African Centre of Excellence in Data Science

College of Business and Economics University

of Rwanda

Topic: Machine learning based prediction of malaria outbreak using environment data in Rwanda

A Dissertation is submitted in full fulfilment to the University of Rwanda in accordance with requirements for the degree of master's degree of data science majoring in Biostatistics in Africa Centre of Excellence in Data Science

By Albert DUKUZUMUREMYI

Reg Number: 219013798

Supervisor: Pierre Claver RUTAYISIRE, PhD

September 2020

DECLARATION

I declare that this dissertation entitled **“Machine learning based prediction of malaria outbreak using environment data in Rwanda”** is the result of my own work and has not been submitted for any other degree at the University of Rwanda or any other institution.

Student Name: Albert DUKUZUMUREMYI

Signature:

APPROVAL SHEET

This dissertation entitled **‘Machine learning based prediction of malaria outbreak using environment data in Rwanda”** written and submitted by Albert DUKUZUMUREMYI in partial fulfilment of the requirements for the degree of Master of Science in Data Science majoring in Biostatistics is hereby accepted and approved. The rate of plagiarism tested using Turnitin is 19% which is less than 20% accepted by ACE-DS.

____
Dr. Pierre Claver RUTAYISIRE
Supervisor

____
Dr. Ignace KABANO
Head of Training

ACKNOWLEDGEMENT

This work is the result of the collective effort of many people to whom I would like to express my gratitude.

First and foremost, I thank Almighty God for grace and unindenting love and for keeping me safe along this journey of master program education of data science.

My profound gratitude goes to the University of Rwanda, African Centre of Excellence in Data Science for giving me the opportunity to get the best knowledge of data science by the support of the Experts around the World.

I would like to express my gratitude to Ministry of Education for enabling the environment for learning, your outstanding support along the years has been a motivation in this journey.

Furthermore, I would like to express my deepest gratitude to my research project supervisor, Pierre Claver RUTAYISIRE, PhD for his thoughtful guidance and cooperation which help me to conduct this research. The outcome of this research project is due to his expert assistance and constructive guidance which should not be forgotten.

My Special thanks to lecturers and classmates, they have been a source of inspiration to acquiring skills of data science and to hard-working. Your collaboration has been the keystone to making this program more successful.

ABSTRACT

Malaria is still one of the common diseases that cause a threat to human population globally, 2019 report of the World Health Organization (WHO), indicated that an estimated 228 million malaria cases were found in 2018 and in 2017 the estimated malaria cases were 231 million. Many studies have highlighted the environment factors as the contributing factors to the variability of malaria prevalence in different regions. Different study revealed that climate factors including temperature, humidity and rainfall should play an important/leading role in the model prediction of malaria outbreak. Furthermore, there is a need for intelligent predictive systems using machine learning techniques which can predict the malaria outbreak, based on historical data on malaria and environment factors, this study was based on building a predictive model of malaria outbreak which should be used as a warning system to health care providers, hospitals, and other health institutions to give a warning on an occurrence of malaria outbreak based on meteorological data. Historical data from 2016 to 2019 of malaria cases from Rwanda Biomedical Center (RBC) and meteorological data including maximum and minimum temperature, rainfall, elevation, latitude ,and longitude from Rwanda Meteorological Agency of ten districts including Nyarugenge, Kicukiro, Nyamagabe, Huye, Gicumbi, Musanze, Karongi, Rusizi, Kayonza, and Nyagatare that constitute the total of 2080 observations have been used to fit six machine learning algorithms including Decision tree, Random Forest, Naïve Bayes, Support vector machine, K-nearest neighbour, and Logistic regression thus the best algorithm model will be selected based on performance metrics of Accuracy, Precision, Recall, F-score and ROC have been applied on each machine learning algorithms to evaluate their performance. Among those classifiers, Random Forest comes with high performance compared to others with more than 90% in all evaluation metrics; it has shown the accuracy of 90.75%, F-score of 90.73%, Precision of 90.69% and Recall at 90.88%. However, the classifiers also have shown the high performance except for Support vector machine which shown only around 60% in all evaluation metrics, but other classifiers scored above 70% on the evaluation metrics. According to this study for malaria outbreak, it is recommended to use Random Forest for malaria prediction. The use of machine learning algorithms models for prediction of malaria outbreak proved to be used as alarming system for health-based, health care providers, and health practitioners thereby they can be well prepared and set up the new prevention measures.

KEY WORDS

a. Machine learning

b. Malaria outbreak

c. Classifiers

d. Support vector machine

e. Logistic Regression

f. Naïve Bayes

g. Decision tree

h. Random Forest

i. K-Nearest Neighbour

j. Training set

k. Test set

LIST OF ABBREVIATION AND ACCRONYMS

WHO:	World Health Organization
GTS:	Global Technical strategies
SDG:	Sustainable Development Goals
LLIN:	Long-lasting insecticidal-treated nets
IRS:	Indoor residual Spraying
ACT:	Artemisinin Combination Therapy
CHWs:	Community Health Workers
RBC:	Rwanda Biomedical Centre
EMR:	Electronic Medical Records
ANN:	Artificial Neural Network
IPTp:	Intermittent preventive treatment of pregnant women
AR:	Auto regressive model
ARIMA:	Auto- Regressive Moving average
SVM:	Support Vector Machine
NB:	Naïve Bayes
KNN:	K-Nearest Neighbour
ROC:	Receiver Operating Characteristic
AUC:	Area Under curve
AUC:	Area Under curve
HMIS:	Health Management Information System
TP:	True Positive
FP:	False Positive
TN:	True Negative
FN:	False Negative

Table of Contents

List of figures

List of tables

1 INTRODUCTION

This chapter introduces the idea on malaria it goes on historical background of malaria and various interventions have been made to eradicate malaria in endemic regions, it shows the malaria picture globally, regionally and its distribution in Rwanda. Moreover, this chapter includes statement of the problem of malaria and shows the contribution of the technology in prevention and control of malaria epidemic, it talks about the research objectives, scope of the study, significance of the study and organization of the study which shows the flow of the study.

1.1 Background of the study

Worldwide malaria is still one of life-threatening for human kind[1], the 2019 report of the World Health Organization (WHO), indicated that globally an estimated 228 million malaria cases were found in 2018 and in 2017 the estimated malaria cases were 231 million. Malaria is highly prevalent in Africa region where the estimated 213 million malaria cases have been found in Africa which constitutes 93% of total cases, followed by Eastern Asia accounted for 5.4% of total cases and then Eastern Mediterranean with 2.1% of all cases. Malaria has a high prevalence in Sous-Sahara Africa whereby only six countries of Sous-Sahara Africa countries carried more than half of the malaria cases globally, those countries include Nigeria accounted for 25%, RDC 12%, Uganda 5%, Niger, Ivory Coast and Mozambique carried 4% per each[2]

However, Malaria has been declined over the years through various strategies have been initiated to controlling and preventing of malaria such that from 2010 to 2018, malaria incidence has been declined from 71 to 57 over 1000 population at risk globally. In Africa, the incidence rate shifted from 294 in 2010 to 229 in 2018 which constitute 22% of incidence reduction globally. Regarding malaria-related deaths; Malaria is among the primary causes of mortality in the World, where 405,000 deaths caused by malaria in 2018, in 2010 were 585,000 deaths, the under-five year's children are the most vulnerable group because they counted 67% of malaria deaths globally. Africa is the most vulnerable region because 94% of global malaria-related deaths have been accounted for African countries Despite the high malaria mortality in Africa, Africa has recognized the great decline of malaria deaths whereby from 2010 to 2018, the malaria deaths reduced from 533 000 to 380000 deaths[2].

In Rwanda, according to the 2017 malaria indicator survey, malaria prevalence was generally at 7% of all population, the eastern province presents the highest prevalence with 13%, and the Northern province accounted for the lowest prevalence of 1%, the prevalence among under five years children is at 7% of all children disproportionately distributed with their family's wealth quintile whereby it accounted for 13% in lowest wealth quintile and 2% of the highest wealth quintile. Malaria prevalence among the children in the age of 5 to 14 years is 11% and the children of rural area present high prevalence (13%) compared to those of urban area (3%) and the prevalence of individual aged 15 and above is 6% with 12% in Eastern province versus 1% in Northern Province. Among women between 15-49 years, malaria prevalence accounted for 5% and the women with no education the prevalence is at 6% versus 3% among women who attained secondary or higher education[3]. Malaria was responsible for 419 malaria-related deaths in 2013 and 715 deaths in 2016[4].

Malaria is transmitted by infected female mosquitoes named anopheles and transferred to a human being through a bite of infected anopheles on the human body, the anopheles 'bites mainly happened between sunset and sunrise; it takes around 45 minutes after biting of mosquito to spread in human blood bite by the mosquito. Globally there are around 600 groups of Anopheles mosquitoes but among them, only 60 groups of anopheles are natural malaria transmitters and only 30 species are more predominant[5]. Plasmodium falciparum, Plasmodium malariae, Plasmodium vivax, and Plasmodium ovale are the four public species of malaria[67].

The most prevalent malaria in Africa regions is the Plasmodium falciparum whereby it accounted for 99.7% of Africa malaria cases in 2018, in south-East Asia, was accounted for 50%, Eastern-Mediterranean region was at 71% as well in western Pacific region was at 65%. Most Plasmodium Vivax cases are predominant in South-East Asia with 53% and its majority is presented in India with 47% of all cases[2]. Plasmodium falciparum is considered as the most severe form of malaria; its clinical features include fever, chills, muscular aching, headache, diarrhoea, vomiting cough, and abdominal pain. Additionally, the symptoms will follows caused by the weakness of the organ, such as pulmonary oedema, renal failure, pulmonary oedema, generalized convulsions, circulatory collapse followed by coma and death. However, if the treatment of Plasmodium falciparum exceeds 24 hours after the onset of clinical symptoms it might be fatal[8].

.

1.1.1 Surveillance for malaria

It is urgent to strengthening the measure for declining malaria morbidity and mortality all over the world by increasing the number of countries, territories, and areas free of malaria by highlighting different approaches aiming to reduce malaria transmission. WHO has initiated Global Technical Strategy for malaria (GTS) from 2016 to 2030 and the main objective was to achieve the elimination of malaria globally, the pillars of GTS include "to ensure universal access to malaria prevention, diagnosis and treatment"; "to accelerate effort towards elimination and attainment of malaria-free status" and "transform malaria surveillance into core intervention". Moreover, Global Technical Strategy (GTS) for malaria also has underlined various goals to be achieved before the end of 2030 comprise "reducing malaria mortality rates globally at least 90% compared with 2015"; "reducing malaria cases incidence globally at least 90% compared with 2015"; "eliminate malaria in at least 35 countries in which malaria was transmitted in 2015", and "Prevent re-establishment of malaria in all countries that are malaria-free"[6,9,].

However, the various prevention measures of malaria are recommended by the World Health Organization (WHO) include "prompt diagnostic testing", "effective treatment" and "following of malaria cases to complement available scaled-up tools, including universal coverage of long-lasting insecticidal-treated nets (LLINs) and indoor residual spraying (IRS) with insecticide and treatment of malaria cases with artemisinin combination therapy (ACT) in malaria-endemic countries"[10].

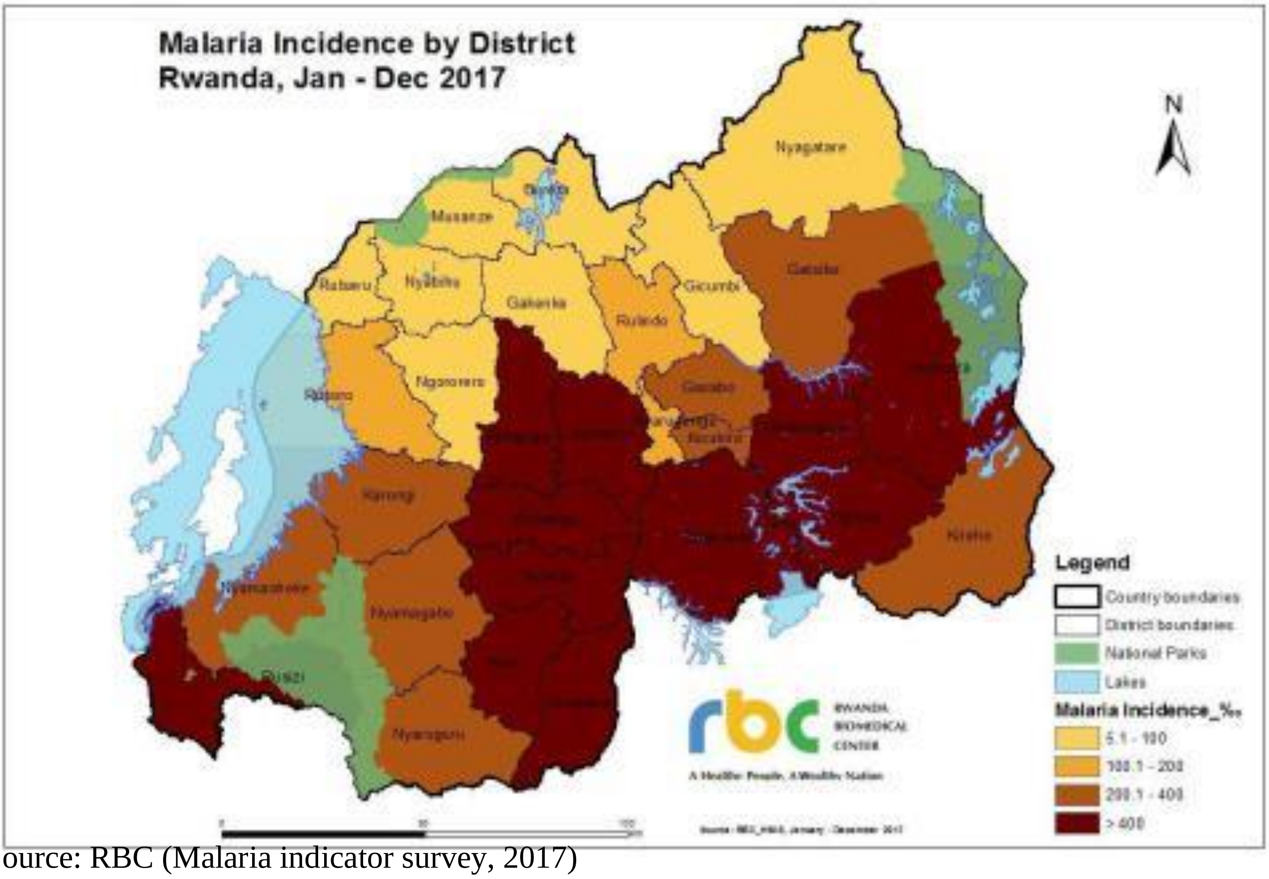
Rwanda has put in place different initiatives to fighting against malaria; among them: Rwanda has extended community-based treatment of malaria from September 2016 by including children with more than five years and adults who give an increase to 56% of all malaria diagnosis treatments are done by community health workers (CHWs) and by November, Government of Rwanda granted free diagnosis treatment to all households belong to Ubudehe 1 and 2 categories. Furthermore, in late 2016 and the beginning of 2017 Government of Rwanda has also distributed in mass distribution campaign, a total of more than five million ITNs and in 2016, IRS was inserted with organophosphate insecticide beginning in September 2016 to pre-empt resistance to carbamate insecticides and implementing expanded from three to five districts[4].

1.1.2 Malaria with environment factors

There are many risk factors contributing to the variability in malaria prevalence in different regions including population immunity, mosquito control measures, social and economic status and environmental factors including climate factors (temperatures, rainfall, relative humidity, etc.) elevation, longitude latitude among others. The environmental factors have a significant impact on the transmission of malaria by giving the favourable environment for breeding of mosquitoes[11]. However, an increase in temperature leads to the acceleration of mosquito metabolic rate, an increase in eggs production and more frequent of blood-feeding. The rainfall has an indirect consequence on longevity and by the effect, it has on humidity, because the relative wet condition creates a favourable habitation for mosquitoes which lead to an increase of them in the geographical area and its seasonal abundance of mosquitoes which are malaria vectors. The excessive rain negatively affects the disease vectors because flooding washes away the breeding site[12].

