Evaluating the Effectiveness of Zero-Inflated Models and Machine Learning Techniques for Malaria Prevalence: A Comparative Study

By

Sumaira Asghar

Department of Statistics

Faculty of Natural Sciences

Quaid-i-Azam University, Islamabad

2023

بسم الله الرحمن الرحيم

In the Name of Allah The Most Merciful and The Most Beneficent

Evaluating the Effectiveness of Zero-Inflated Models and Machine Learning Techniques for Malaria Prevalence: A Comparative Study

By

Sumaira Asghar

A THESIS SUBMITTED IN THE PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF MASTER OF PHILOSOPHY IN STATISTICS

Supervised By

Dr. Muhammad Yousaf Shad

Department of Statistics

Faculty of Natural Sciences

Quaid-i-Azam University, Islamabad

2023

Declaration

I "Sumaira Asghar" hereby solemnly declare that this thesis titled, "Evaluating the Effectiveness of Zero-Inflated Models and Machine Learning Techniques for Malaria Prevalence: A Comparative Study".

- This work was done wholly in candidature for a degree of M.Phil Statistics at this University.
- Where I got help from the published work of others, this is always clearly stated.
- Where I have quoted from the work of others, the source is always mentioned. Except of such quotations, this thesis is entirely my own research work.
- Where the thesis is based on work done by myself jointly with my supervisor, I have made clear exactly what was done by others and what I have suggested

Dated:____________________ Signature:____________________

Dedication

I am feeling great honor and pleasure to dedicate this research work to

My beloved Family

Whose endless affection, prayers and wishes have been a great source of comfort for me during my whole education period and my life

CERTIFICATE

Evaluating the Effectiveness of Zero-Inflated Models and Machine Learning Techniques for Malaria Prevalence: A Comparative Study

By

Sumaira Asghar

(Reg. No. 02222113007)

A THESIS SUBMITTED IN THE PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF M.PHIL. IN

STATISTICS

We accept this thesis as conforming to the required standards

1.____________________________
Dr. Muhammad Yousaf Shad
(Supervisor)

2.____________________________
Dr. Muhammad Akbar Awan
(External Examiner)

3____________________________ 27/11/23
Prof. Dr. Ijaz Hussain
(Chairman)

DEPARTMENT OF STATISTICS
QUAID-I-AZAM UNIVERSITY
ISLAMABAD, PAKISTAN
2023

Declaration

I "Sumaira Asghar" hereby solemnly declare that this thesis titled, "Evaluating the Effectiveness of Zero-Inflated Models and Machine Learning Techniques for Malaria Prevalence: A Comparative Study".

- This work was done wholly in candidature for a degree of M.Phil Statistics at this University.
- Where I got help from the published work of others, this is always clearly stated.
- Where I have quoted from the work of others, the source is always mentioned. Except of such quotations, this thesis is entirely my own research work.
- Where the thesis is based on work done by myself jointly with my supervisor, I have made clear exactly what was done by others and what I have suggested

Dated:____________________ Signature:____________________

Dedication

I am feeling great honor and pleasure to dedicate this research work to

My beloved Family

Whose endless affection, prayers and wishes have been a great source of comfort for me during my whole education period and my life

Acknowledgments

First and foremost I praise and acknowledge Allah Almighty, The Lord and Creator of the Universe. All respect and gratitude goes to the Holy Prophet Hazrat Muhammad (SAW) who has always been a wonderful source of inspiration for us and whose manner of living illuminates our hearts with the light of Islam. My sincere appreciation goes out to my esteemed supervisor, Dr. Muhammad Yousaf Shad, for his constant guidance and support during the entire research process. I owe him for his advice, especially as it related to my thesis. I feel extremely fortunate to have his leadership. Many thanks to him. I extend my sincere gratitude, admiration, and respect to all other professors Prof. Dr. Ijaz Hussain Kherani (Chairman), Dr. Sajid Ali, Dr. Abdul Haq, Prof.Dr. Javed Shabbir, Dr. Manzoor Khan who instilled within me the knowledge and guided me throughout my research activities. I want to sincerely thank my parents, Mr. and Mrs. Muhammad Asghar for their affection and assistance throughout my life. It would not have been feasible for me to finish this work without their help and encouragement. I want to thank all of my friends and classmates for their assistance and cooperation.

