J. Ali, R. Khan, N. Ahmad, and I. Maqsood, “Random Forests and Decision Trees,” no. December, 2013.

[73] S. Zhang, C. Zhang, and Q. Yang, “Data preparation for data mining,” *Appl. Artif. Intell.*, vol. 17, no. 5–6, pp. 375–381, 2003, doi: 10.1080/713827180.

[74] Z. S. Abdallah and G. I. Webb, “Encyclopedia of Machine Learning and Data Mining,” *Encycl. Mach. Learn. Data Min.*, no. January, 2017, doi: 10.1007/978-1-4899-7687-1.

[75] F. Nargesian, A. Asudeh, and H. V. Jagadish, “Responsible Data Integration: Next-generation Challenges,” *Proc. ACM SIGMOD Int. Conf. Manag. Data*, pp. 2458–2464, 2022, doi: 10.1145/3514221.3522567.

[76] Y. B. Wah, N. Ibrahim, H. A. Hamid, S. Abdul-Rahman, and S. Fong, “Feature selection methods: Case of filter and wrapper approaches for maximising classification accuracy,” *Pertanika J. Sci. Technol.*, vol. 26, no. 1, pp. 329–340, 2018.

[77] T. Hastie, “▶ Features the results of the NIPS 2003 workshop on feature extraction,” vol. 207, no. 0, p. 6221, 2006.

[78] I. Iguyon and A. Elisseeff, “An introduction to variable and feature selection,” *J. Mach. Learn. Res.*, vol. 3, no. April, pp. 1157–1182, 2003, doi: 10.1162/153244303322753616.

[79] B. Wu, M. Zhou, X. Shen, Y. Gao, R. Silvera, and G. Yiu, “Simple profile rectifications go a long way statistically exploring and alleviating the effects of sampling errors for program optimizations,” *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 7920 LNCS, no. 97, pp. 654–678, 2013, doi: 10.1007/978-3-642-39038-8-27.

[80] S. Mamman, A. Adamu, Y. Ado, and R. Muhammad, “An Overview of the Algorithm Selection Problem,” *Int. J. Comput.*, vol. 26, no. 1, pp. 89–98, 2017.

[81] P. H. C. Chen, Y. Liu, and L. Peng, “How to develop machine learning models for

healthcare," *Nat. Mater.*, vol. 18, no. 5, pp. 410–414, 2019, doi: 10.1038/s41563-019-0345-0.

[82] G. Varoquaux and O. Colliot, "value To cite this version : Evaluating machine learning models and their diagnostic value," 2023.

[83] W. Yip, "Lifecycle of machine learning models 1 2," 2020.

[84] N. Lachiche, "Receiver Operating Characteristic (ROC) Analysis," *Encycl. Data Warehous. Mining, Second Ed.*, vol. 5, no. 3, pp. 1675–1681, 2011, doi: 10.4018/978-1-60566-010-3.ch255.

[85] X. Ying, "An Overview of Overfitting and its Solutions," *J. Phys. Conf. Ser.*, vol. 1168, no. 2, 2019, doi: 10.1088/1742-6596/1168/2/022022.

[86] D. Berrar, "Cross-validation," *Encycl. Bioinforma. Comput. Biol. ABC Bioinforma.*, vol. 1–3, no. January 2018, pp. 542–545, 2018, doi: 10.1016/B978-0-12-809633-8.20349-X.

[87] G. C. Cawley and N. L. C. Talbot, "Efficient approximate leave-one-out cross-validation for kernel logistic regression," *Mach. Learn.*, vol. 71, no. 2–3, pp. 243–264, 2008, doi: 10.1007/s10994-008-5055-9.

[88] R. R. Nadikattu, "Fundamental Applications of Machine," vol. 6, no. 1, pp. 31–40, 2018.

[89] S. J. Cunningham, J. Littin, and I. H. Witten, "A PPLICATIONS OF MACHINE LEARNING IN INFORMATION RETRIEVAL 1 Introduction," *Inf. Retr. Boston.*.

[90] N. A. K. Rosili, N. H. Zakaria, R. Hassan, S. Kasim, F. Z. C. Rose, and T. Sutikno, "A systematic literature review of machine learning methods in predicting court decisions," *IAES Int. J. Artif. Intell.*, vol. 10, no. 4, pp. 1091–1102, 2021, doi: 10.11591/IJAI.V10.I4.PP1091-1102.