In Rwanda, the increase in temperature lead to an increase in the probability of the transmission of malaria, it is not only the temperature influence the spaces of anopheles mosquitoes occupies but also the presence of rain influence the variability of malaria prevalence, the malaria vector spend their life cycle wet habitat, therefore, and the change in mosquitoes quantity depends on hydrological change from rainy season to dry season with the decisive criterion for insect's development being linked to the duration of wet phase with intervening of the dry period. Moreover, the high humidity affects the metabolism system of anopheles[13]. According to Rwanda Biomedical Center (RBC), 19 districts have been classified as high endemic regions because of climate, altitude, and high prevalence, hence both eastern and southern province accounted for 79% of disease burden and 11 districts accounted for 59% of malaria cases. Also, high transmission of malaria occurred twice a year: one from May to June and another from November to December following to the Rwanda rainy season which connects the impact of climate conditions on the variability of malaria in Rwanda[14]. The following Rwanda Map shows the distribution of incidence of malaria among the districts of Rwanda:

Figure 1:Malaria incidence by District of Rwanda



Source: RBC (Malaria indicator survey, 2017)

Figure 1 Indicate the distribution of malaria incidence among the districts of Rwanda, it shows that there is a high incidence rate in southern and Eastern Province such that ten districts in South and Eastern province have registered incidence of more than 400 over 1000 population at risk in 2017. It also shows the low malaria incidence is found among the northern province districts, whereby most districts presented incidence, ranged between 5.1 to 100 per 1000 population at risk. This low malaria incidence is explained by low temperature ever found in North province which is negatively affect living condition

Given the association between malaria transmission and climate, towards the control and prevention of malaria, it is very essential to take care of the contribution of that climate and environmental factors to malaria prevalence. Therefore, for retaining and strengthening the surveillance system for malaria control, long period data quality data for malaria are very needed to build the models, which relate climate to malaria[14]. It is expected that without the mitigation, climate change will result in an increase in malaria burden in many malaria-endemic regions especially in highly densely populated tropical highlands[9].

1.1.3 Application of machine learning in health sector

Machine learning is used in many fields and sometime the users are not aware that they are using it, Application of machine learning is applied in different domains including face recognition, speech recognition, disease prediction, self-driving cars, web search, and anomaly detection among others[15]. Artificial intelligence (AI) has begun to create change in healthcare across developed markets and has potential to drive game-changing improvements for underserved communities in global health. From enabling community-health workers to better serve patients in remote areas to helping governments in low and middle-income countries (LMICs) prevent deadly disease outbreaks before they occur, there is growing recognition of the tremendous potential of AI tools to break fundamental trade-offs in health access, quality, and cost[15]. Furthermore, machine learning is widely used in checking health condition, in diagnosis of different disease such as cancer and in prediction of occurrence of disease. In pharmacology, machine learning is used to find the right formula and reliable drugs to cure for certain disease and it is widely used in choosing the effective therapeutic treatment. Among machine learning techniques the artificial neural network (ANN) has proved to be very powerful due the black box and volatile learner concepts and deep learning has attained the exponential growth due the ability of insightful decision making therefore the Deep learning results to high-level abstraction in data. Deep learning also provides an improved performance when dealing with irregular and non-stationary time series data because it discovers and characterize the complex features of data set and back-propagation of neural network is one of the most used methods in deep learning. However, decision trees have high ability in discovery, in accuracy, preciseness and reliability[16].

Furthermore, one of the applications of machine learning in health care used for the processing of large and complex dataset so that the clinical insights found through the health data, the initiation of machine learning in health care also led to the increase of patient satisfaction. Moreover, machine learning is largely used in disease prediction whereby by using machine learning predictive algorithms the disease prediction made possible then health care be smarter. Hence, by implementation of machine the prediction of disease and epidemic outbreak lead to its early initiation of preventive measures[17].

There is a growing interest in building predictive model for malaria prediction to give support for clinical and public health practitioners to strategically implementation of the prevention and control measures ahead of time. Machine learning algorithms have been very successful by characterized with high performance metrics compared with the traditional methods; also, machine learning models do not require the deep knowledge of statistics[1].

1.2 Problem statement

The effect of malaria is more profound in developing countries where there are limited medical resources, equipment, and hospital facilities. In developing countries poverty economic instability are still impeding the prevention and reduction of risk factors of the malaria to the population[18]. Various preventive and treatment measures for controlling malaria cases have been put in place include insecticide-treated mosquito nets (ITNs); indoor residual spraying (IRS); accurate diagnosis and prompt treatment with artemisinin-based combination therapies (ACTs); and intermittent preventive treatment of pregnant women (IPTp)[4]. Study has shown that Malaria elimination requires the robust health systems where the community access to quality health services, health information systems for tracking progress, and system for public health response[17]. Despite various malaria control and preventive measures that have been put in place, malaria still pose threats to human life, however, one of the efficient ways for malaria prevention is the development predictive models with high performance which can make the early prediction of malaria outbreak and help health players to take prevent and control measures in advance.

Furthermore, nowadays there is increased using malaria prediction models to support the clinical and public health services to implement the preventive and control measures[17]. Previously, several traditional methods of prediction of malaria outbreak have been utilized such as season forecast model, Auto Regressive Model (ARM), Auto- Regressive Moving average (ARIMA); with all prediction models there was always a problem of low accuracy of those they require a lot of time in data analysis[19]. Therefore, there is an urgent need for intelligent predictive systems using machine learning techniques which predict the malaria outbreak, based on historical data on malaria with high precision and accuracy. This study focuses on building a predictive model of malaria outbreak using the environment data, it would be used as a warning system to health care providers, hospitals, health-based institutions as well health care practitioners to give alert whether the has been an increase of malaria depending on the environment data occurred. As the environment factors have the significant effect on malaria, machine learning models trained using environment data will lead a model with good performance metrics which will be used to predict the period of malaria outbreak depending on presence of environment condition. The well-defined malaria outbreak parameters are sufficient to predict the outbreak and with comparing the traditional methods. The support vector machine (SVM), Naïve Bayes, Decision tree, and other classification algorithms are the classifications of machine learning techniques which are largely used in health care decision as decision support techniques[19].

1.3 Objective of the study

The main objective of the study is to build machine learning predictive model of malaria outbreak using environment data.

The specific objectives of this study are the following:

- To apply machine learning techniques in prediction, prevention, and control of malaria outbreak by using environment data
- To identify the contribution of environment factors to the persistence of malaria in Rwanda.
- To develop an early warning system, which can predict malaria outbreak thus, set up an early response.

1.4 Scope of the study

The scope of this study is as follow:

- The dataset of monthly malaria cases in Rwanda of 2016 to 2019 from the Rwanda Biomedical Center (RBC) and the data set of meteorological data composed by weekly average maximum and minimum temperature and weekly rainfall from 2016 to 2019, elevation, latitude, and longitude of ten selected districts got from Rwanda Meteorological Agency merged in one dataset.
- The combined dataset has been extracted into Python 3.7 programming for being pre-processed then cleaned for the errors, missing and outliers, thereafter, various classification algorithms include Support vector machine (SVM), Random Forest, Naïve Bayes, Logistic Regression, K-Nearest Neighbour and Decision tree have been trained on the dataset of malaria cases and these environment data.
- The performance metrics of accuracy, recall, precision, and F1- score as well as Receiver Operating Characteristic (ROC) curve and Area under curve applied to evaluate performance of these machine-learning algorithms and the best model recommended used in prediction of malaria outbreak.
- The data processed in Python3.7 programming language Jupiter notebook

1.5 Limitation of the study

This study is only limited to prediction of malaria outbreak using environment data by using machine learning algorithms; classification algorithms, it has used historical malaria cases data and environment data from ten districts for 2016-2019 period which have been sampled based on the convenience of the data, it has also used meteorological factors. This study has used six classification machine learning algorithms for predicting the outbreak of malaria and the performance of model has been evaluated using Recall, Precision, Accuracy, F score and ROC.

1.6 Significance of the study

Nowadays there is an increase of using computer-based technology such as machine learning application in health sector such in diagnosis and treatment of disease such as cancer, machine learning also plays an indispensable role in prediction of risk factors associated to the presence of certain disease. This study makes possible prediction of malaria outbreak by considering the environment data and past malaria cases using machine learning techniques and the best algorithm based on accuracy and other performance metrics selected to be used for malaria outbreak prediction. This study gives contribution to health care players, hospitals, and health organization to aware the future occurrence of malaria outbreak so that they may take more precautions and control measure in advance to save people lives.

1.7 Organization of the study

This study is made by five chapters. Chapter one contains the background of the study, the problem statement, objectives of the study, scope, and significance of the study. Chapter two is made by the literature on prediction of malaria using machine learning and the basic concepts of machine learning. Chapter three assess the methodology used to extract data, train, and build model and evaluation of machine learning algorithms used in the study. Chapter four presents the results; analyses building of machine learning algorithms and discussion of the findings got using the stated methodology. Chapter five concludes the study and gives the recommendations to focus on based on findings got.

2 LITERATURE REVIEW

This part details the concepts of machine learning techniques, it defines the theories of machine learning and malaria. This part also takes review assess literatures written on prediction of malaria using traditional methods versus malaria prediction using machine learning techniques.

2.1 Definition of key terms

2.1.1 Machine learning

Machine learning is a branch of artificial intelligence, and it is software which works like brain of human being, machine learning learns from the data and applies it in making decision. It can improve software and its ability to solve problem through gaining experience like human memory. Machine learning also defines as a study which existed in artificial intelligence that uses probabilistic, statistical and optimization techniques to train computer system to scrutinize and learn discern and hard patterns in complex, large and noisy data. It is learning how to perform better based on experience, for example the prediction of disease based on past observations and the goal is to create an algorithm which can learn automatically without any of human assistance or intervention. Machine learning uses different algorithms that learn from data to improve, describe data, and predict outcomes. As the machine learning algorithms use training data that make possible to produce more precise models based on those data and machine learning model defined as the result generated when we train machine learning algorithm with data[20].

Machine learning is widely applied in medical sector because it provides an alternative solution to medical problem by using various techniques such as clustering and classification which take past data to predict the status disease and this approach was inspired many researchers trying to use medical data to predict disease[7].

Machine is required to increase the accuracy to increase the predictive models. There are various machine learning approaches depending on problem being addressed based on type and the volume of the data. Those approaches are supervised learning; unsupervised learning; Reinforcement learning and Deep learning. The Supervised learning approaches has two algorithms' categories include Regression and classification.

2.1.2 Classification

Classification is supervised machine learning technique which used to forecast the group membership for data instance, it is widely used among other techniques of machine learning[21]. The classification algorithm is an algorithm that receive training set and learn from a function of the form f: Rn→{+1,−1}. This function will be applied to new inputs and the class where the inputs will be classified. The common classification algorithms are: Naïve Bayes, Decision tree, Fisher Linear Discriminant and Support Vector machine (SVM), Logistic regression and K-Nearest Neighbour. The performance of classification algorithm depends greatly on the characteristics of the data to be classified; there is no single classifier works better in all problems. The measure of precision and recall are most popular metric used to evaluate the quality of classification system, recently, receiver operating curve (ROC) is used to evaluate the true positive rate and true negative rate[22].

2.1.3 Malaria

According to WHO, Malaria is a life-threatening disease caused by parasites that are transmitted to people through the bites of infected female Anopheles mosquitoes. It is preventable and curable. Malaria is more severe on some category of people and those groups are considered as the riskier than others, those groups include infants, children under 5 years of age, pregnant women and patients with HIV/AIDS, as well as non-immune migrants, mobile populations and travellers[23].

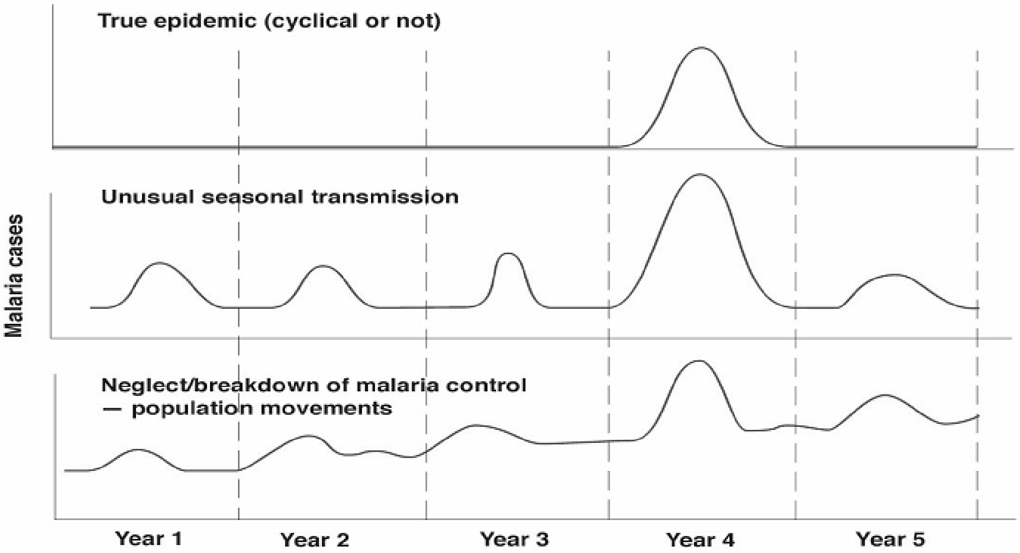
2.1.4 Malaria cases

Malaria case is the presence of malaria sickness or disease in human being and that presence is confirmed by parasitological testing. Based on the origin of the infection, malaria cases can be classified as autochthonous, indigenous, induced, introduced/imported, or relapsing. A suspected case is not taken as malaria case until the parasitological confirmation. In the place of malaria control setting, malaria case is the occurrence of conformed malaria infection while in the region of malaria elimination, malaria case is the occurrence of confirmed malaria infection with or without the symptoms[24].

2.1.5 Malaria outbreak

Until today there is no universal definition of a malaria outbreak/epidemic. Generally, Malaria is defined as sudden increase in malaria incidence among populations in which the disease is rare or a seasonal increase in community is subject in clinical malaria in areas of low to moderate transmission constitute a malaria epidemic. Even though it is more complicated to give the definition of "normal" occurrence it can be defined only for a particular population in a specific area and time. Thus, malaria epidemics can generally be considered as a disturbance of the existing epidemiological equilibrium. Malaria outbreak can be also defined as the excess increase in malaria cases which surpass the existing health facilities to handle it[25].

Different efforts have been made to put the thresholds of malaria outbreak which define the outbreak based on the experience of malaria disease. Yet those thresholds can only be possible when there are data have existed for some years and the population remained stable in the area. In that situation the declaration of an epidemic is very clear. However, sometimes maria outbreaks can occur in situation where the previous data are unavailable, or they are irrelevant due to significant changes happened. In this circumstance, the precise thresholds will not be possible, and an epidemic situation is more practically defined by the rapid increase in numbers, a high case-fatality rate and the fact that the existing health services are overwhelm[25]. Here ae the major types of malaria outbreak:



Source: WHO, 2002

- **True epidemics:** infrequent/cyclical outbreaks in relatively non-immune populations related to climatic anomalies. They occur mainly in arid and semi-arid zones with little or no seasonal fluctuation where infection is normally rare.
- **Strongly seasonal transmission**: variable but relatively predictable transmission influenced by normal climatic variations.
- Malaria transmission exacerbated by population movements and country political instability. The pattern can be either explosive or exaggerated seasonal variation.
- Neglect/breakdown of control: a general upward trend in endemicity and transmission in areas where malaria has re-emerged because of neglected control activities (not necessarily linked to a complex emergency).

2.2 Prediction of malaria using time series analysis

In building predictive tool for malaria surveillance in Afghanistan Anwar et al. Malar J, (2016) developed an autoregressive integrated moving average (ARIMA) models for predicting the future malaria incidence in Afghanistan. Malaria data from January 2015 to December 2015 whereby the environment and climate data have been only used to assess whether they have improved the predictive power models. Therefore, the study came up with two models, one for near-term prediction which showed that malaria incidence can be predicted using the previous cases of four previous months and 12months prior while model 2 showed that for the long-term malaria prediction; malaria incidence can be predicted using the rates 1 to 12 months prior. Eventually, the researchers concluded that the ARIMA can work as the complement to existing surveillance systems as they provide better understanding of malaria dynamics in limited resources settings with minimal data inputs but getting the malaria prediction which can be used for public health planning[26].

Afshin Ostovar et al (2016) have developed the Early warning system for predicting the malaria incidence in South-eastern Iran. By considering the contribution of meteorological factors on malaria, he used monthly data over the last 6 years from 2003 to 2009, were analysed to assess the relationship between meteorological variables including temperature, rainfall, and relative humidity with morbidity data, and malaria cases. Moreover, by using the Univariate autoregressive integrated moving average for making two models one based one for predicting weekly malaria incidence and the other for predicting monthly malaria incidence. Finally, the study came with better fit weekly model with R^2=0.863, for monthly model R^2=0.424. However, the meteorological variables were not statistically significant except the minimum and maximum temperature in monthly model[27].

Mohammed I. Musa (2015) developed ARIMA and ARIMAX models by using climate variables and the past diagnosed malaria cases from 2006 and 2011 as training set and data from 2021 as test set and created the best models fitted for predicting the malaria cases in Sudan from 2013 and 2014. The ARIMAX model used to examine the relationship between malaria cases and climate data using least Bayesian Information System (BIC) values. The result showed that ARIMA has four different models where the average for all states is (1,0,1) (0,1,1) and ARIMAX models revealed that there is a significant variation between the states in Sudan[28].