Sumaira Asghar

Abstract

In this study, Malaria is still a serious global health issue that demands new approaches for precise prevalence estimation and efficient control measures. This study undertakes a comprehensive investigation into the evaluation of statistical models and machine learning techniques for estimating malaria prevalence in regions where zero counts are prevalent. The research focuses on zero-inflated data typical of malaria prevalence and seeks to identify the most appropriate modelling strategy for precise prevalence estimation in such circumstances. The research question driving this study is: ”Which zero-inflated model exhibits superior performance in handling zero-inflated malaria data?” To address this question, the study engages in a comparative analysis of various modeling techniques. Zero-inflated models including Zero-Inflated Poisson (ZIP) and Zero-Inflated Negative Binomial (ZINB) are employed alongside traditional Generalized Linear Models (GLM) such as Poisson and Negative Binomial (NB) models. Additionally, hurdle models including Poisson Hurdle and Negative Binomial Hurdle models are incorporated into the analysis to account for the excess zeros in the data. The study uses a comprehensive approach that involves fitting and assessing these various models to actual data on the prevalence of malaria. The primary focus was to assess the efficacy of these models in accurately capturing the characteristics of the data. Our findings revealed varying degrees of model performance. Notably, the ZINB and ZIP models exhibited limitations in effectively handling the complexity of the zero-inflated data, resulting in suboptimal fits. However, contrasting results emerged when employing the negative binomial (NB) negative binomial hurdle (NBH), poisson (P) and poisson hurdle (HP) models, as they consistently demonstrated robust performance across different aspects of the data. These results underscore the importance of carefully selecting an appropriate modeling approach for zero-inflated data analysis. The success of the NB, NBH, P, HP models in our study suggests their potential as effective tools for modeling zero-inflated data with greater accuracy and reliability. This work contributes to the understanding of model selection in the context of zero-inflated data, offering valuable insights for researchers and practitioners working with similar data types. Moreover, advance machine learning algorithms such as support vector machine and random forest model also effectively captured zero inflated data.

Contents

List of Tables

List of Abbreviations

WHO	World Health Organization
AIC	Akaike Information Critrea
RF	Random Forest
SVM	Support Vector Machine
ZIP	Zero-inflated Poisson
ZINB	Zero-inflated negative binomial
HP	Hurdle Poisson
NBH	Negative Binomial Hurdle
GLM	Generalized Linear Model
PMF	Probability Mass Function

Chapter 1

Introduction

Malaria is a dangerous parasitic disease by female Anopheles mosquitoes that have been infected to bite humans. There are five distinct types of parasites, namely Plasmodium malariae, Plasmodium falciparum, Plasmodium ovale, Plasmodium vivax, and Plasmodium knowlesi, which have the capability to cause malaria in humans. The most lethal species among them are Plasmodium vivax and Plasmodium falciparum. Plasmodium falciparum is the most common and fatal parasite in Africa, especially in the Sub-Saharan region. The prevalence of illness and mortality in the general population remains alarmingly high, with the World Health Organization (WHO) estimating an annual occurrence of 300-500 million clinical malaria cases. Among these cases, over 100 million individuals experience severe manifestations, leading to a staggering number of deaths exceeding 100 million. The majority of these cases (80%) and fatalities (over 95%) are concentrated in tropical Africa. Furthermore, approximately 280 million people are carriers of the malaria parasite, and the situation is progressively worsening. The parasites that cause malaria to be drug-resistant is steadily advancing, contributing to the emergence of new malaria outbreaks in previously unaffected regions.

Eritrea, situated in eastern Africa, shares its borders with Djibouti to the southeast, Ethiopia to the south, and Sudan to the west. Positioned within the Horn of Africa, it possesses a considerable coastline along the Red Sea to the north and east.Eritrea's administrative divisions consist of six zones: Gash Barka, Ma'akel (Central), Anseba, Debub (South), Debubawi K'eyih Bahri (Southern Red Sea), and Semenawi K'eyih Bahri (Northern Red Sea). Within these zones, there are a total of 58 sub-zones. Eritrea spans across 117,600 square kilometers and supports a population of 3.72 million. The population density in Eritrea is 45 people per sq km,which is higher in highlands and low in lowlands. Accoring to this the rank of Eritrea in the world in terms of population density is 154th. In the central highlands of Eritrea, the elevation ranges from more than 3000 metres above sea level to below sea level in the dry south. The climate in Eritrea spans from hot and arid along the Red Sea to temperate conditions in the highlands, with isolated micro-catchment areas in the eastern escarpment experiencing a sub-humid climate. In 2015, around 40% of the country experienced precipitation ranging from 300 to 600 mm.

The prevalence of malaria, an infectious disease, is greatly influenced by climatic conditions. The prevalence of malaria has been linked in numerous studies to meteorological variables, particularly temperature and rainfall. For example, rainfall performs an important role in the creation of stagnant water bodies that serve as favorable breeding grounds for disease-carrying vectors. The transmission of malaria is strongly influenced by climatic factors. The distribution and abundance of the malaria parasite and its mosquito vectors are greatly influenced by temperature and rainfall patterns. The prevalence of malaria exhibits an exponential increase in response to moderate rainfall and slightly elevated temperatures, highlighting the disease's sensitivity to climate factors.