[91] B. Sekeroglu, R. Abiyev, A. Ilhan, M. Arslan, and J. B. Idoko, "Systematic literature review on machine learning and student performance prediction: Critical gaps and possible remedies," *Appl. Sci.*, vol. 11, no. 22, 2021, doi: 10.3390/app112210907.

[92] D. Salcedo *et al.*, "Machine Learning Algorithms Application in COVID-19 Disease: A Systematic Literature Review and Future Directions," *Electron.*, vol. 11, no. 23, 2022, doi: 10.3390/electronics11234015.

[93] A. Raj, N. Dehingia, A. Singh, J. McAuley, and L. McDougal, "Machine learning analysis of non-marital sexual violence in India," *EClinicalMedicine*, vol. 39, p. 101046, 2021, doi: 10.1016/j.eclinm.2021.101046.

[94] T. B. Ayuk *et al.*, “Prevalence of diarrhoea and associated risk factors among children under-five years of age in Efoulan health district- Cameroon, sub-Saharan Africa,” *MOJ Public Heal.*, vol. 7, no. 6, pp. 259–264, 2018, doi: 10.15406/mojph.2018.07.00248.

[95] D. Mosisa, M. Aboma, T. Girma, and A. Shibru, “Determinants of diarrheal diseases among under five children in Jimma Geneti District , Oromia region , Ethiopia , 2020 : a case-control study,” *BMC Pediatr.*, pp. 1–13, 2021, doi: 10.1186/s12887-021-03022-2.

[96] C. Town, S. Africa, K. Carden, and M. A. Dalvie, “Diarrhoea among Children Aged under Five Years and Risk Factors in Informal Settlements : A Cross-Sectional Study in,” 2021.

[97] K. Ghosh, A. Sinha, and M. Mog, “Prevalence of diarrhoea among under five children in India and its contextual determinants : A geo-spatial analysis,” *Clin. Epidemiol. Glob. Heal.*, vol. 12, no. March, p. 100813, 2021, doi: 10.1016/j.cegh.2021.100813.

[98] D. Bekele and E. Merdassa, “Determinants of Diarrhea in Under-Five Children Among Health Extension Model and Non-Model Families in Wama Hagelo District , West Ethiopia : Community-Based Comparative Cross-Sectional Study,” no. October, 2021.

[99] E. T. Solomon, S. R. Gari, H. Kloos, and B. Mengistie, “Diarrheal morbidity and predisposing factors among children under 5 years of age in rural East Ethiopia,” *Trop. Med. Health*, vol. 48, no. 1, 2020, doi: 10.1186/s41182-020-00253-4.

[100] H. Lanyero *et al.*, “Antibiotic use among children under five years with diarrhea in rural communities of Gulu , northern Uganda : a cross-sectional study,” pp. 1–9, 2021.

[101] D. N. Gessesse and A. A. Tarekegn, “Prevalence and associated factors of diarrhea among under-five children in the Jawi district , Awi Zone Ethiopia , 2019 . Community based comparative cross-sectional study,” no. August, pp. 1–9, 2022, doi: 10.3389/fped.2022.890304.

[102] R. M. Hartman *et al.*, “Risk Factors for Mortality Among Children Younger Than Age 5 Years With Severe Diarrhea in Low- and Middle-income Countries : Findings From the World Health Organization-coordinated Global Rotavirus and Pediatric Diarrhea Surveillance Networks,” *Clin. Infect. Dis.*, vol. 76, no. 3, pp. 1047–1053, 2023, doi: 10.1093/cid/ciac561.

[103] N. H. Son, “Data cleaning and Data preprocessing,” 2011, [Online]. Available: http://www.mimuw.edu.pl/~son/datamining/DM/4-preprocess.pdf

[104] R. Barandela, J. S. Sánchez, V. García, and E. Rangel, “Strategies for learning in class

imbalance problems," *Pattern Recognit.*, vol. 36, no. 3, pp. 849–851, 2003, doi: 10.1016/S0031-3203(02)00257-1.

[105] N. A. Nnamoko, F. N. Arshad, D. England, J. Vora, and J. Norman, "Evaluation of Filter and Wrapper Methods for.pdf," no. September, 2014.