2.3 Malaria outbreak prediction using machine learning techniques

In study of Africa malaria's epidemic prediction(Muthoni Masinde, 2020); she used past malaria incidence, climate data (average annual temperature and annual rainfall) of all Africa countries exposed to malaria disease in past 18 years (2000-2017) to predict the future malaria incidence, therefore the nine machine learning algorithms (Logistic regression, Decision tree, Naïve Bayes, Deep Learning, Support vector machine, Gradient boosted tree, Random forest, Fast large margin and General linear model) have been applied to build the prediction model in MATLAB software and the data set of 272,832 rows of data for the climate data from World's knowledge portal have been used. Furthermore, the algorithms were all evaluated based the accuracy, F-measure, AUC, precision, sensitivity, recall, specificity, and classification error have been used to evaluate the ensemble of algorithms. The overall performance showed Logistic regression, Fast Large Margin, decision tree and general linear model as the four best algorithms based on that performance metrics[15].

Babagana et al (2017), developed the intelligent malaria outbreak early warning system which is a mobile application used to predict the malaria outbreak based on the climatic factors by using machine learning and the study have been conducted in Ghana (Kumasi). The contribution of ecological factors such as temperature, relative humidity, solar radiation wind speed and precipitation on malaria incidence were showcased. The study comprised of four stages; the first stage was the collection of data from repositories, the second was to identify the hidden ecological factors and to identify the causal relationship among the ecological factors, the third stage was to use machine learning algorithms, the ten machine learning algorithms have been used in building a predictive model of malaria outbreak and the performance metric of accuracy has been used to evaluate the best model; among those machine learning algorithms, support vector machine had higher accuracy than others. The last stage was to develop the mobile application which embedded the best predictor which is support vector machine; the application reads the climatic information from free weather and geographic application programming interface and predict the malaria outbreak several days in advance[1].

Thakur et al (2019) studied the prediction of malaria abundances using artificial neural network in four provinces of India; he used the big data of climate composed of temperature, relative humidity, rainfall, and vegetation index from 1995- 2014 combined with symptomatic malaria cases and the prediction model was developed using a feed-forward neural network of artificial neural network and root mean square of error was used as the evaluator of the prediction model fixed to 150%. Therefore, the results confirmed the variability of prediction between areas according to clinical variables and precipitation. The average error ranges from 18% to 117% the research ended by suggesting more exploration of data in malaria prediction to increase the accuracy in real practice[29].

Study conducted on malaria outbreak prediction using machine learning in India, Sharma et al (2015) built a malaria outbreak prediction model which played the role of early warning tool to identify the potential malaria outbreak. By using two machine learning classifiers which were Support vector machine (SVM) and artificial neural network (ANN) and the comparison of them were made to find the best model to be utilized in malaria outbreak prediction. The large data set of Maharashtra from 2011 to 2014 have been used as training data. Therefore, he used the climate variables which were temperature, rainfall and humidity with combination of clinical data which were the total positive cases of malaria, the total number of Plasmodium Falciparum(pF) and outbreak occur in binary outcome yes/no. Furthermore, he used the Receiver operating curve (ROC) and Root mean square mean square of error (RMSE) as a performance metric to the prediction model. He observed that, the model prediction using support vector machine (SVM) is more accurate than artificial neural network (ANN) hence he adopted the SVM as the prediction the model which can predict the outbreak 15 to 20 days in advance[19].

Comparative study on prediction of symptomatic and climatic based malaria parasite counts using machine learning conducted by Opeyemi (2018), the study adopted the experimental study of malaria parasite counts in Mina Metropolis, Niger state and Nigeria, meanwhile the climate data from NICOPE, Bosco, Niger state, FUTI Minna and Nigeria where the total of 1200 experimental data was collected to be used in research. However, the two classifiers Support vector machine and artificial neural network do the prediction of the model. The performance metric of accuracy, sensitivity, specificity, and false-positive rate and false the negative rate has been used to evaluate the performance of the two models, and finally, the support vector machine has been demonstrated to have high performance compared to artificial neural network[16].

3 METHODS

This part gives the guidelines on the methodology used in this study, it details on different machine algorithms used to the malaria outbreak prediction using environment data, this part also deals with the details on data source; where came from the data used in the study and how they have been collected, the machine learning techniques used to build the model. Furthermore, this part details the how the data has pre-processed to get ready for the model building which consists of data wrangling and data transformations, this part also details on different performance metrics have used to evaluate the performance of the machine learning algorithms.

3.1 Research design

This retrospective study looking into past malaria cases experience registered in HMIS and reported at RBC, with also environment data recorded on Rwanda Meteorology Agency, its objective is to give a prediction on malaria outbreak by using environment data and past malaria cases in ten selected districts by applying machine learning algorithms and all machine learning algorithms are evaluated using performance metric. The sampling methods used to select ten districts is by convenience, it is non-probabilistic methods whereby the selection criteria based on the availability of the data[16]. The study has used sampling methods based on the availability of meteorology data. Those ten meteorological stations are only ready to provide all these needed data for malaria prediction

3.2 Data sources

Two datasets were merged and used in this study, one from Rwanda Meteorological Agency which constitutes the historical environment data include weekly rainfall, average weekly minimum temperature, and average weekly maximum temperature from 2016 to 2019, climate data also includes latitude, longitude and elevation of selected meteorological stations, all these data have been collected from two meteorological stations per each Rwanda province and aggregated to ten districts including Gitega-Nyarugenge, International Kigali Airport- Kicukiro, Huye airfield, Gikongoro station-Nyamagabe, Rubengera-Karongi, Kamembe-Rusizi, Musanze airfield-Musanze, Byumba station - Gicumbi, Nyagatare station - Nyagatare and Kawangire - Kayonza. All these meteorological data which were collected daily has been computed on weekly basis and the data of station is representing the district of where it located.

Moreover, the meteorological data on these stations are linked to weekly malaria cases found in these ten districts of selection from 2016 to 2020 period. On the other side, the data of malaria cases collected from all health facilities in Rwanda from 2016 to 2020 period to represent the malaria status in Rwanda for that period. Data on malaria cases which have been collected in all health facilities and those which have collected by community. In addition, by using the data of malaria cases help to identify the occurrence of malaria outbreak; malaria outbreak defined as the sudden increase in malaria cases for the period so that by using malaria cases at district lever help to create another variable which is a dichotomous variable (two outcomes) and considered as the label in machine learning classification. Nevertheless, malaria cases were reported through the Health Management Information system (HMIS) from the malaria test done at health centres, hospitals, clinics, and test are done by health community workers (CHWs).

All data that have been used in the study were comply with the rules and regulations governed by Rwanda Biomedical Centre (RBC) and Rwanda Meteorology Agency institutions. To getting access to climate data from Meteorological Agency, I requested to present all required documents giving permission to use those data in academic research include letter to the directorate of the institution explaining the purpose of the study, the recommendation from university, personal identification card and filling out of the application form requesting data. For malaria data, the data have been collected through the health reporting system and they have accessed by request submitted at Rwanda Biomedical Centre (RBC).

Table below shows the summary statistics of features used for the study; all these features will be used to predict whether they will happen malaria outbreak.

Table 1: Summary statistics of features

Variable	Observation	Mean	Std. Dev.	Min	Max
Rainfall	2,080	19.85	24.38	0	182.9
Longitude	2,080	29.82	0.44	28.91	30.43
Latitude	2,080	-1.925	0.37	-2.4	1.28
Elevation(m)	2,080	1687.8	250.25	1377	2235
Minimum temperature(^{O}C)	2,080	15.06	1.59	9.34	26.14
Maximum temperature(^{O}C)	2,080	25.42	2.27	18.53	63.33

Source: RBC and RMA

For another side, the label (dependent variable) which is a malaria outbreak has been created based on malaria cases were found on a weekly basis, according to the Rwanda Biomedical Centre

(RBC), the malaria outbreak is calculated by computing the average of malaria cases in five years, therefore, we warned for malaria the outbreak, if malaria cases found, are greater than the average cases for other cases there is no outbreak. Therefore, the average of malaria cases was 1178.

3.3 Research framework

This study analysis part is mainly divided into three parts which are data pre-processing and cleaning, application of machine learning algorithms as well as for performance evaluation of machine learning algorithms. By using various performance metrics applied to each of the algorithms help to identify the best classifier recommended for future prediction of malaria outbreak

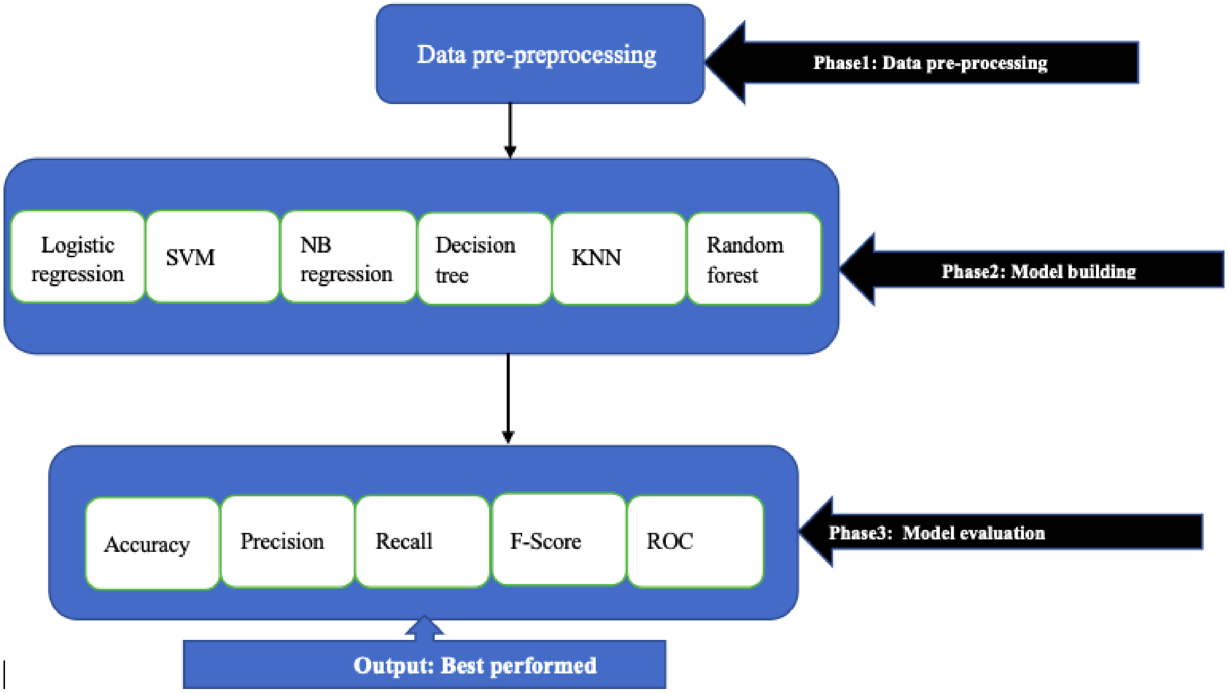


Figure 2: Research framework

3.4 Data pre-processing and data cleaning

Data pre-processing defined as the transformation or encoding of dataset in such way the machine Data pre-processing defined as the transformation or encoding of the dataset in such a way the machine learning algorithms can easily parse it, that transformation makes the features are interpreted by machine algorithms in easy way[35]. The dataset used in the machine learning model the building comprises of features (inputs) and labels (output) whereby the features are independent variables and labels are the dependent variable. For this study, the features are weekly average minimum temperature, average maximum temperature, rainfall, longitude, latitude and elevations, all features are quantitative variables and have been used to predict the dependent variable(label) which is a malaria outbreak. All features and label are combined in one dataset formatted in Ms Excel Csv and dataset must be exported in Python 3.7 programming language Jupiter notebook so that the data wrangling starts after data being imported. Data wrangling techniques have applied to the data set to handling missing values, outliers in the data set and other noise which can impact the performance of the model, the data wrangling is made in Python programming language.

3.5 Model building

After data pre-processing, the five supervised machine algorithms were applied on data set to build a predictor model, those algorithms including Support vector machine (SVM), Random Forest, Naïve Bayes (NB), Logistic regression and decision tree. However, the data set was split into two parts including the training set and test set, the training set was composed of 80% of the whole data while the test was made by remained part of 20% of the whole data set. The training set used to train the predictive model and test set to use for evaluation of a predictive model by applying the performance classifier performance metrics. The set of codes for classification were written in Python programming language (python 3.7) Jupiter 5.6 notebook and used to build the malaria prediction model.

3.5.1 Logistic regression

Logistic regression is the basis of the classification algorithm for supervised machine learning and has a similar relationship to the neural network. Logistic regression is used to assign an experiment into one of two (binary logistic regression) or several (multi-class logistic regression) types, which implies that the category of the dependent variable must be categorical.

It is one of machine learning algorithms that uses probabilistic theories, and the goal of binary logistic regression is to train an algorithm that will be able to decide on binary classification of new input observation by using sigmoid classifier[30]. The binary logistic regression is our interest in this analysis. By using the logistic regression classifier and the environment data, logistic regression classifier predicts whether malaria outbreak will occur.

Here is the expression of logistic regression mathematically.

Consider a vector X with the function [x_1 , x_2, x_3, ..., x_n] and the value of the classifier y should be 1, meaning that there will be an outbreak of malaria or 0, meaning that there will be no outbreak of malaria.

The logistic regression wants to know the probability P(y=1|x) =P, the probability that; the observation is belonging to malaria outbreak class and the probability P(y=0|x) =1-P (q), the observation belongs to no outbreak class

Odds will be the ratio of these two probabilities

Odds=$\frac{P}{1-P}$ and the Logit function will be equals to the logarithm of the Odds.

Z=Log$\left(\frac{P}{1-P}\right)$ **Logit function**

P=$\frac{1}{1+e^{-Z}}$ **Logistic function** with **Z=α+βX=β1X1+… +βdXd**

As linear regression uses Ordinary Least square to estimate the coefficients of the model, the Maximum likelihood estimation used to obtain the coefficient of the logistic regression model that relate the predictor to the target. Here is the graphical representation of Logistic regression:

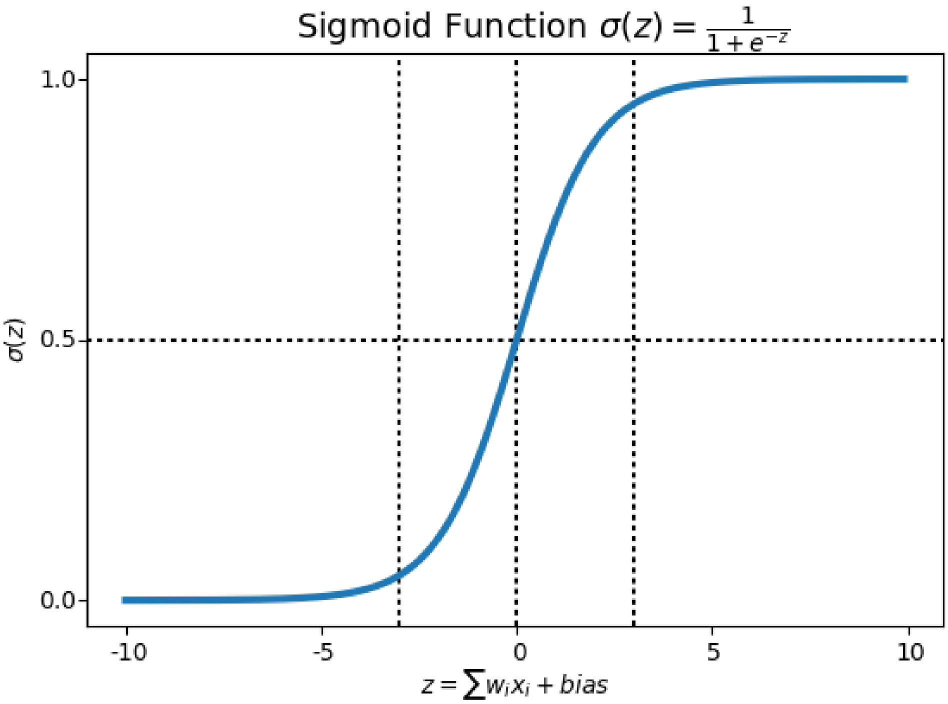


Figure 3: Logistic Regression

3.5.2 K-Nearest Neighbours

K nearest neighbour is machine learning algorithm which is only used for classification. For given training data and new data point, the K- nearest neighbour algorithm will classify new data based on the class of the training data that is close to and the closeness is determined by the distance metric such as Euclidian distance (mostly used for continuous features), Manhattan distance, absolute distance, hamming distance, etc. which applied to the feature space. The objective is to get the K-closest data point in whole training data and the classification of new data point based on majority class of K-nearest training data[31].

The training examples are needed to be in memory at run-time that why it is called sometimes Memory-Based Classification and because induction is delayed to run-time is also considered as Lazy Learning technique. It is also named Examples-Based classification or Case-Based classification because it is classification based directly on training examples[32].