Following a widespread malaria outbreak that impacted the entire nation in 1998, Eritrea has made remarkable advancements in key health indicators after implementing the National Malaria Control Program (NMCP) and other comprehensive strategies. As a result, there has been a notable reduction in both malaria incidence and mortality rates in the country. According to Kifle et al. (2019), In Eritrea, considerable advancements in the fight against malaria have been noted. The study reported a substantial decrease in malaria incidence, with a reduction of 89% in 2012 compared to 1998, dropping from 157 cases per 1000 at-risk individuals to 17 cases per 1000 people in 2016. Moreover, there was a remarkable decline of 98% in malaria-specific mortality, decreasing from 0.198 deaths per 1000 individuals in 1998 to 0.004 deaths per 1000 people in 2012. The National Malaria Control Programme (NMCP) has created a thorough national strategy plan to eradicate malaria from Eritrea by 2030 in response to the favourable trend. The decline in malaria death rates, which supports the zero-inflated nature of the response variable, led to the use of zero-inflated models in the data analysis.

1.1 symptoms

When an individual without prior immunity is bitten by a mosquito carrying the malaria parasite, symptoms typically do not appear for at least seven days, and sometimes as long as 10-15 days. The initial indications of malaria, including fever, headache, chills which may be mild and difficult to recognize. However, Plasmodium falciparum malaria, which can result in serious sickness and even death, can turn lethal if untreated within 24 hours. Malaria's most typical signs and symptoms include:

- Sweats.
- Body aches.
- Generally feeling sick.
- Nausea.
- vomiting.
- Anemia.
- Abdominal pain.
- Muscle or joint pain.

- Respiratory distress.
- Rapid heart rate.

In the event that an individual develops a fever while residing in or shortly after visiting a location with a significant risk of malaria, it is crucial for them to promptly seek medical advice from a doctor. Should they encounter severe symptoms, it is essential to seek immediate medical assistance.

1.2 Prevention and treatment

Malaria prevention is crucial in combating this life-threatening disease that affects millions of people worldwide. Implementing effective preventive measures can significantly reduce the incidence and spread of malaria. Here are a few key strategies for malaria prevention:

1. Mosquito control:

   Since infected mosquito bites are the main method of transmission for malaria, managing mosquito populations is crucial. This can be accomplished by utilising bed nets sprayed with pesticides, indoor residual insecticide spraying, and environmental management to get rid of breeding grounds.

2. Antimalarial medication:

   Taking antimalarial medication as prescribed is another critical aspect of prevention. Medications such as chloroquine, artemisinin-based combination therapies (ACTs), and other prophylactic drugs can help prevent the development of malaria parasites in the body.

3. Protective clothing:

   A physical defence against mosquito bites can be created by wearing long sleeves, long pants and socks. Additionally, using repellents on exposed skin and clothing can further deter mosquitoes.

4. Travel precautions:

   It is crucial to take certain measures when visiting areas where malaria is prevalent. Which may include taking antimalarial medication before, during and after the trip, using mosquito nets and repellents, and being aware of the peak mosquito-biting times.

5. Environmental awareness:

   Creating awareness about the importance of maintaining a clean environment is vital. Stagnant water, particularly in puddles, ponds, or discarded containers, serves as breeding grounds for mosquitoes. By promoting good sanitation practices, proper waste management, and drainage systems, communities can significantly reduce mosquito breeding sites.

6. Indoor residual spraying (IRS):

   A crucial vector control technique that can rapidly and efficiently reduce malaria transmission is indoor residual spraying (IRS). This approach entails applying a residual insecticide to the inner walls and ceilings of residential buildings, targeting areas where malaria-carrying vectors are likely to encounter the insecticide.

7. Insecticide-treated mosquito nets (ITNs):

   These are crucial tool in the fight against malaria. These mosquito nets are specially treated with insecticides that repel, disable or kill the mosquitoes carrying the malaria parasite. ITNs can act as a physical barrier to keep mosquitoes from biting and infecting people while they sleep when used appropriately. The insecticides used in ITNs are safe for humans but lethal to mosquitoes, making them an effective means of reducing malaria transmission. We can reduce the drastically spread of malaria and save lives by early detection of malaria and by doing their treatments. ACT is very effective in combating malaria and is the only medication that is 100% effective against the disease. (ACT) consists of two or more drugs that function in different ways against the malaria parasite. The appropriate treatment approach for the patient depends on their clinical condition:

- Age
- The severity of symptoms
- Use of other medications
- The type of parasite causing the infection

1.3 Background of the problem

In this research, Excessive zeros are a common problem when counting the frequency of specific events in research projects, such as the number of fatalities, cigarettes smoked, disease cases and hospitalisations. Count data are the name given to this kind of data, which typically consists of discrete, non-negative numbers with a bottom bound of zero. It is critical to develop models that take these difficulties into account because count data are often used in many disciplines, including public health, medicine and epidemiology. Poisson, Negative binomial (NB) models and logistic models have historically three main models used to analyse count data. The classical Poisson model may not be useful since the zero-inflated count data may show the issue of over-dispersion, under-dispersion, or zero-inflation. Even while NB regression can be used to address over-dispersion, it frequently falls short when addressing zero inflation. Then in addition to logistic, Random Forest (RF) model and Support Vector Machine (SVM), it is also possible to utilise zero-inflated negative binomial (ZINB), zero-inflated Poisson (ZIP) , Negative binomial hurdle (NBH) and Hurdle Poisson model (HP). Both the hurdle and zero-inflated models use count procedure and binomial type process.