[106] M. Birgersson, G. Hansson, and U. Franke, "Data Integration Using Machine Learning," *Proc. - IEEE Int. Enterp. Distrib. Object Comput. Work. EDOCW*, vol. 2016-Septe, no. September 2016, pp. 313–322, 2016, doi: 10.1109/EDOCW.2016.7584357.

[107] N. Silva and C. Mena, "Identifying the underlying risk factors of local communities in Chile," *Disaster Prev. Manag. An Int. J.*, vol. 29, no. 5, pp. 681–696, 2020, doi: 10.1108/DPM-04-2020-0105.

[108] J. Bergstra, "Algorithms for Hyper-Parameter Optimization Algorithms for Hyper-Parameter Optimization," no. December, 2011.

[109] M. Dash and H. Liu, "Feature selection for classification," *Intell. Data Anal.*, vol. 1, no. 3, pp. 131–156, 1997, doi: 10.3233/IDA-1997-1302.

[110] S. Maheshwari, "A Review on Class Imbalance Problem: Analysis and Potential Solutions," *Int. J. Comput. Sci. Issues*, vol. 14, no. 6, pp. 43–51, 2017, doi: 10.20943/01201706.4351.

[111] T. Aljuaid and S. Sasi, "Proper imputation techniques for missing values in data sets," *Proc. 2016 Int. Conf. Data Sci. Eng. ICDSE 2016*, no. September 2017, 2017, doi: 10.1109/ICDSE.2016.7823957.

[112] L. Himmelspach and S. Conrad, "The effect of noise and outliers on fuzzy clustering of high dimensional data," *IJCCI 2016 - Proc. 8th Int. Jt. Conf. Comput. Intell.*, vol. 2, no. Ijcci, pp. 101–108, 2016, doi: 10.5220/0006070601010108.

[113] Y. Li and A. Ngom, "Data integration in machine learning," *Proc. - 2015 IEEE Int. Conf. Bioinforma. Biomed. BIBM 2015*, no. November 2015, pp. 1665–1671, 2015, doi: 10.1109/BIBM.2015.7359925.

[114] F. Muhlenbach, R. Rakotomalala, F. Muhlenbach, R. Rakotomalala, C. Attributes, and J. Wang, "Discretization of Continuous Attributes To cite this version : HAL Id : hal-00383757 Discretization of Continuous Attributes," pp. 397–402, 2009.

[115] M. Hasnain, M. F. Pasha, I. Ghani, M. Imran, M. Y. Alzahrani, and R. Budiarto, "Evaluating Trust Prediction and Confusion Matrix Measures for Web Services Ranking," *IEEE Access*, vol. 8, pp. 90847–90861, 2020, doi: 10.1109/ACCESS.2020.2994222.

[116] R. Blagus and L. Lusa, “SMOTE for high-dimensional class-imbalanced data,” *BMC Bioinformatics*, vol. 14, 2013, doi: 10.1186/1471-2105-14-106.

[117] S. Maldonado, R. Weber, and F. Famili, “Feature selection for high-dimensional class-imbalanced data sets using Support Vector Machines,” *Inf. Sci. (Ny).*, vol. 286, pp. 228–246, 2014, doi: 10.1016/j.ins.2014.07.015.

[118] M. Koziarski, “CSMOUTE: Combined Synthetic Oversampling and Undersampling Technique for Imbalanced Data Classification,” pp. 1–8, 2021, doi: 10.1109/ijcnn52387.2021.9533415.

[119] A. Rácz, D. Bajusz, and K. Héberger, “Effect of dataset size and train/test split ratios in qsar/qspr multiclass classification,” *Molecules*, vol. 26, no. 4, pp. 1–16, 2021, doi: 10.3390/molecules26041111.

[120] B. H. Shekar and G. Dagnew, “Grid search-based hyperparameter tuning and classification of microarray cancer data,” *2019 2nd Int. Conf. Adv. Comput. Commun. Paradig. ICACCP 2019*, no. February, pp. 1–8, 2019, doi: 10.1109/ICACCP.2019.8882943.

[121] H. M and S. M.N, “A Review on Evaluation Metrics for Data Classification Evaluations,” *Int. J. Data Min. Knowl. Manag. Process*, vol. 5, no. 2, pp. 01–11, 2015, doi: 10.5121/ijdkp.2015.5201.