For K-Nearest Neighbour, the general idea is to compare each data point in classification data set to all of data in training set and compute the distance between the data that we wish to classify and all of data of training data. Furthermore, the two stages for K-Nearest Neighbour are including the determination of the nearest neighbours and the determination of class using these neighbours.

In case of this study, the K-nearest Neighbour play the role of binary classifier, it classifies if there is an outbreak of malaria or not based on new environment data by calculating the Euclidian distance between the training data and new input data and the result get classified accordingly. Here is the mathematical expression of K-nearest model.

Suppose we have training dataset D made up of $(x_i)_i$ Є [1, |D|] training samples. There is also feature F in which any numerical value is normalized to the range of [0, 1]. Each training example is labelled with class label y_j Є Y. Therefore, the objective is to classify an unknown example q, for each xi Є D; the distance between xi and q will be calculated as follow:

$$d(q,x_i) = \sum_{f \epsilon F} w_f \, \delta(q_{f,} \, x_{i\,f})$$

The three possible values should be taken by distance metric

$$(q_{f,}x_{i\,f}) = \begin{cases} 0 \ = f \ is \ discrete \ \ and \ q_f = x_{i\,f} \\ 1 = f \ is \ discrete \ and \ q_f \neq x_{i\,f} \\ \ |q_f \ - x_{i\,f} = \ f \ is \ continuous \end{cases}$$

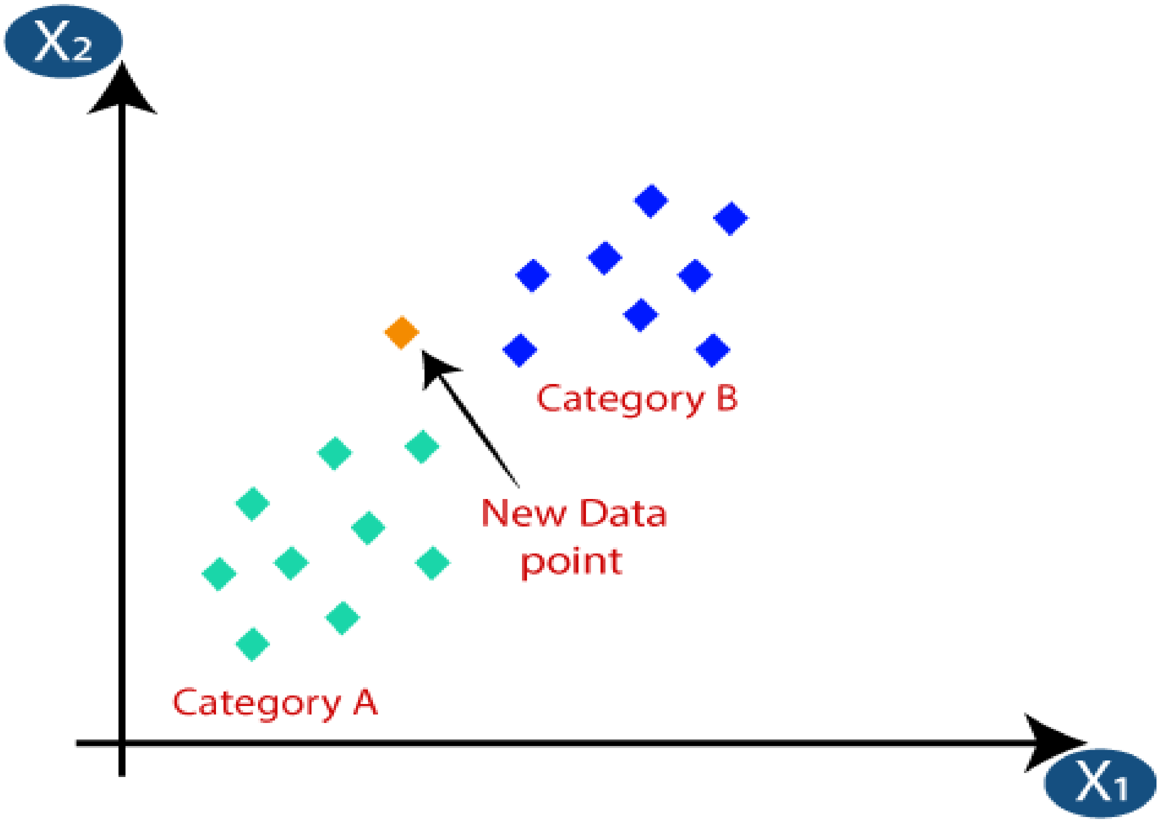


Figure 4: K-Nearest Neighbour

3.5.3 Support Vector Machine (SVM)

Support vector machine (SVM) is supervised machine learning technique associated with the algorithm that analyses data to know the pattern. SVM is used for classification and regression analysis. For a given training data, which must be categorized on of the two categories, SVM training algorithm builds a model that assigns new examples into one of the two categories, making it a non-probabilistic binary linear classifier. Given some training data D, a set of n points of the form

D= {(X_i Y_i) | X_i ∈ Rp, Y_i € {-1,1}} i =1 to n

Where the Y_i is either 1 or −1, indicating the class to which the point X_i belongs. Each X_i is a P-dimensional real vector. We want to find the maximum-margin hyperplane that divides the points having y_i= 1 from those having y_i=-1. Any hyperplane can be written as the set of points x satisfying, w. x – b=0, where '.' denotes the dot product and w the (not necessarily normalized) normal vector to the hyperplane. The aim of Support Vector machine is to place the hyperplane in such way to be as far as possible to closest points of both classes. The parameter b/||w|| determines the offset of the hyperplane from the origin along the normal vector w. If the training data are linearly separable, we can select two hyperplanes in a way that they separate the data and there are no points between them, and then try to maximize their distance. The region bounded by them is called "the margin". These hyperplanes can be described by the equations w.x – b=1, and w. x – b=-1[16].

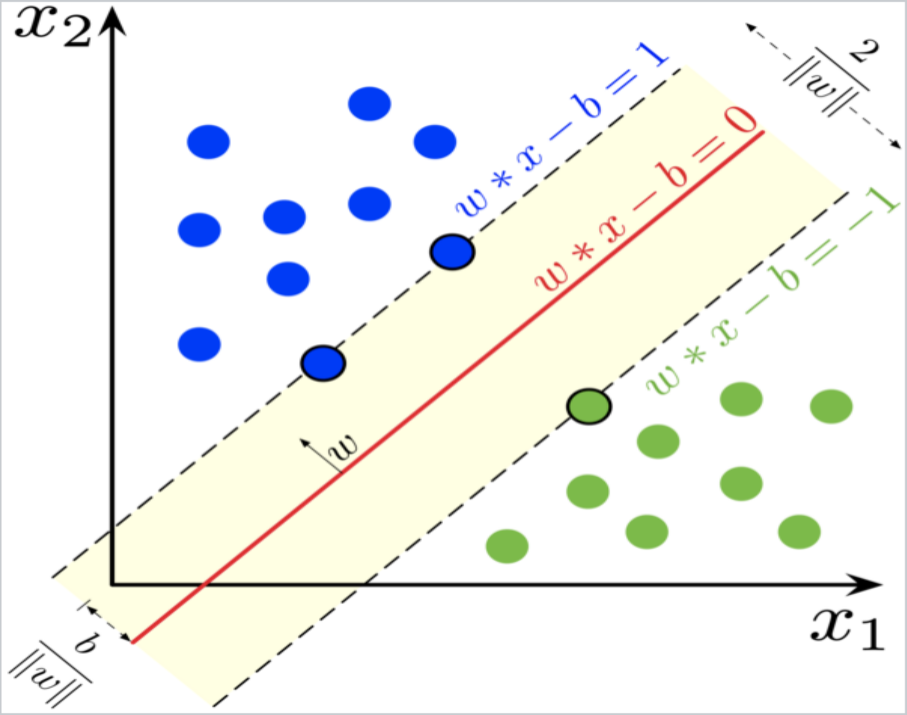


Figure 5: Support Vector Machine

3.5.4 Random Forest

Random forest is supervised machine learning used for regression where the response is continuous variable and classification whereby the response is categorical variable, mainly random forest is used for classification problem and it is made by decision trees and the more decision trees, the more it becomes more robust. On building every decision tree, it uses bagging and feature randomness by creating the uncorrelated forest of tree. Random forest creates the decision on sample data and makes prediction on each of the sample data and finally the best prediction made by voting. Random forest considered as ensemble methods, and it gives the best prediction compared to a single decision tree because it reduces the overfitting by making average of the results[33].

Let X be a random vector in such that X= $(x_1, x_2,\dots, x_p)^T$ which represents the inputs or predictor and random variable Y and it is response and we assume the joint distribution $P_{XY}(X,Y)$. As other classifications, the objective is to find the prediction function f(x) to predict Y. Hence that prediction function is given by loss function L (Y, f(x)) and it has the purpose to minimize the expected loss value, E_{XY} (L (Y, f(x)).

L (Y, f(x)) is the measure of how close f(x) to Y, for classification

$$L\ (Y, f(x)) = I\ (Y \neq f(x)) \\ = \begin{cases} 1\ if\ Y = f(x) \\ 1\ Otherwise \end{cases}$$

Let the set of possible values of Y represented by ¥ and by minimizing E_{XY} (L (Y, f(x))) for zero to one loss gives **f(x) =** $\boldsymbol{argmax_{\text{Y€¥}} = P(Y = y|X = x)}$ which is Bayes Rules

For the prediction f in terms of collection of so called “base learners” $h_1(x), \dots, h_j$ (x) and the bases are combined to give “the ensemble predictor” f(x) which the most frequently predicted class[30]. Here is the graphical representation of Random Forest

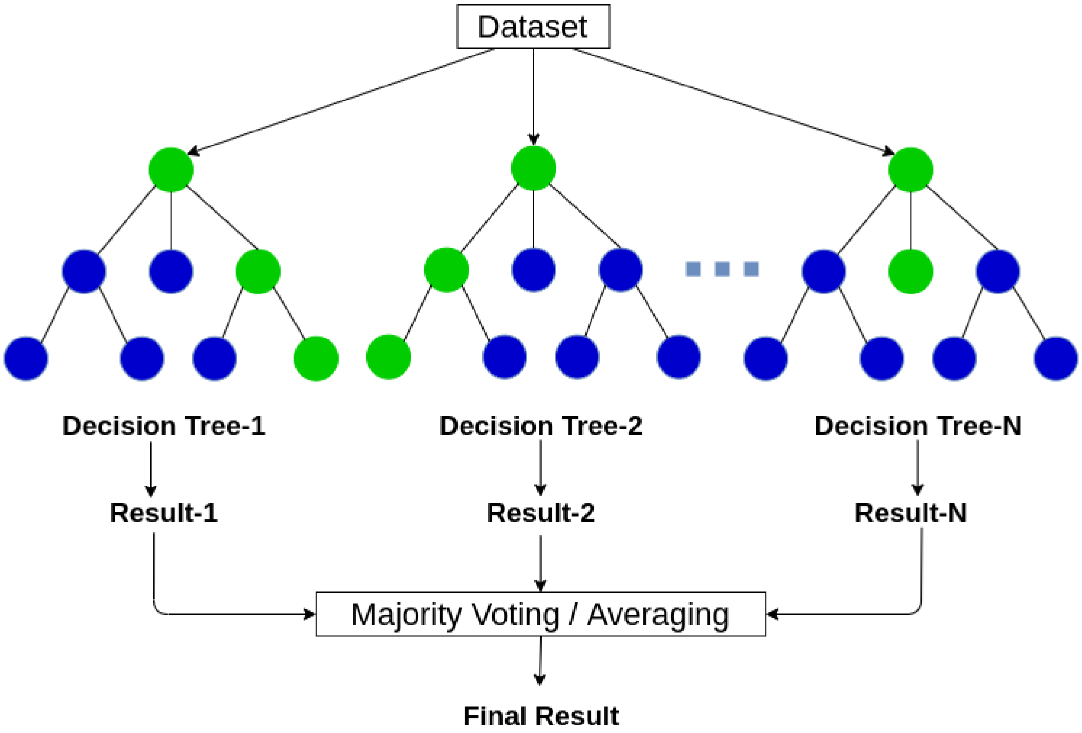


Figure 6: Random Forest

3.5.5 Naïve Bayes

The naive Bayes classifier is supervised machine learning algorithm uses the Bayes' Theorem which assumes that; each feature only depends on the class. This means that each feature has only the class as a parent. Naïve Bayes is attractive as it has an explicit and sound theoretical basis which guarantees optimal induction given a set of explicit assumptions. There is one disadvantage on which the independency assumptions of features with respect to the class are violated in some real-world problems. However Naïve Bayes is remarkably in violation of assumptions, it is fast and easy to implement with the simple and effective structure. It is also useful for high dimensional data as the probability of each feature is estimated independently. The practical use of machine learning, Naïve Bayes classifiers has different qualities combined into one Naïve classifier include: an intuitive approach, its ability to work with small data, low computation cost for training and prediction and finally it gives the solid results in variety of setting. Let *C* denote the class of an observation X. To predict the class of the observation X by using the Bayes rule is possible by using posterior probability

$$P(C\backslash X) = \frac{P(C)P(X\backslash C)}{P(X)}$$

In the Naïve Bayes classifier, using the assumption that features X_1, X_2, . . ., X_n are conditionally independent of each other given the class; we get

$$P(C|\mathbf{X}) = \frac{P(C)\prod_{i=1}^{n} P(X_i|C)|}{P(\mathbf{X})}.$$

Where:

- P(C\X) is the posterior probability of the class
- P(C) is the prior probability of the class
- P(X\C) is likelihood which is the probability of predictor
- P(X) is the prior probability of the predictor or marginal probability

In Naïve Bayes, we compare the posterior's observation for each possible class and the because the marginal probability remains the same for all comparison, the comparison goes to numerators of posterior for each class. Therefore, the class has greatest posterior numerator will be the predicted class[34’35].

3.5.6 Decision tree

A decision tree is machine learning used in classification and regression expressed as a recursive partition of the in-stance space. Decision tree composes with nodes. It has three kinds of nodes; the node which has no incoming edge called roof, the other nodes that have one incoming edge and other nodes with outcoming edge called internal or test node while the others without outcomings are leaves or terminal/decision nodes. In a decision tree, each internal node splits the instance space into two or more sub-spaces according to a certain discrete function of the input attributes values.

In decision tree, each decision rule appears to the decision. With decision rules it will create other branches which lead to new nodes. In fact, the decision tree is more popular because of its easy interpretability. For classification use of decision tree, each leaf is assigned to one class representing the most appropriate target value. However, each leaf may have the probability vector indicating the probability of certain value that the target attribute has, and the new observations are classified by navigating from the root of decision tree to the leaf according to the outcome of the tests along the path[18].

Decision tree learning is based on finding the decision rule that produce the greatest decrease in impurity at the node and the decision tree classifier uses by default Gini impurity which its formula written as:

$$G(t) = 1 - \sum_{i=1}^{c} p_i^2$$

Where G(t) is Gini impurity at node t and p_i is the proportion of observation of class c at note t. The process of finding decision rule continue that create splits to increase impurity is repeated recursively until all nodes are pure or cut-off reached[35].

3.6 Evaluation of model

Model evaluation in machine learning is an integral part of building model because it helps to know the best model fit the data that will be used for future prediction. The evaluation of model performance is not done on the training for avoiding the problem of overfitting, but model evaluation uses test set[36]. There are so many performance metrics used for the evaluation of machine learning models depending on the type of machine learning algorithms (classification or regression) used. Moreover, it is better to use more evaluation metrics for a single model because one model may perform well by using one evaluation metrics and perform poorly by using other evaluation metrics so by using different evaluation metrics will help to conclude if the model performs correctly and optimally[37].

For this study evaluation of the performance of machine algorithms classifiers will be made to measure the performance of the machine learning algorithm, to apply this, the test set will be used, each machine learning algorithm is trained using the training set and the evaluation of the machine learning algorithms are measured on the test set. The evaluation metrics that have been used in the study, they include Accuracy, Recall, Precision, F-score, and ROC all those evaluation metrics have been used to measure the performance of six classifiers.

3.6.1 Confusion matrix

A confusion matrix is used to evaluate the quality of the output of a classifier. The diagonal elements are the number of points where the predicted label is equal to the true label. The off-diagonal on the other hand, are the labels that the classifier mislabelled. It is better when the diagonal values of the confusion matrix are higher since it indicates that many predictions are correct[38].

Table 2: Confusion matrix

		Predicted class	
		Class 1= Yes	Class 2= No
Actual class	Class 1=Yes	True Positive	False Negative
	Class 2 = No	False Positive	True Negative

True Positives tells us the number of cases that the classifier correctly predicted that the person is positive. In this case Total number cases predicted as outbreak whereby they happened outbreak

True Negative tells us the number of cases that the classifier correctly predicted that the person is negative, for this study it indicates the total number of cases classified as no outbreak whereby no outbreak happened.

False Positives are also called Type 1 error. It tells us the number of cases that the classifier incorrectly predicted, it tells us the number cases predicted as outbreak whereby no outbreak happened.