The two models' treatment of zeros is where they diverge most. The count distribution does not contribute to the extra zeros, however, because the count process in the hurdle model is zero-truncated. When there are too many sampling and structural zeros in the data, zero-inflated models are employed, whereas hurdle models are used when there are only too many structural zeros. We need zero-inflated models because the abundance of zeros prevents us from using the traditional models for data on counts.

In this research, we used four zero-inflated models (ZIP, ZINB, HP and NBH) to look into how the weather particularly the temperature and rainfall, affects malaria deaths in Eritrea. For comparison, we have also taken into account the classical Poisson, logistic and NB models. We also apply SVM, RF to check how this type of data works on these models.

1.4 Research motivation

Investigating the effects of temperature and rainfall on malaria fatalities is the driving force behind this investigation. More specifically:

- Applying classical models, zero-inflated, hurdle models, Random forest model, and Support vector machineto malaria data in order to examine how temperature and precipitation affect malaria fatalities.

- Utilizing AIC to compare the outcomes of all models and describing which model performs the best.

- Comparing the results of traditional models against zero-inflated models using the Vuong test.

Chapter 2

Literature Review

A literature review is a crucial part of academic research because it offers a thorough overview and critical analysis of the body of work already written on a certain subject. It entails reading, analysing, and synthesising a variety of academic materials, including books, journal articles, and conference papers, in order to pinpoint important themes, concepts, and disagreements pertinent to the research issue or field of study. The goals of a literature review are to demonstrate the researcher's understanding of the subject's current state of knowledge and to pinpoint any gaps or ambiguities that warrant further investigation. Critical thinking skills are required to assess the validity, applicability and limitations of the literature under review. Researchers can identify important theories, approaches and research methods applied in their field by doing a literature study. It aids in the development of a research plan, research questions and research objectives that add to the body of information already known in the field.

The parasite and the mosquito species that spread malaria are very sensitive to climate variables. Numerous studies have examined how climate influences mosquito abundance and malaria transmission. Jury and Kanemba (2007) stated that at a lead time of six months, it was discovered that upper-level winds over the Western Pacific could forecast 57% of malaria variance. Craig et al. (1999) It was said that rainfall produces water bodies that are suitable vector breeding habitats. But simply because there are more bodies of water after a rainstorm does not mean that there are more malaria cases. An adequate amount of rain, however, can increase the number of vectors. Also the growth and biting of the anopheles mosquito are said to be regulated by temperature. Additionally, they stated that malaria transmission was not good at temperatures below 18°C, but was favourable at temperatures above 22°C; temperatures above 32°C also result in high vector populations and significant mortality. Mopuri et al. (2020). (2020) The polynomial model successfully predicted malaria cases in the study area. Caminade et al. (2014) It showed that the decision regarding the malaria impact model has an important bearing on the malaria outcome metrics. Akpan et al. (2019) The higher magnitude of change in species predominance predicted for the latter part of the 21st century under the high emission scenario may result in a more notable increase in malaria burden as compared to previous periods and scenarios within the century. Loevinsohn (1994) stated that malaria and other vector-borne illnesses are predicted to become more prevalent because of the climatic shift. Chuang et al. (2017) indicated that Lubombo and Hhohho administrative regions have warmer temperatures and more precipitation, which encourage the spread of malaria in Swaziland. Dlamini et al. (2018) looked into for effective surveillance, the gaps in reporting and inquiry must be filled. Jury and Kanemba (2007) Upper-level winds over the Western Pacific predicted 57% of the variance in malaria cases with a six-month lead time. Hay et al. (2002)indicated that throughout the previous century or during the time of the supposed malaria comeback in East Africa, there hasn't been a major change in temperature, rainfall or the amount of months that are ideal for P. falciparum transmission. Thomson et al. (2017) the effectiveness of interventions in three countries may be understated in current malaria evaluations. Abiodun et al. (2018) said that using the mosquito model to anticipate mosquito abundance could offer additional information about how to eradicate or control malaria in Africa. Mbokazi et al. (2018) Anopheles merus's greater prevalence and wider distribution in the Ehlanzeni District throughout the research period may have had a significant impact on the regional transmission of malaria in the Mpumalanga Province. Wardrop et al. (2013) showed that rising temperatures raised the risk of malaria in each of the four regions. Wei et al. (2020) To examine the "1-3-7" approach's implementation and the epidemiological aspects of malaria in Yunnan Province in order to collect data for the creation of post-elimination surveillance interventions. Kim et al. (2018) Zero-inflated regression models was employed by researcher as most effective statistical strategy for predicting the number of weekly heat deaths in South Korea.

Feng (2021) They emphasized the distinction in data generation methods between zero-inflated models and hurdle models. To evaluate the models' effectiveness, they conducted simulation studies.

Zeileis et al. (2008) Compared to their classical equivalents, both model classes are more adept at incorporating extra zeros and over-dispersion. Abiodun et al. (2019) demonstrated that temperature and rainfall have a substantial impact on malaria incidence in the study locations of Limpopo Province, South Africa. They utilised the Vuong test to examine if the ZINB regression model or traditional NB regression performed better. The outcome showed that the ZI model performed better. Makinde et al. (2020) The research area's lowest temperature, total annual precipitation and relative humidity all significantly affect the growth in the number of inpatients and outpatients. Khan et al. (2011) They discussed six count models (Poisson, Negative Binomial (NB), Zero-Inflated Poisson (ZIP), Zero-Inflated Negative Binomial (ZINB), Hurdle Poisson (HP), and Hurdle Negative Binomial (NBH)) were employed to evaluate fall count data derived from a prospective cohort study involving women aged 40 to 80. Their findings revealed that the models based on the Negative Binomial distribution outperformed those based on the Poisson distribution. Among the models, NBH demonstrated the closest alignment with the data, surpassing ZINB and NB models. To compare the models, they utilized the Voung test, as well as metrics such as AIC and BIC. Hu et al. (2011) Models such as ZINB, NB, HP, NBH and ZIP were used to analyse how HIV risk reduction initiatives affected the quantity of unprotected sex encounters. ZINB model had best fit when goodness of fit of each model was compared. Feng (2021) To fit such data, hurdle and zero-inflated models are frequently used. Tüzen and Erbaş (2018) studied how many cigarettes young people in Turkey smoke every day. Data from the cigarette consumption count analysis was done using the models mentioned above. NBH and ZINB had superior results. Tan and Yen (2017) In Malaysia, social-demographic traits are strongly correlated with the consumption habits of nonsmokers, occasional smokers and regular smokers. Purnama (2021) They suggested that the Zero Inflated Negative Binomial regression model stands out as the optimal selection for analyzing data pertaining to daily cigarette smoking among individuals in Indonesia. Liaqat et al. (2020) explained the zero-inflated models are very reliable model for calculating and reporting the number of afflicted lymph nodes in primary breast cancer. Lambert (1992) They discussed that response count variable is considered to follow a Poisson distribution, with an additional distribution featuring a probability of p concentrated at the value of zero.

Greene (1994) For examinaion of data with too many zeros and too much dispersion, the ZINB model was proposed. Mullahy (1986) The proposed modified distributions end up as parameter-restricted variations of the known distribution types. Zeileis et al. (2008) described in depth how to do these studies using the R statistical software's pscl module. Vuong (1989) Under the broadest assumptions, the likelihood ratio statistic is directional. Biau (2010) The process is repeatable and accommodates sparsity. Buri and Hothorn (2020)) Existing random forest variants for ordinal outcomes are assessed when prognostic variables have a non-proportional odds influence. Zhang et al. (2013) Even with a limited sample of data, the support vector machine can accurately extract the defect categories. le Cessie and van Houwelingen (1994) When dealing with paired correlated data, the whole likelihood can be assessed.

Chapter 3

Material and methods

3.1 Study Area and Data Description

To address research objectives or test hypotheses, data collecting is a fundamental research procedure that involves gathering and getting appropriate data. The selection of data sources depends on the research topic and may include secondary sources like existing databases, literature, or archive records in addition to primary sources like surveys, interviews, observations or experiments. Effective data collection also helps to the overall validity and validity of study findings. Eriteria, situated within the Horn of Africa, shares its borders with Sudan to the west, Ethiopia to the south, and Djibouti as well as the Red Sea to the southeast. Its varied geography includes hills, semi-arid plains and deserts. The prevalence of diseases like malaria can be influenced by the climate, which ranges from arid to semiarid. As of the year 2021, there are about 6 million people living in Eritrea. In contrast to rural areas, which have a lower population density, urban places, like Asmara, have a larger population density. In Eritrea, malaria has been a major public health issue. The rainy season is when the sickness is most common. Environmental factors that affect mosquito breeding and the transmission of malaria include stagnant water, temperature and humidity. There is an interest in researching the state of Eritrea's health, as seen by your data collection on malaria mortality. Collaborations with regional health authorities, academics and organisations can give your study insightful information and support.

In this research, we looked at the connection between Eritrea's rainfall, temperature and malaria fatalities in Eritrea. We gathered the monthly data from six zones (zoba), which are further divided into 53 sub-zones (sub-zoba), throughout a nine-year period (2009–2017).

Since the supervisor had previously worked on malaria in Eritrea Mihreteab et al. (2020) he was able to give data on malaria deaths. The Health Management Information System (HMIS) set up by the Eritrean Ministry of Health provided them with information on the epidemiology of malaria.

Infrared Rainfall with Stations for Climate Hazards (CHIRPS) project provided data on monthly rainfall (mm). Collaboratively initiated, the project was crafted through the partnership of the United States Geological Survey (USGS) and the University of California, Santa Barbara. To create CHIRPS databases, they integrated in-situ station data with high-resolution satellite images.

You can find monthly temperature and rainfall information for each sub-zoba in Eritrea at https://earlywarning.usgs.gov/fews/ewx/index.html.