From the confusion matrix, we can obtain evaluation metrics that will enable us to compare different machine learning models according to their performance. These metrics include:

- Accuracy
- Precision
- Recall
- F-score

a. Accuracy is known as the correctness of the model which is the sum of true predicted malaria outbreak over the total number of total malaria outbreak cases

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

b. Precision is measure of accuracy of the correct malaria outbreak cases over the total malaria outbreak including correct classified and wrong classified.

$$\text{Precision} = \frac{TP}{TP+FP}$$

c. Recall is also called true positive rate is the measure of correct malaria outbreak cases over all cases should have been classified as malaria outbreak cases.

$$\text{Recall} = \frac{TP}{TP+FN}$$

d. F- score is the combination of Recall and precision of the model, it is considered as the harmonic mean of recall and precision of the model.

$$\text{F-score} = 2.\frac{\text{Recall . Precision}}{\text{Recall+Precision}}$$

ε. **Receiver Operating Characteristics (ROC) curve and Area under Curve (AUC)**

ROC is defined as the plot of the test of sensitivity or true positive which is plotted on Y-axis versus the 1-specifity or the false positive on X-axis. It has been efficient to evaluate the performance or the quality of diagnostic test and mostly it is used in radiology test. ROC has a good performance through the decrease of standard error and as the number of test sample and Area Under Curve (AUC) increases as well as increase sensitivity when the analysis of variance test is performed[39]. The area under the curve (AUC) is defined as the area under the ROC curve and it measures how good prediction it is. It is also the gauge for the quality of separation[7].

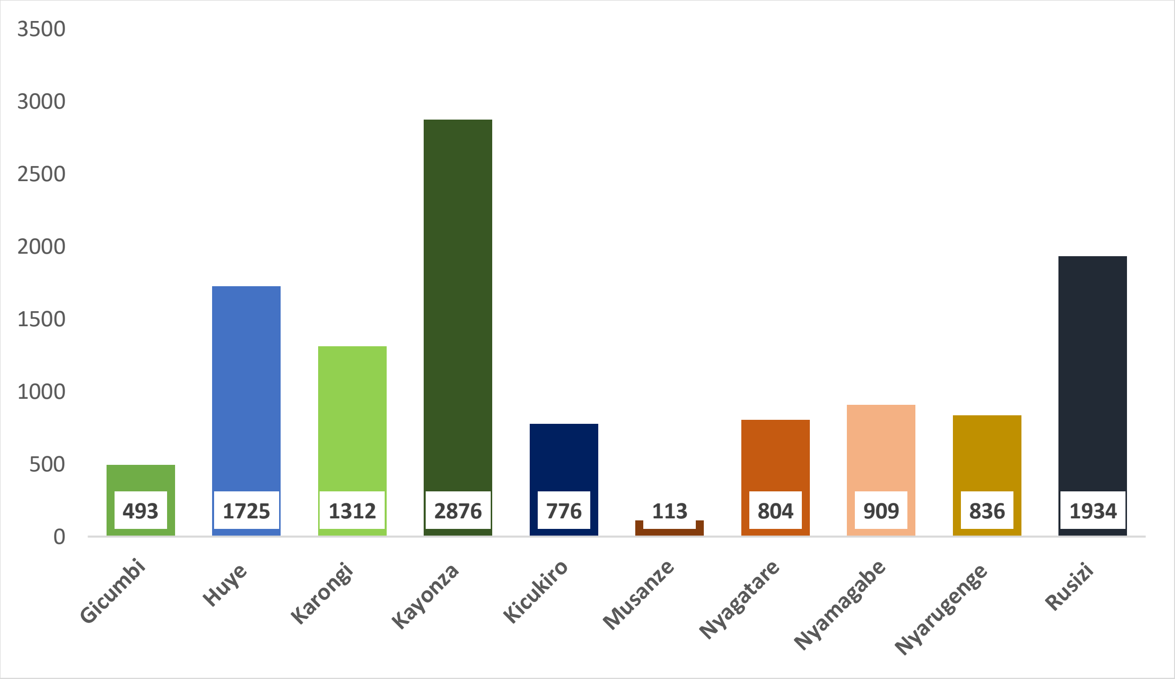
4 RESULTS AND DISCUSSION

This chapter gives the details about the process from data extraction in Python to the discussion of the findings, it starts with, the exploration of features by using descriptive statistics and looks patterns among features by showing correlation among features as well as looking into distribution of the features. It also covers the application of machine learning algorithms to the prediction of malaria outbreak, presents, and interprets the results on the evaluation metric used for evaluation of classifiers. Furthermore, it ends with the discussion of each classifier according to its the performance then the best model is recommended for malaria outbreak prediction.

4.1 Results

4.1.1 Description of features

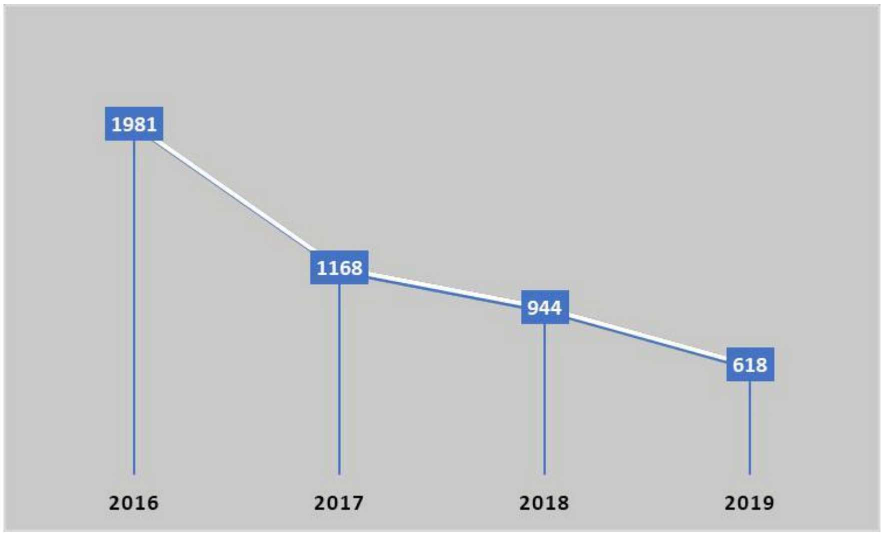
This part shows the characteristics of features used for prediction of malaria outbreak, it describes each feature by using descriptive statistics and presented either by graph or table which makes easy to interpret each feature in quantitative way



Source: RBC data

Figure 7: Distribution of weekly average malaria cases by district (weekly)

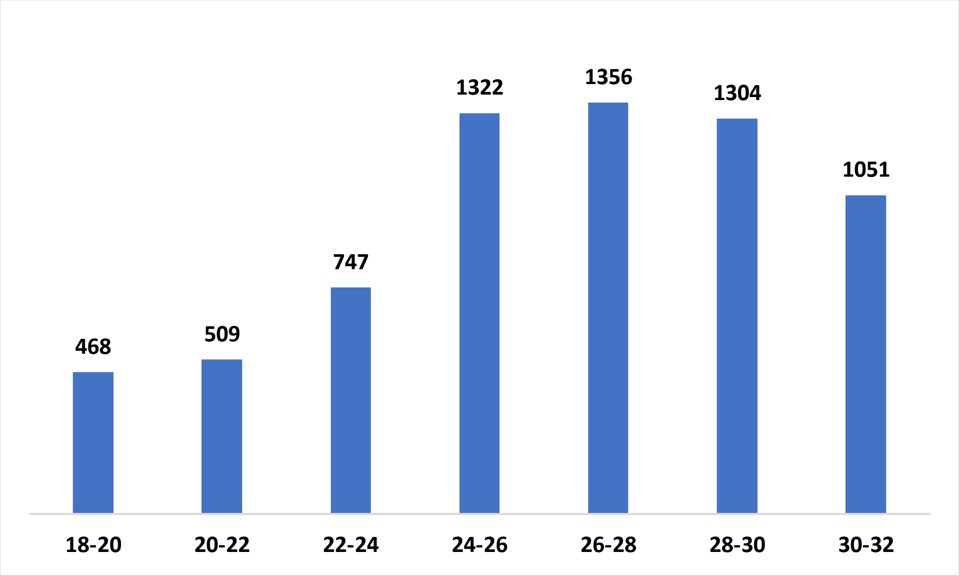
The figure 7 shows the occurrence of malaria cases in ten sampled districts, among all the districts, Kayonza presents more weekly average malaria cases compared to other districts with 2876 cases per week, followed by Rusizi, Huye, and Karongi with respectively 1934, 1725 and 1312 malaria cases, the districts with lower malaria cases are Musanze and Gicumbi which presents 113 and 493 registered cases respectively.



Source: RBC data

Figure 8: Trend on malaria cases since 2016 to 2019

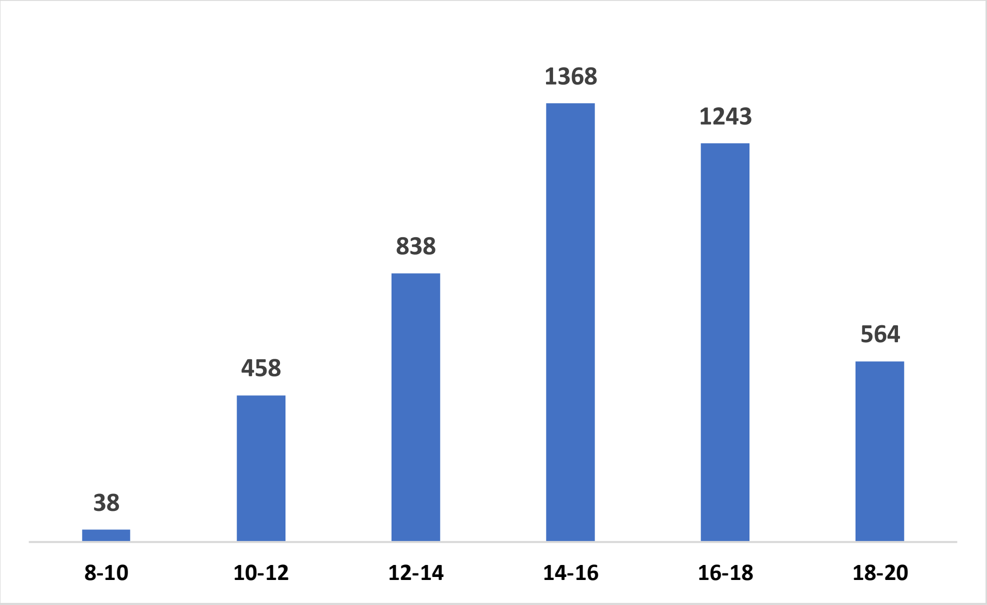
Due to the different interventions made to reduce the malaria prevalence in Rwanda has significantly reduced as figure 8 shows. Since 2016 in ten districts the weekly average cases of malaria were 1981 cases, and they have declined in such way in 2017 became 1168 and in 2019 they became 618 average malaria cases per week.



Source: RBC and Rwanda Meteorological Agency

Figure 9: Presentation of malaria cases on maximum temperature

The figure 9 shows the distribution of weekly average malaria cases with maximum temperature, it gives the rough picture of relationship of maximum temperature to malaria. According to the figure, most of the cases of malaria have been registered at the average of maximum temperature ranged from 24 to 30^{0}C. Between 26-28^{0}C registered the most malaria cases (1356 cases) and 18-20^{0}C is associated with least malaria cases which were 468 weekly average cases.

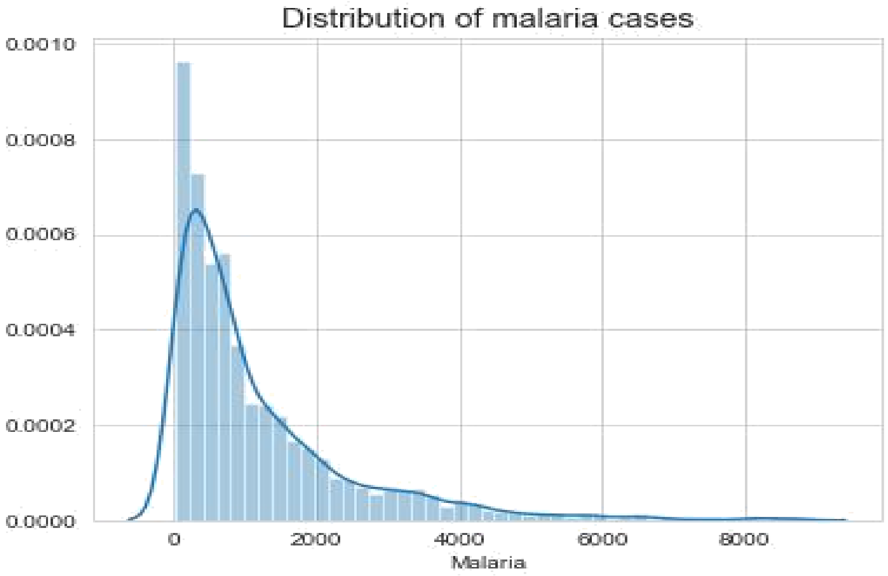


Source: RBC and Rwanda Meteorological Agency

Figure 10: Presentation of malaria cases with minimum temperature

The figure 10 shows the weekly average malaria cases with the daily minimum temperature, it gives the picture of relationship between the minimum temperature and malaria cases. According to the figure, it is more likely to get more malaria cases if the minimum temperature ranged between 14 to 16^{0}C they have registered the average of 1368 weekly malaria cases. Between 8 to 10^{0}C they have registered the lower malaria cases where 38 average malaria cases were recorded.

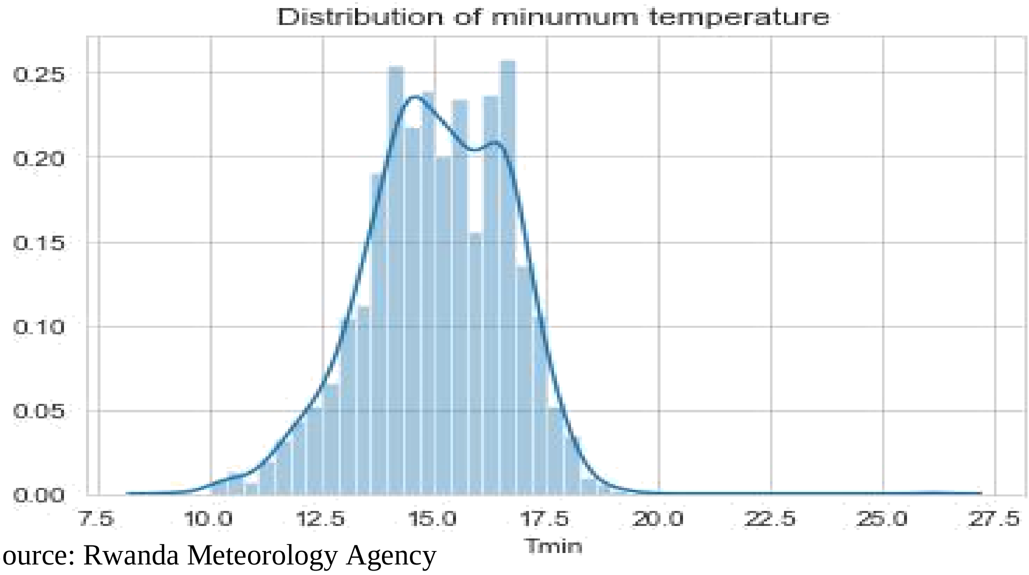
4.1.2 Distribution of the features



Source: RBC

Figure 11: Distribution of malaria cases

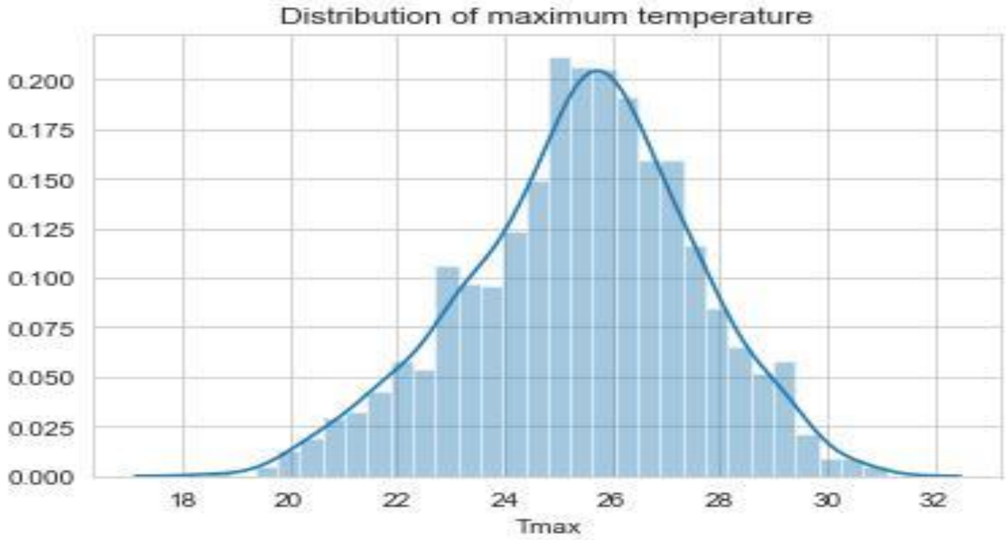
According to the graph above the malaria cases data are not normally distributed it is right or positive skewed and its mean is likely to be greater than median and mode, weekly average malaria cases are more concentrated between 0 to 2000 and it has some outliers which can heavily affect the mean. This is explaining the variability of malaria cases over the period of the year as the inequal distribution of malaria in sampled districts.