3.2 Generalized Linear Model

The integration of a generalized linear model (GLM) into the linear regression framework expands the scope to accommodate a wide range of response variables, even those that do not conform to a normal distribution. It is versatile and effective method frequently employed for data analysis when the underlying premises of ordinary least squares regression are broken. The response variable in a GLM can be categorical, continuous, binary, count, or even countable.

In the context of the linear regression model, the response variable 'y' is established as a linear composite of individual predictors 'x'. However, this model's effectiveness can be limited in count-based scenarios due to its inherent linearity constraint, which regrettably hinders its ability to encompass various real-world situations. Nelder and Wedderburn (1972) therefore proposed the Generalised Linear Model (GLM), an advanced statistical modelling technique to construct a linear relationship between 'y' and 'x' if the distribution of response variable is from the class of non-linear models. In this instance, it is presumptive that distribution of 'y' is member of the exponential family. The exponential family is thought to include probability distributions that satisfy the following function:

$$f(y;\lambda,\psi) = \exp\left\{\frac{(y\lambda - b(\lambda))}{a(\psi)} + c(y,\psi)\right\} \tag{3.1}$$

Where the fixed scale or dispersion parameter λ and the canonical parameter ψ are both variables. Choose an exponential family member, such as the Poisson or NB distribution to use and by adjusting the values of the functions a(.), b(.) and c(.) for various Y distributions.

Distribution of 'y' and the connection function together determine the GLM. The linear combination of the explanatory variable is related to E(Y_i) = μ_i according to link function. It links linear predictor and probability distribution parameter. The function is represented by

$$\eta_i = g(\mu_i) \quad = x_i^T\beta \tag{3.2}$$

In this $g(\mu_i)$ is link function and β is regression coefficient's vector. The link function is also known as the canonical link function if linear predictor η and canonical parameter λ are same,i.e. η=λ

3.2.1 Poisson Model

The Poisson model represents a discrete probability distribution that quantifies the probability of a specific count of events transpiring within a given time frame (or other fixed intervals like distance or area), provided these events transpire at a consistent rate, independent of the time that has elapsed since the preceding event.

If anticipated the interval contains *mu* occurrences then there is probability that will be exactly 'y' occurrences (y being a non-negative integer, y = 0, 1, 2,... The formula of probability distribution can be written as:

$$f_i(y_i) = \frac{e^{-\lambda_i}\lambda_i^{y_i}}{y_i!} \qquad y = 1, 2, 3, \ldots, \lambda > 0 \tag{3.3}$$

When the rate parameter λ is used. Because the Poisson distribution has an equal mean and variance, a larger mean corresponds to a larger standard deviation. The GLM is used to transform the above equation into Poisson regression. The correlation between the prediction function η and the mean of Yi is logarithmic.

$$\log(\lambda_i) = \eta \tag{3.4}$$

The log link makes sure that no values are ever negative when they are fitted. The dispersion parameter ψ in GLM Poisson is also fixed at 1.

3.2.2 Negative Binomial Model

The quantity of events or instances of a particular event is the outcome of interest in negative binomial. It analyses count data using a statistical model. It is a subset of the binomial distribution that represents the overall success rate across a predetermined number of unique Bernoulli trials.

Instead of imposing a fixed number of trials, the negative binomial model focuses on modelling the number of trials required to attain a particular number of successes. This makes it appropriate for circumstances where the number of attempts needed to achieve a particular success rate might vary and the outcome is a count variable. Across various fields such as biology, economics, finance, and epidemiology, the Negative Binomial model finds common application for data analysis in situations where the variance surpasses the mean, leading to over-dispersion within count data. Over-dispersion denotes increased variability in the data, a phenomenon not adequately captured by standard Poisson and binomial distributions.

The negative binomial model can be fitted to the data in practise by using maximum likelihood estimation or other estimation techniques. The model can be expanded to incorporate covariates or predictor variables to investigate the effects of other variables on the count outcome. It provides us best understanding between the relationship of predictor variables and the count outcome of interest. The density function of NB distribution:

$$f(y;\theta,\lambda) = \frac{\Gamma(y+\theta)}{\Gamma(\theta)\Gamma(y+1)}\left(\frac{\theta}{\lambda+\theta}\right)^{\theta}\left(1-\frac{\theta}{\lambda+\theta}\right)^{y} \tag{3.5}$$

In this equation θ and λ are two parameters. E(Y)= λ is mean and Var(Y) $= \lambda + \frac{\lambda^2}{\theta}$ is variance of the distribution. The function of the dispersion parameter, *theta*, is to convert the Negative Binomial distribution into a Poisson distribution. If θ is smaller, over-dispersion becomes larger. The log link function is widely used to verify that fitted values are always non-negative. Other link functions include identity and sqrt. Additionally, it is the **glm.nb** function's default R link in the MASS package. Similar to GLM Poisson, the dispersion parameter ψ in GLM NB is fixed at 1.

Overall, the negative binomial model offers a flexible and interpretable framework for understanding the link between predictor factors and the desired count outcome, making it a helpful tool for analysing count data with over-dispersion.

3.2.3 Logistic Model

Referred to as logistic regression, the logistic model is a statistical approach used to investigate the association between a binary-dependent variable and one or multiple independent variables. The binary nature of the logistic model's dependent variable signifies that it can exclusively assume one of two potential outcomes, often symbolized by the values 0 or 1. Calculating the likelihood that the dependent variable will equal 1 given the values of independent variables is goal of logistic regression.

$$p(x) = \frac{1}{1 + e^{-(x-\mu)/s}} \tag{3.6}$$

In this equation 'μ' is the location parameter, where's' is the scaling parameter and p(mu)=1/2.Now, this phrase can be written as follows:

$$p(x) = \frac{1}{1 + e^{-(\beta_0+\beta_1 x)}} \tag{3.7}$$

In this $\beta_0 = -\frac{\mu}{s}$which is also referred to as intercept (it is the y-intercept or vertical intercept of the line $y = beta_0 + beta_1 x$) and $beta_1 = frac1s$ (inverse scale parameter or rate parameter).

Since around 1970, the binary regression model that has been used most frequently is the logistic model. If a binary variable has more than two possible values, it can be generalized to a categorical variable, and a multinomial logistic regression can be created from a binary logistic regression. Logistic regression is used extensively in the social sciences, machine learning, and the majority of medical fields. For instance, the Trauma and Injury Severity Score (TRISS), which is often used to predict death in injured patients, was created using logistic regression. Using logistic regression, the likelihood of developing a particular disease can be predicted.

3.3 Zero-Inflated Count Data Regression Models

They are frequently used when working with count data that have more zeros than would be predicted by a typical count distribution. For statistical analysis, the distribution of the numbers is frequently represented using negative binomial distribution and Poisson distribution. It is noticed that Poisson regression is typically regarded the fundamental count model that all other count models are built upon. It is noted that several other count models, including Poisson regression, which is sometimes regarded as the fundamental count model, are based on Poisson regression. In a Poisson model, count response is the random variable y, and mean is represented by the lambda parameter ($lambda$). Lambda frequently referred to as the rate or intensity parameter. Lambda is also written as mu (mu) in statistical literature when referring to Poisson and traditional negative binomial models. Zero-inflated data have high percentage of 0 counts. Zero-inflated count data regression models consists of two components: count component and zero-inflation component

Count Component: To represent the count data, this component uses a count distribution like the Poisson or negative binomial distribution. It captures relationship between the independent variables and the counts when they are greater than zero.

Zero-Inflation Component: It simulates the extra zeros present in the data. It accounts for the possibility of observing a zero count as a result of a second operation that produces too many zeros. The existence of a distinct sub population with zero counts or structural factors may be the cause of this additional procedure.

To capture these qualities, Zero-inflated Negative Binomial (ZINB), Zero-inflated Poisson (ZIP), Hurdle Poisson (HP) and Negative Binomial Hurdle (NBH) model widely utilised. Two part and mixture model respectively, these are also known as hurdle models and zero-inflated models. A count process is binomial process (which contrasts zeros and non-zeroes) are included in both models. A mixed model and a two-part model differ primarily in how they handle excess zeros and how they interpret parameters. The distribution of the hurdle count process is zero-truncated, but count process of zero-inflated model can have zeros. According to Lambert (1992) This component use a count distribution to explain the count data, such as the Poisson or negative binomial distribution. One in which participants regularly give zero counts, also called as "structural or excessive zeros" for instance, let's say there were no malaria-related deaths or participants were not at risk of developing malaria. "Sampling zeros" refers to a situation where subjects were made aware of the conclusion but chose not to inform the authorities. The hurdle model Mullahy (1986) makes the assumption that all zeros in data are "structural zeros". When modelling whether or not the answer variable is a positive number, one portion of the model is a zero count process (binary model), and when modelling positive observations of the data, a count procedure utilising a shortened model, such as a truncated Poisson or truncated NB distribution, makes up the remaining component. Mullahy (1986) says the concept behind hurdle formulations is that a binary outcome of rather variable count has a zero or a non-zero outcome is controlled by a model of binomial probability. "Hurdle is crossed" if realisation is non-zero (positive) and conditional distribution of the positives is controlled using data model with truncated counts at zero". (R) offers zeroinfl and hurdle functions in package pscl for analysis of these models. Below is a brief summary of various models.