Source: Rwanda Meteorology Agency

Figure 12: Distribution of minimum temperature

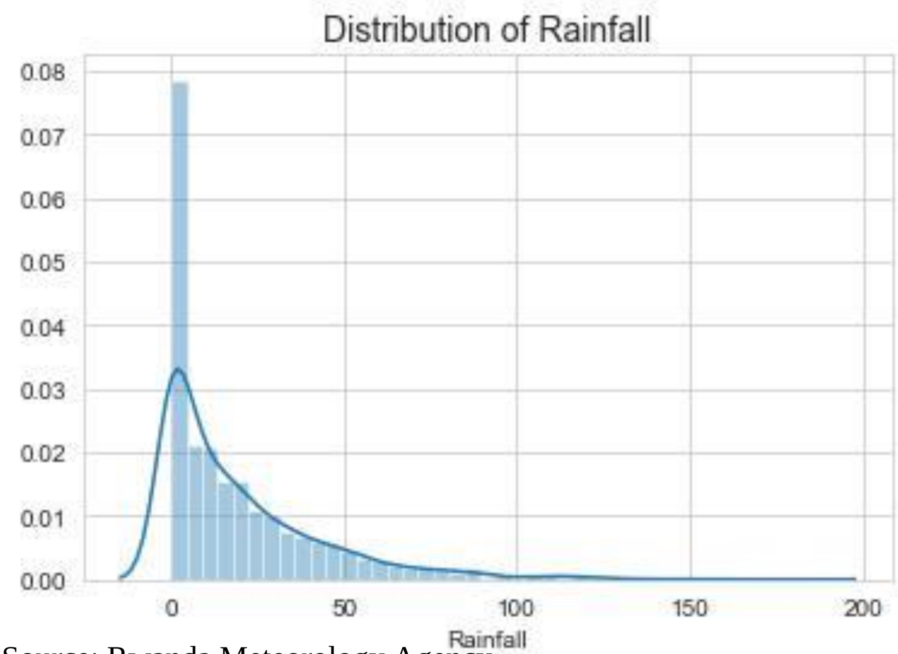
Figure displays the distribution of minimum temperature registered in the ten sampled districts, as the graph shown, minimum temperature is approximately normally distributed where mode, mean, and median converge at around 15^{0}C. In ten districts, the minimum temperature is from the 9.3^{0}C to 19.4^{0}C.



Source: Rwanda Meteorology Agency

Figure 13: Distribution of maximum temperature

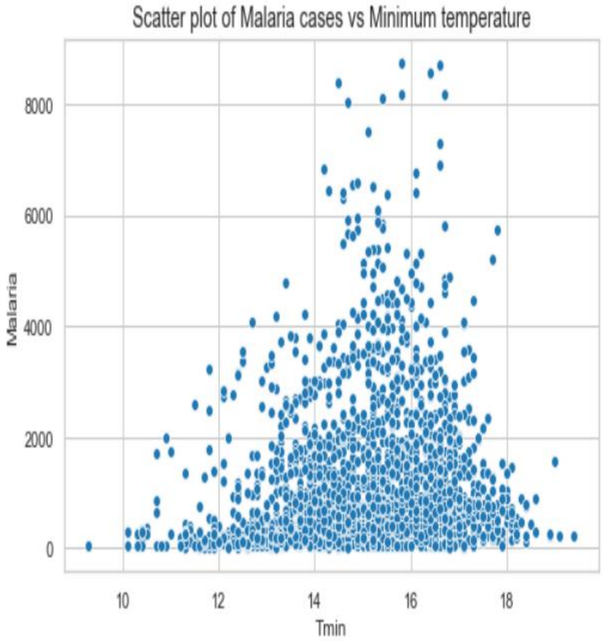
Average weekly maximum temperature data are normally distributed according to the graph above, the median and mode are almost symmetric to the mean; there is no outlier among the maximum temperature. The mean, mode and median are likely to turn around 25^0C. The daily maximum temperature is ranged from 18.1 to 31.1^0C.

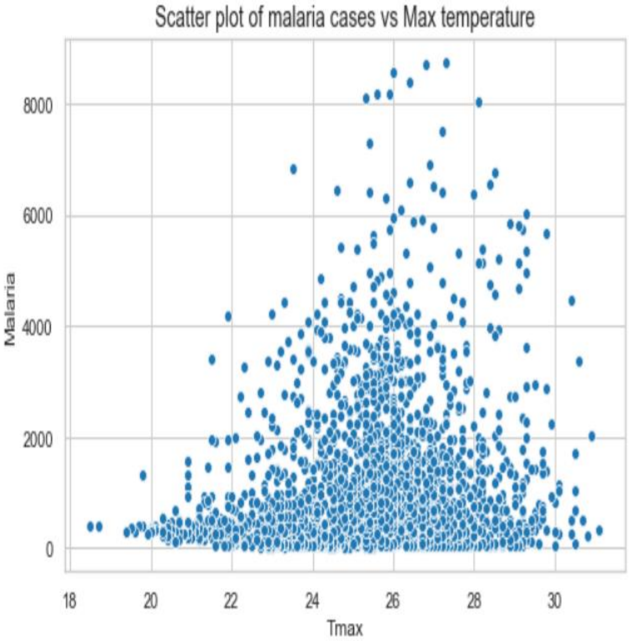


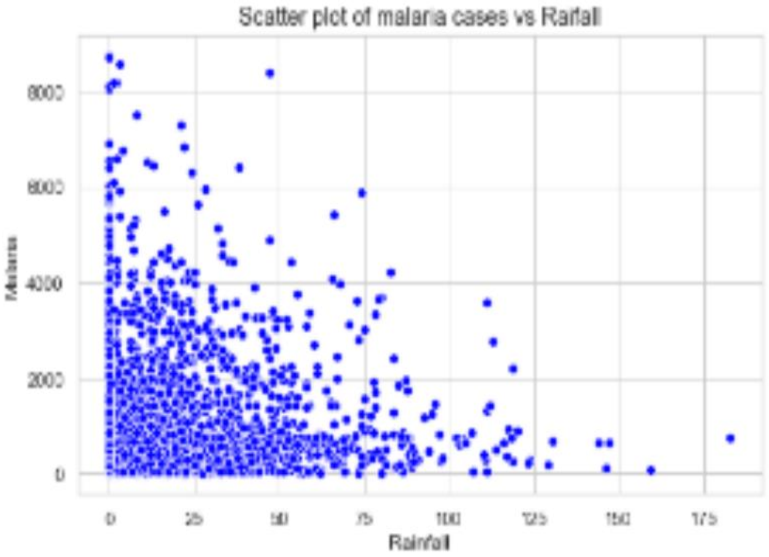
Source: Rwanda Meteorology Agency

Figure 14: Distribution of weekly rainfall

Weekly rainfall data are not normally distributed, it is right skewed, whereby mean, mode and median are not equal, most of the weeks have registered 0 mm of rain, comes as an outlier to this data set of rainfall and it explained by the seasonality of rain in Rwanda.


Scatter plot of Malaria cases vs Minimum temperature
Malaria
8000
6000
4000
2000
0
10
12
14
16
18
Tmin

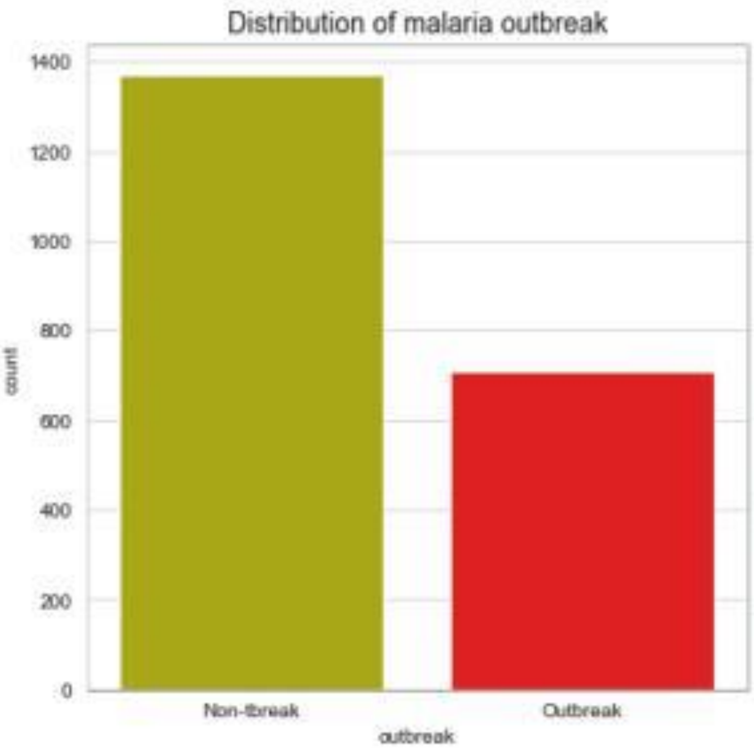

Scatter plot of malaria cases vs Max temperature
Malaria
8000
6000
4000
2000
0
18
20
22
24
26
28
30
Tmax


Scatter plot of malaria cases vs Rarfall
Malaria
8000
6000
4000
2000
0
0
25
50
75
100
125
150
175
Rainfall

Source: RBC and Rwanda Meteorology Agency

Figure 15: Scatter plot of malaria cases with temperatures and Rainfall

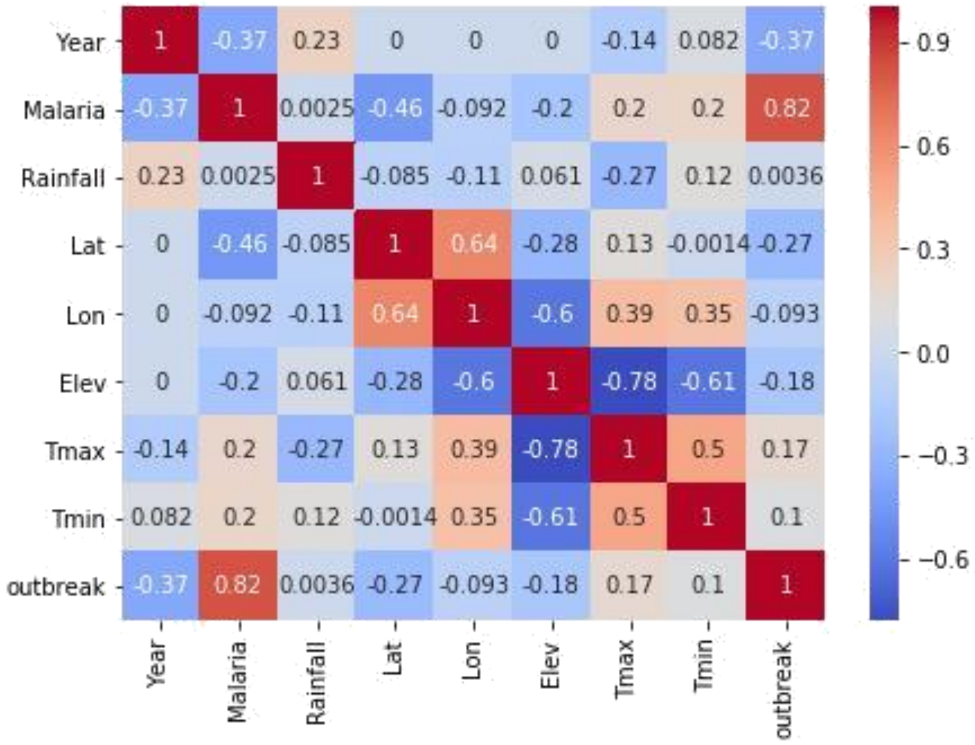
The figure 10 shows the relationship between the occurrence of malaria cases with climate data, that relationship is presented in scatter plots, as shown in figure 10 there is a weak positive relationship between malaria cases and average maximum, the same as average weekly minimum temperature, it has weak positive correlation between two which implies the increase of temperature may associate to slight increase in malaria cases. According to the scatter plot of malaria cases and rainfall, in the data we have the relationship between malaria cases and rainfall in ten districts is insignificant, the increase or decrease of rain do not contribute a lot in variability of weekly malaria cases.



Source: RBC

Figure 16: Presentation of the predictor

After applying the used formula of calculation of malaria outbreak: the malaria outbreak happened whereas the number of malaria cases occurred are greater than the average of weekly malaria cases calculated over the past five years. According to the malaria data from ten districts of 2080 observations, there are 710 cases of malaria outbreak and other 1370 cases declared no malaria outbreak. However, this predictor will serve as machine learning predictor or label (Y) whereby all features (X) have been used to predict this predictor.



Source: RBC and Rwanda Meteorology Agency

Figure 17: Correlation matrix among all variable

The graph of the correlation matrix intends to show how the variables are correlated themselves, this correlation has been calculated by using the Spearman coefficient, it lets us know the variables which are highly correlated and show to what degree of multicollinearity among the features. According to the correlation matrix, most of the features are not strong collected except the minimum and maximum temperature as well as temperature and elevation whereas they are highly negatively correlated with more than -0.6, the longitude and latitude is also highly positively correlated with more than 0.6. Briefly, this the correlation matrix shows that environmental data are correlated among themselves they have a degree of multicollinearity. However, most of the classifiers are not highly affected by multicollinearity so that we get assurance our features should be used for prediction of malaria outbreak.

4.1.3 Evaluation of the model

After training machine learning algorithms, the models are evaluated to measure their performance in prediction role, the models are evaluated by using the test set, looking into precision, recall, accuracy and F1 score, confusion matrix and ROC, all score between 0 to 1 whereby 0 is the worst model while 1 represent the perfect model, those metrics show about effectiveness and the efficiency of the machine learning algorithm is to the prediction for new inputs. The following table shows the evaluation for each of the five classifiers.

Table 3: Evaluation metrics with machine learning algorithms

Classifier	Precision	Recall	F score	Accuracy	Confusion matrix
Logistic Regression	0.838286	0.837935	0.83805	0.8382	[[332, 70], [63, 357]]
Random Forest	0.908805	0.906965	0.907345	0.907543	[[354, 48], [28, 392]]
Gaussian NB	0.733348	0.686158	0.672771	0.690998	[[187, 215], [39, 381]]
Decision Tree	0.886594	0.88511	0.885425	0.885645	[[346, 56], [38, 382]]
Support Vector Machine	0.600169	0.561336	0.518035	0.568127	[[101, 301], [54, 366]]
K Neighbours	0.821104	0.816311	0.816541	0.817518	[[306, 96], [54, 366]]

According to the table, among these all classifiers, Random Forest presents good performance metrics with more 90% of accuracy, precision, recall and F the score which means it can make 90% correct predictions, it is followed by a decision tree with more 88% of accuracy and F score, recall and precision. The logistic regression also presents a good performance metric where all their performance metrics are around 83% that make it among the good binary classifiers. The least performer model is Support Vector Machine algorithms with 56.8% of accuracy, 51.8% of F score, 56.1% of recall and 60% on precision. In general, these machine learning algorithms perform well because most of them can make more 70% good classification prediction.

4.1.3.1 Receiver operating characteristic curve (ROC) and Area Under Curve (AUC)

Another performance metric used for classification is Receiver Operating Characteristics (ROC) and Area Under curve, it helps to know how the model can separate classes of the predictor, it is plotted the true positive rate versus the false-negative rate. For this predictor, it helps to know to which degree our model can distinguish the outbreak and non-outbreak of malaria. If the AUC is closer to 1 the better those models can distinguish the malaria outbreak and non-outbreak. The following graphs are ROC and AUC obtained during the evaluation of those models by ROC-AUC

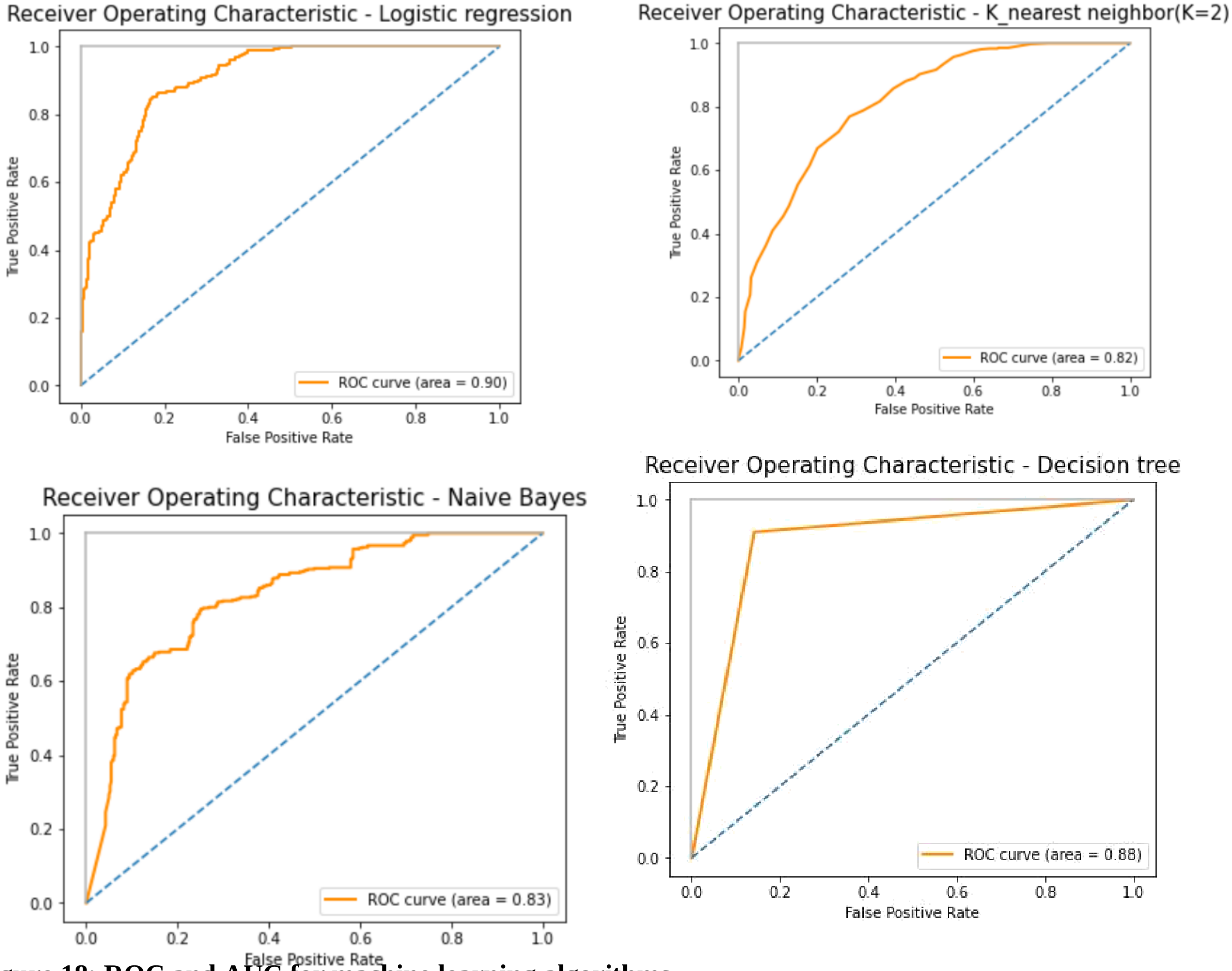


Figure 18: ROC and AUC for machine learning algorithms

The ROC and AUC above they all show the degree of discrimination of non-outbreak and the outbreak of malaria, by plotting the true positive rate versus false positive rate; the true positive rate is the true positive over the summation of true positive and false negative. Meanwhile, the false-positive rate calculated by false-positive over the summation of a true negative and false positive. According to the graphs above, those all classifiers are represented by good AUC score which is greater than 80%, this simply means that; those entire models are capable to distinguish whether they have predicted malaria outbreak or non-malaria outbreak at more than 80%. The model that has the best AUC score is Random Forest with 97% which means its degree of separation is at 97%, logistic regression model also presents a very good score of 90%, Decision tree gets 88% followed by Naïve Bayes with 83%, the least performer among another model is K nearest neighbour with 82%. Furthermore, this additional proof to the good performance of our models and the capability to the classification of malaria outbreak using climate data.

4.2 Discussions

The malaria transmission is strongly depended to many factors including the vector abundances of the anopheles mosquitos' category: the transmitter of malaria disease, it also depends on mosquitoes' bite frequency to human beings and its susceptibility to the parasite, the population immunity, the longevity of mosquito, the rate of development of mosquitoes and human behaviors. The social factors also affect a lot on the transmission of malaria including housing conditions and the control measures for mosquito. Moreover, the climatic condition and its variability have a significant impact on the incubation of plasmodium and the breeding of Anopheles which is the key environmental contributor of malaria transmissions[12]. The limitation

of this study is that only focused on the contribution of climate and environment vectors, it does not ignore the contribution of other factors to the variability of malaria as highlighted above but the purpose of this study is to show the contribution climate factors to the variability of malaria.

The study has found the positive correlation between malaria prevalence weekly average minimum temperature, average maximum temperature, and the rainfall, compared to other, maximum temperature is very correlated to malaria outbreak with Spearman correlation of 0.17 which means that that; with an increase of increase in maximum temperature malaria tends to increase as well. Furthermore, the elevation has demonstrated the negative correlation, the more elevation increases malaria prevalence decrease in Rwanda that justifies the lower prevalence of malaria in Gicumbi and Musanze as Figure 5 indicates; it means that; as long as the elevation increases the environment becomes unfavorable for the breeding of mosquitoes, also longitude and latitude showed negative correlation with malaria outbreak.

According to the performance metric used to measure the classifiers, almost of all classifiers are scored above 70% on the metric of performance. All classifiers have been trained and tested by climate data including average minimum temperature, average maximum temperature and rainfall. The prediction model of malaria outbreak also has involved the year of data collection, longitude, latitude, and elevation in selected ten districts of Rwanda (Nyarugenge, Kicukiro, Huye, Nyamagabe, Gicumbi, Musanze, Karongi, Rusizi and Kayonza and Nyagatare). All these features have used to predict whether they occur malaria outbreak or not. However, the malaria outbreak as the sudden increase in malaria cases, Rwanda Biomedical has computed the malaria outbreak whereby the available malaria cases in specific period surpass the average computed malaria cases occurred in the past five years, after computation we found that malaria outbreak happened whether the weekly malaria cases are greater than or equal 1178 malaria cases over the week, in another case there is no outbreak of malaria.

After fitting the model on the train set and evaluate them using the test set; Random Forest has demonstrated better performance according to the performance metric whereby it has shown the accuracy of 90.75%, F score of 90.73%, Recall of 90.60%, Precision of 90.88% and AUC score of 97%, there is more than 90% assurance that our model makes correct classifications, it presents in a good performance because it is not highly affected by multicollinearity compared to other classifiers. Decision Tree classifier also comes with high performance, according to the

table of performance metric Decision Tree has the accuracy of 88.5%, the F score of 88.5%, the Recall of 88.5%, the precision of 88.6% and the AUC score of 88%, by the Decision tree used for classification, it mostly gives a good prediction. For our model, the climate data displays good performance to predicting malaria outbreak. Furthermore, the logistic regression has shown good performance for the prediction of malaria outbreak, by using Sigmoid function it makes 83.82% of accuracy and 83.79% of precision as well as 90% of AUC score. The logistic has been a good binary classifier; it has set the threshold of 0.5 whereby the probability below then 0.5 classified as 0 represents non-outbreak and the probability greater than 0.5 classified as 1 represents the malaria outbreak. For the K-Nearest neighbour the performance is pretty good, it has an accuracy of 81.75% and AUC score of 82% which guaranty to make the more than 80% correct classification, it makes a prediction by calculating the Euclidian distance of the closest points and the closest point is the class of the label. The least performer classifier is Support vector machine compared to other classifiers, it has the accuracy of 56.81%, F score of 51.80, Recall of 56.33% and Precision of 60.01%, this the classifier was highly affected by the multicollinearity among the climate data therefore, this model is not recommended for using SVM.

According to the comparison made for all classifiers, Random Forest has proved to be the best in all performance, it displays the good accuracy, precision, F score recall and AUC whereby more 90% of the prediction made by using Random Forest to predict malaria outbreak by using all those features are more likely to be correct, however, the other classifiers showed the good performance but Random Forest was the best compared to other therefore the recommended model that should be used to predict the malaria outbreak is Random Forest.

5 CONCLUSION AND RECOMMENDATION

5.1 Conclusion

For this study, we have built the prediction of malaria outbreak by using climate data and environment data, the total of 2080 observations from environmental factors and malaria cases from ten selected districts have been used. The study has also used machine learning algorithms especially classifiers, we have assessed six classifiers including Random Forest, Decision tree, Logistic regression, K-Nearest neighbour, Naïve Bayes, and Support vector machine. The performance of these classifiers is evaluated using accuracy, Recall, Precision, F-score, and AUC score. Among the classifiers, Random Forest comes with high performance compared to other classifiers with more than 90% in all evaluation metrics; it has shown the accuracy of 90.75%, F-score of 90.73%, The precision of 90.69% and Recall of 90.88%. However, the other classifiers also have shown the high performance except for Support vector machine which shown only around 60% in all evaluation metrics, but other classifiers scored above 70% on the evaluation metrics. This high performance explained the linkage between malaria and the climate factors but also it has shown the efficiency of machine learning in prediction especially in classification.

The study provides the early warning system of malaria outbreaks based on environment factors using machine learning algorithms, this system should be used by hospitals, health care providers, health involved organizations, to be aware ahead of time whether there might happen the malaria outbreak so that they can take precautions and make available the resources ahead of time so that the human lives be saved. In addition, the system can be more useful to identify the place might have a heavy malaria burden according to its climate condition so that more interventions should be allocated in that place. It is a contribution to the public health, and it might be used as one of malaria control system so that burden caused by malaria should decrease as we use this model in the correct way. This study was only limited on building model but there is a still a room for building mobile application whereby it may become more helpful to the wide community.

This study has only used the environment factors including temperature, rainfall, altitude, latitude and longitude to predict whether there occurs the malaria outbreak or not, it does ignore the contribution of other factors contributing to the variability of malaria including mosquito vectors, parasite strain and human host activities, but the focus of this study was to use the environmental factors to prediction of malaria outbreak. Moreover, this study has only covered only 10 districts because it was only feasible to find both climate and malaria data and it has been a challenge because if It was possible to access the data of the whole the country would give the opportunity to train a big dataset which might increase the performance of algorithms so that the models would be more precise and accurate.

According to the results as displayed and discussed, the prediction of malaria outbreak using environmental data has been successful based on machine learning algorithms of classification so by using the data, we have in the study and Random Forest algorithm we are sure at more than 90% to make the correct prediction of malaria outbreak. Therefore, I would like to call different health organizations to adopt and use it as one of the control and mitigation measures of malaria outbreak, so that we can achieve the global target of reducing malaria by 90% by 2030

5.2 Recommendation

The study gives the following recommendations on different perspectives which should be addressed:

- I would like to call, the Rwanda Biomedical Center (RBC), Rwanda Meteorological Agency and other institutions to increase the dissemination of more and good data so that the research should get access to those data and that will affect the quantity of research produced and the insights from the research should contribute to development in socio health economic sector.
- Machine the learning-based prediction has been impressive in dealing with a complex dataset and giving the high-performance predictive models, so I would like to call more research to take part in it, therefore they can start to build the good performer models in different sectors.
- The more sample size increase, the more we get more precise and accurate models, I would like call to further research, to make coverage of the whole country therefore we can best accurate precise model.
- I encourage the technological practitioners the good collaboration in the development of the mobile application used to malaria outbreak prediction, whereas it will be built on the best performer model (Random forest) therefore according to the climate condition, we can predict whether there occurs the malaria outbreak or not, and the community should benefit from it by letting they know when more likely to occur malaria outbreak so that the individual measures should be taken to prevention.
- I would like to call other research to increase the health-related research so that their findings should be the evidence on the different decision taken in health sector.
- I would like to call different health organizations or other organizations to make available more research funds so that it can contribute to the increase in quality and quantity of research produced which can contribute to the well-being of society

6 REFERENCES

1. Modu B, Polovina N, Lan Y, Konur S, Asyhari AT. applied sciences Towards a Predictive Analytics-Based Intelligent Malaria Outbreak Warning System †. :1-20. doi:10.3390/app7080836
2. World Health Organization. *World Malaria Report 2019. Geneva.*; 2019. https://www.who.int/publications-detail/world-malaria-report-2019
3. RBC RM of HM and OPDD-. Rwanda Malaria Indicator Survey 2013. TT -. Published online 2014. http://dhsprogram.com/pubs/pdf/MIS16/MIS16.pdf
4. President U, Initiative M. Rwanda - Malaria Operational Plan FY 2019. Published online 2019.
5. Jepsen S. Malaria control. *Tidsskr Nor Laegeforen*. 2000;120(14):1665-168.
6. WHO. Global technical strategy for malaria 2016-2030. *World Heal Organ*. Published online 2015:1-35. http://apps.who.int/iris/bitstream/10665/176712/1/9789241564991_eng.pdf?ua=1
7. Aliyu A. A HYBRID MODEL FOR PREDICTING MALARIA USING DATA MINING TECHNIQUES Aminu Aliyu. Published online 2017.
8. Malaria 7.1.
9. Partnership RBM. Action and investment to defeat malaria 2015-2030. *Geneva World Heal Organ*. 2008;Available.
10. Ingabire CM, Hakizimana E, Kateera F, et al. Using an intervention mapping approach for planning, implementing and assessing a community-led project towards malaria elimination in the Eastern Province of Rwanda. *Malar J*. 2016;15(1):1-12. doi:10.1186/s12936-016-1645-3
11. Bi P, Tong S, Donald K, Parton KA, Ni J. Climatic variables and transmission of malaria: A 12-year data analysis in Shuchen County, China. *Public Health Rep*. 2003;118(1):65-71. doi:10.1016/S0033-3549(04)50218-2
12. Surveillance CD, Malaria RB. Using Climate to Predict Infectious Disease Outbreaks : A Review. *English*. Published online 2004:118 ST-Using climate to predict infectious dise. http://www.who.int/globalchange/publications/en/oeh0401.pdf
13. Henninger SM. Local climate changes and the spread of malaria in Rwanda. *Health (Irvine Calif)*. 2013;05(04):728-734. doi:10.4236/health.2013.54096
14. Initiative M. President ' s Malaria Initiative Strategy. 2015;(April).
15. Masinde M. Africa ' s Malaria Epidemic Predictor : Application of Machine Learning on Malaria Incidence and Climate Data. Published online 2020:29-37.
16. A. Abisoye O, G. Jimoh R. Comparative Study on the Prediction of Symptomatic and Climatic based Malaria Parasite Counts Using Machine Learning Models. *Int J Mod Educ Comput Sci*. 2018;10(4):18-25. doi:10.5815/ijmecs.2018.04.03
17. Agrawal A, Agrawal H, Mittal S, Sharma M. Disease Prediction Using Machine Learning. *SSRN Electron J*. 2018;(May):6937-6938. doi:10.2139/ssrn.3167431
18. Galitskaya EG, Galitskkiy EB. Classification trees. *Sotsiologicheskie Issled*. 2013;(3):84-88. doi:10.4018/978-1-60960-557-5.ch006
19. Sharma V, Kumar A, Panat L, Karajkhede G. Malaria Outbreak Prediction Model Using Machine Learning. 2016;(January).
20. Kirsch D. *Machine Learning by Judith Hurwitz And.*; 2018.
21. Soofi A, Awan A. Classification Techniques in Machine Learning: Applications and Issues. *J Basic Appl Sci*. 2017;13(September):459-465. doi:10.6000/1927-5129.2017.13.76
22. Dönmez P. Introduction to Machine Learning The Wikipedia Guide. *Nat Lang Eng*. 2013;19(2):285-288. doi:10.1017/s1351324912000290

23. Malaria definition .

24. World Health Organization. WHO malaria terminology. *Who*. Published online 2019:31. http://www.who.int/malaria%0Ahttps://apps.who.int/iris/bitstream/handle/10665/208815/WHO_HTM_GMP_2016.6_eng.pdf?sequence=1

25. WHO. Field guide for malaria epidemic assessment and reporting. *World Health*. 2004;1(September):1-27.

26. Anwar MY, Lewnard JA, Parikh S, Pitzer VE. Time series analysis of malaria in Afghanistan: using ARIMA models to predict future trends in incidence. *Malar J*. 2016;15(1):1-10. doi:10.1186/s12936-016-1602-1

27. Ostovar A, Haghdoost AA, Rahimiforoushani A, Raeisi A, Majdzadeh R. Time series analysis of meteorological factors influencing Malaria in south eastern Iran. *J Arthropod Borne Dis*. 2016;10(2):222-237.

28. Musa M. Malaria Disease Distribution in Sudan Using Time Series ARIMA Model. *Int J Public Heal Sci*. 2015;4:7. doi:10.11591/.v4i1.4705

29. Santosh T, Ramesh D. Artificial neural network based prediction of malaria abundances using big data : A knowledge capturing approach. 2019;7(October 2017):121-126. doi:10.1016/j.cegh.2018.03.001

30. Poirot H. Logistic Regression. Published online 2019.

31. Neeb H, Kurrus C. Distributed K-Nearest Neighbors. Published online 2016:1-17.

32. Singh S, Haddon J, Markou M. Nearest-neighbour classifiers in natural scene analysis. *Pattern Recognit*. 2001;34(8):1601-1612. doi:10.1016/S0031-3203(00)00099-6

33. Breiman L. ST4_Method_Random_Forest. *Mach Learn*. 2001;45(1):5–32. doi:10.1017/CBO9781107415324.004

34. Taheri S, Mammadov M. Learning the naive bayes classifier with optimization models. *Int J Appl Math Comput Sci*. 2013;23(4):787-795. doi:10.2478/amcs-2013-0059

35. Joshi P. *Python Machine Learning Cookbook*.; 2016. http://proquest.safaribooksonline.com.ezproxy.lib.vt.edu/9781786464477

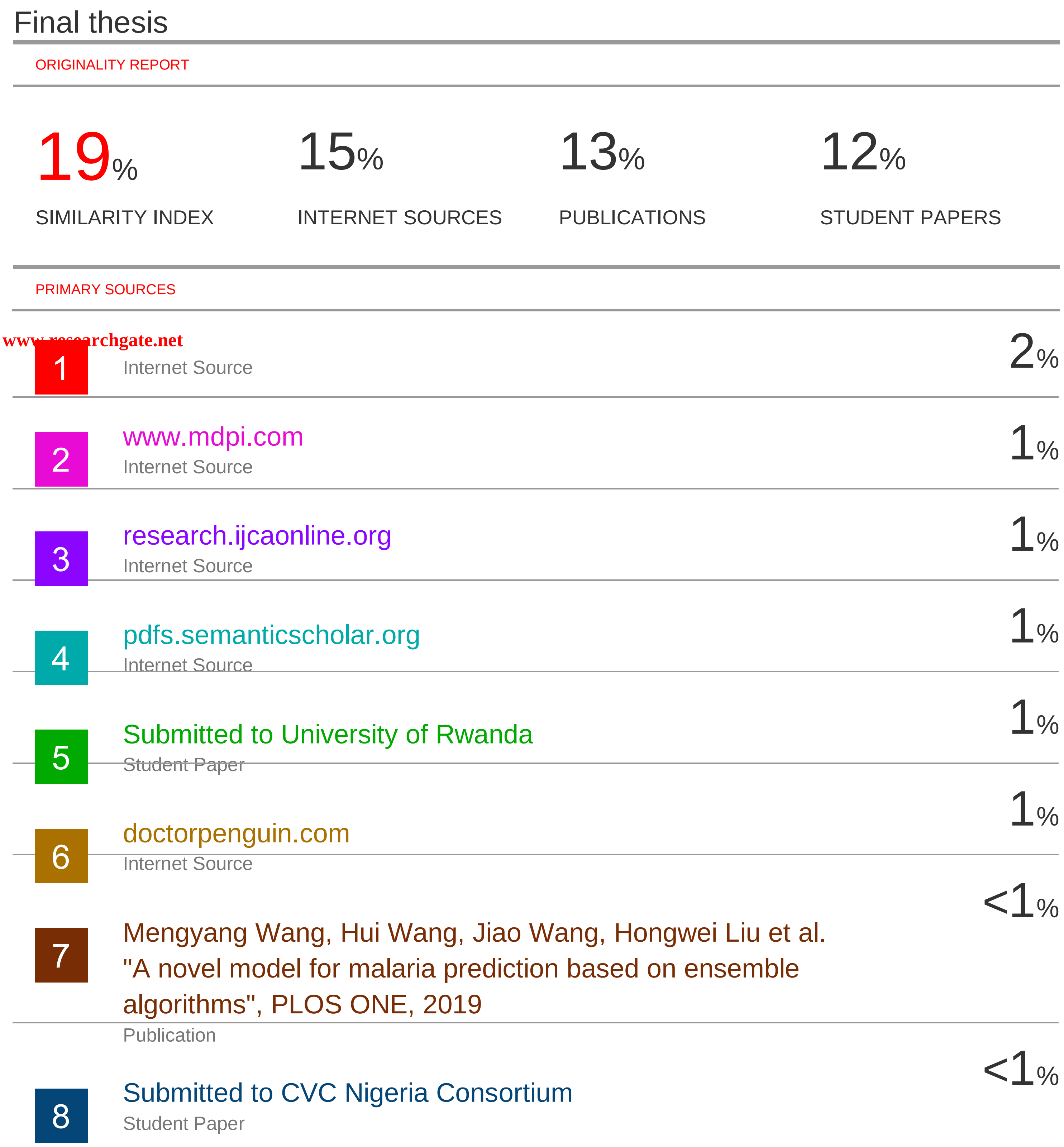
36. Etter PC. Model evaluation. *Underw Acoust Model*. Published online 2010:261-278. doi:10.4324/9780203475652_chapter_11

37. Evaluation Metrics Definition | DeepAI. https://deepai.org/machine-learning-glossary-and-terms/evaluation-metrics

38. Narkhede S. Understanding Confusion Matrix | by Sarang Narkhede | Towards Data Science. *Towar Data Sci*. Published online 2018. https://towardsdatascience.com/understanding-confusion-matrix-a9ad42dcfd62

39. Park SH, Goo JM, Jo CH. Receiver operating characteristic (ROC) curve: Practical review for radiologists. *Korean J Radiol*. 2004;5(1):11-18. doi:10.3348/kjr.2004.5.1.11

Annexes



Final thesis

ORIGINALITY REPORT

19%	15%	13%	12%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	www.researchgate.net Internet Source	2%
2	www.mdpi.com Internet Source	1%
3	research.ijcaonline.org Internet Source	1%
4	pdfs.semanticscholar.org Internet Source	1%
5	Submitted to University of Rwanda Student Paper	1%
6	doctorpenguin.com Internet Source	1%
7	Mengyang Wang, Hui Wang, Jiao Wang, Hongwei Liu et al. "A novel model for malaria prediction based on ensemble algorithms", PLOS ONE, 2019 Publication	<1%
8	Submitted to CVC Nigeria Consortium Student Paper	<1%

7 web.stanford.edu

9	Internet Source	<1%
10	www.rollbackmalaria.org Internet Source	<1%
11	Submitted to Ghana Technology University College Student Paper	<1%
12	Submitted to Gujarat Technological University Student Paper	<1%
13	Submitted to University of Melbourne Student Paper	<1%
14	www.pmi.gov Internet Source	<1%
15	Submitted to Mount Kenya University Student Paper	<1%
16	Submitted to University of East London Student Paper	<1%
17	Submitted to The Robert Gordon University Student Paper	<1%
18	link.springer.com Internet Source	<1%
19	Babagana Modu, Nereida Polovina, Yang Lan, Savas Konur, A. Asyhari, Yonghong Peng. "Towards a Predictive Analytics-Based	<1%

	Intelligent Malaria Outbreak Warning System", Applied Sciences, 2017 Publication	
20	"Proceedings of ICRIC 2019", Springer Science and Business Media LLC, 2020 Publication	<1%
21	worldwidescience.org Internet Source	<1%
22	www.cs.bme.hu Internet Source	<1%
23	ijcrsee.com Internet Source	<1%
24	apanursingpapers.com Internet Source	<1%
25	www.essay.uk.com Internet Source	<1%
26	Submitted to Coventry University Student Paper	<1%
27	Submitted to Babes-Bolyai University Student Paper	<1%
28	Submitted to uva Student Paper	<1%
29	Submitted to essex Student Paper	<1%

30	Submitted to Institute of Professional Studies Student Paper	<1%
31	Submitted to Visvesvaraya Technological University Student Paper	<1%
32	Submitted to uvt Student Paper	<1%
33	es.scribd.com Internet Source	<1%
34	Submitted to University of Southampton Student Paper	<1%
35	www.researchsquare.com Internet Source	<1%
36	"Eco-friendly Computing and Communication Systems", Springer Science and Business Media LLC, 2012 Publication	<1%
37	docplayer.net Internet Source	<1%
38	www.discoverdatascience.org Internet Source	<1%
39	Chang-Yu Wang, Tsair-Fwu Lee, Chun-Hsiung Fang, Jyh-Horng Chou. "Fuzzy Logic-Based Prognostic Score for Outcome Prediction in	<1%

	Esophageal Cancer", IEEE Transactions on Information Technology in Biomedicine, 2012 Publication	
40	Marilyn Milumbu Murindahabi, Domina Asingizwe, P. Marijn Poortvliet, Arnold J.H. van Vliet et al. "A citizen science approach for malaria mosquito surveillance and control in Rwanda", NJAS - Wageningen Journal of Life Sciences, 2018 Publication	<1%
41	doc.lagout.org Internet Source	<1%
42	publications.waset.org Internet Source	<1%
43	tel.archives-ouvertes.fr Internet Source	<1%
44	Submitted to North West University Student Paper	<1%
45	etd.uum.edu.my Internet Source	<1%
46	citeseerx.ist.psu.edu Internet Source	<1%
47	www.aflatoxinpartnership.org Internet Source	<1%
48	Submitted to University of Derby	

	Student Paper	<1%
49	soe.rutgers.edu Internet Source	<1%
50	eprints.whiterose.ac.uk Internet Source	<1%
51	Submitted to Vrije Universiteit Amsterdam Student Paper	<1%
52	Submitted to uu Student Paper	<1%
53	Submitted to Stockholm University Student Paper	<1%
54	Almutairi , Yasamiyan Aweed. "Impact of Sentiment Analysis and Reviewer Profile on Reviews Helpfulness = تأثير تحليل الآراء والملف الشخصي8 للمقيم على جودة التقييم", King Abdulaziz University : Scientific Publishing Centre, 2020 Publication	<1%
55	thesai.org Internet Source	<1%
56	peerj.com Internet Source	<1%
57	www.coursehero.com Internet Source	<1%

58	Submitted to University of Glasgow Student Paper	<1%
59	Submitted to King Saud University Student Paper	<1%
60	Submitted to Pondicherry University Student Paper	<1%
61	Submitted to Hoa Sen University Student Paper	<1%
62	Submitted to University of KwaZulu-Natal Student Paper	<1%
63	Submitted to The University of Manchester Student Paper	<1%
64	repository.tudelft.nl Internet Source	<1%
65	Submitted to National College of Ireland Student Paper	<1%
66	www.fp.utm.my Internet Source	<1%
67	Noman Ali, Nadeem Ullah Khan, Shahid Waheed, Syed Mustahsan. "Etiology of acute undifferentiated fever in patients presenting to the emergency department of a tertiary care center in Karachi, Pakistan", Pakistan Journal of Medical Sciences, 2020	

	Publication	<1%
68	Submitted to MAHSA University Student Paper	<1%
69	Submitted to University of Liverpool Student Paper	<1%
70	biochempress.com Internet Source	<1%
71	ixa2.si.ehu.es Internet Source	<1%
72	Submitted to CSU, San Jose State University Student Paper	<1%
73	Hum Yan Chai, Liang Kim Meng, Hamam Mohamed, Hon Hock Woon, Khin Wee Lai. "Elimination of character-resembling anomalies within a detected region using density- dependent reference point construction in an automated license plate recognition system", Journal of Electronic Imaging, 2016 Publication	<1%
74	"Recent Advances in Information and Communication Technology 2019", Springer Science and Business Media LLC, 2020 Publication	<1%
75	cse.buffalo.edu Internet Source	

76	vc.bridgew.edu Internet Source	<1%
77	Sofia Benbelkacem, Farid Kadri, Baghdad Atmani, Sondès Chaabane. "Machine Learning for Emergency Department Management", International Journal of Information Systems in the Service Sector, 2019 Publication	<1%
78	Submitted to University of Lincoln Student Paper	<1%
79	Sheeba, J. I., and K. Vivekanandan. "Low frequency keyword and keyphrase extraction from meeting transcripts with sentiment classification using unsupervised framework", Proceedings of the Second International Conference on Computational Science Engineering and Information Technology - CCSEIT 12 CCSEIT 12, 2012. Publication	<1%
80	Turkoglu, I.. "A hybrid method based on artificial immune system and k-NN algorithm for better prediction of protein cellular localization sites", Applied Soft Computing Journal, 200903 Publication	<1%
81	arxiv.org Internet Source	

82	www.icongen.in Internet Source	<1%
83	scholarcommons.usf.edu Internet Source	<1%
84	Rainer Lutze, Klemens Waldhor. "Smartwatch based tumble recognition — A data mining model comparision study", 2016 IEEE 18th International Conference on e-Health Networking, Applications and Services (Healthcom), 2016 Publication	<1%
85	Xilei Dai, Junjie Liu, Xin Zhang. "A review of studies applying machine learning models to predict occupancy and window-opening behaviours in smart buildings", Energy and Buildings, 2020 Publication	<1%
86	Submitted to University College London Student Paper	<1%
87	www.ijpttjournal.org Internet Source	<1%
88	Submitted to Eiffel Corporation Student Paper	<1%
89	www.slideshare.net Internet Source	<1%

90 hdl.handle.net
Internet Source

<1%

91 Gao, Hong-Wei, Li-Ping Wang, Song Liang, Yong-Xiao Liu, Shi-Lu Tong, Jian-Jun Wang, Ya-Pin Li, Xiao-Feng Wang, Hong Yang, Jia-Qi Ma, Li-Qun Fang, and Wu-Chun Cao. "Change in Rainfall Drives Malaria Re-Emergence in Anhui Province, China", PLoS ONE, 2012.
Publication

<1%

<1%

92 arno.uvt.nl
Internet Source

<1%

93 Zhongjie Li, Qian Zhang, Canjun Zheng, Sheng Zhou et al. "Epidemiologic features of overseas imported malaria in the People's Republic of China", Malaria Journal, 2016
Publication

<1%

<1%

94 www0.sun.ac.za
Internet Source

95 Rebecca Grealy, Jasper Herruer, Carl L. E. Smith, Doug Hiller, Luke J. Haseler, Lyn R. Griffiths. "Evaluation of a 7-Gene Genetic Profile for Athletic Endurance Phenotype in Ironman Championship Triathletes", PLOS ONE, 2015
Publication

<1%

96 van den Berg, Francesca T., Michael B. Thompson, and Dieter F. Hochuli. "When hot

	rocks get hotter: behavior and acclimatization mitigate exposure to extreme temperatures in a spider", Ecosphere, 2015. Publication	<1%
97	bura.brunel.ac.uk Internet Source	<1%
98	www.edureka.co Internet Source	<1%
99	orca.cf.ac.uk Internet Source	<1%
100	www.who.int Internet Source	<1%
101	realestates.uonbi.ac.ke Internet Source	<1%
102	Farid Zayeri, Masoud Salehi, Hasan Pirhosseini. "Geographical mapping and Bayesian spatial modeling of malaria incidence in Sistan and Baluchistan province, Iran", Asian Pacific Journal of Tropical Medicine, 2011 Publication	<1%
103	"Data Mining and Knowledge Discovery Handbook", Springer Nature, 2010 Publication	<1%
104	Nick Chater. "What is the type-1/type-2 distinction?", Behavioral and Brain Sciences,	

1997
Publication

105 Yishi Zhang, Haiying Wei, Yaxuan Ran, Yang Deng, Dan Liu. "Drawing openness to experience from user generated contents: An interpretable data-driven topic modeling approach", Expert Systems with Applications, 2020 <1%
Publication

<1%
106 Ali Aljofey, Qingshan Jiang, Qiang Qu, Mingqing Huang, Jean-Pierre Niyigena. "An Effective Phishing Detection Model Based on Character Level Convolutional Neural Network from URL", Electronics, 2020
Publication

<1%
107 T. C. Olayinka, S. C. Chiemeke. "Predicting Paediatric Malaria Occurrence Using Classification Algorithm in Data Mining", Journal of Advances in Mathematics and Computer Science, 2019
Publication

<1%
"Machine Learning and Big Data", Wiley, 2020
Publication

<1%
108 Jorge Enrique Rodríguez Rodríguez, Víctor Hugo Medina García, Nelson Pérez Castillo. "Chapter 21 Webpages Classification with Phishing Content Using Naive Bayes Algorithm",

109

Springer Science and Business Media LLC, 2019
Publication

110 Siuly Siuly, Yan Li, Yanchun Zhang. "EEG Signal Analysis and Classification", Springer Science and Business Media LLC, 2016
Publication
<1%

111 Patanjali Kashyap. "Machine Learning for Decision Makers", Springer Science and Business Media LLC, 2017
Publication
<1%

112 Yan Bi, Weiwei Yu, Wenbiao Hu, Hualiang Lin, Yuming Guo, Xiao-Nong Zhou, Shilu Tong. "Impact of climate variability on Plasmodium vivax and Plasmodium falciparum malaria in Yunnan Province, China", Parasites & Vectors, 2013
Publication
<1%

113 Submitted to Intercollege
Student Paper
<