3.3.1 Zero-Inflated Models

As discussed earlier, zero counts are present for both "structural" and "sampling" origins in zero-inflated distributions. It is believed that the sample zeros, which derive from the standard Poisson distribution, occurred at random. The general formula for the zero-inflated:

$$P(Y_i = y_i) = \begin{cases} \pi_i + (1-\pi_i) \cdot p(y_i = 0; \lambda_i), & \text{if } y_i = 0 \\ (1-\pi_i) \cdot p(y_i; \lambda_i), & \text{if } y_i > 0 \end{cases} \tag{3.8}$$

In this equation $\pi - i$ is probability of structural zero and $p(y_i; \lambda_i)$ is distribution assessed at zero, and $p(y_i = 0; \lambda_i)$ is the PMF of the regular count distribution. The ZIP is written as count distribution exhibits Poisson distribution.

$$P(Y_i = y_i) = \begin{cases} \pi_i + (1-\pi_i) \cdot e^{-\lambda_i} & \text{if } y_i = 0 \\ (1-\pi_i) \cdot \frac{e^{-\lambda_i}\lambda_i^{y_i}}{y_i!} & \text{if } y_i > 0 \end{cases} \tag{3.9}$$

Where λ is mean of standard Poisson distribution. Mean and variance of ZIP:

$$E(y_i) = (1 - \pi_i)\lambda_i \tag{3.10}$$

$$\text{Var}(y_i) = (1 - \pi_i)\lambda_i(1 + \pi_i\lambda_i) \tag{3.11}$$

Over-dispersion would happen if the variance of the data exceeded what ZIP model predicted. ZINB model might then used to simulate both zero inflation and over-dispersion in data. ZINB model is written as:

$$P(Y_i = y_i) = \begin{cases} \pi_i + (1 - \pi_i)\left(\frac{\Gamma(y_i+\theta)}{\Gamma(\theta)\Gamma(y_i+1)}\left(\frac{\theta}{\lambda_i+\theta}\right)^{\theta}\left(1 - \frac{\theta}{\lambda_i+\theta}\right)^{y_i}\right) & \text{if } y_i = 0 \\ (1 - \pi_i)\left(\frac{\Gamma(y_i+\theta)}{\Gamma(\theta)\Gamma(y_i+1)}\left(\frac{\theta}{\lambda_i+\theta}\right)^{\theta}\left(1 - \frac{\theta}{\lambda_i+\theta}\right)^{y_i}\right) & \text{if } y_i > 0 \end{cases} \tag{3.12}$$

Here, mean and shape parameters of negative binomial distribution are λ_i and θ, respectively. This indicates how much over-dispersion there is. Mean and Variance of ZINB are:

$$E(y_i) = (1 - \pi_i)\lambda_i \tag{3.13}$$

$$\text{Var}(y_i) = (1 - \pi_i)\lambda_i\left(1 + \frac{\lambda_i}{\theta} + \pi_i\lambda_i\right) \tag{3.14}$$

When $\theta \to \infty$ ZINB reduces to ZIP model. Typically, to model pi_i and $lambda_i$, respectively, logistic regression and log-linear regression are utilised.

Zero-inflated model can be written as:

$$\log(\lambda_i) = x_{Ti}\alpha, \quad \text{logit}(\pi_i) = z_{Ti}\beta \tag{3.15}$$

The regression coefficients for the covariates $x_i^\top$ and $z_i^\top$ in this equation are denoted by α and respectively.

3.3.2 Hurdle Model

A random variable is separated into two components in a hurdle model, the first of which is the probability of achieving a value of 0, and the second of which is the likelihood of achieving a value other than zero. Adoption of hurdle models is frequently prompted by a data excess of zeros that more conventional statistical models are unable to effectively account for. The hurdle model is likewise referred to as Zero-altered models was introduced by Mullahy (1986). For all non-zero observations, it uses a zero-truncated fraction and treats all zeros as "structural" zeros. The following is a model for a random variable x:

$$\Pr(x = 0) = \theta \tag{3.16}$$

$$\Pr(x \neq 0) = p_{x\neq 0}(x) \tag{3.17}$$

$p_{x\neq 0}(x)$ shows the truncated probability distribution function which truncated at 0. Hurdle model's standard form is written as as:

$$P(Y_i = y_i) = \begin{cases} p_i & y_i = 0 \\ (1 - p_i) \cdot \frac{p(y_i;\mu_i)}{1-p(y_i=0;\mu_i)} & y_i > 0 \end{cases} \tag{3.18}$$

the normal count distribution's PMF is $p(y_i; mu_i)$, and the distribution's evaluation at zero is $p(y_i = 0; lambda_i)$. The Probability distribution for HP is expressed as:

$$P(Y_i = y_i) = \begin{cases} p_i & \text{if } y_i = 0 \\ (1 - p_i) \cdot \frac{\frac{e^{-\lambda_i}\lambda_i^{y_i}}{y_i!}}{1-e^{-\lambda_i}} & \text{if } y_i > 0 \end{cases} \tag{3.19}$$

John G. Cragg introduced hurdle model in 1971, when the non-zero values of x were modelled using a normal model and the zeros were modelled using a probit model. The hurdle model was named because the probit component of the model was designed to simulate the existence of "hurdles" that must be surmounted in order for the values of x to reach non-zero values. Geometric, Poisson and negative binomial models for non-zero counts were later developed as hurdle models for count data.

The NBH is written as follows to take into account over-dispersion:

$$P(Y_i = y_i) = \begin{cases} p_i & \text{if } y_i = 0 \\ \frac{1-p_i}{1-\left(\frac{\theta}{\lambda_i+\theta}\right)^{\theta}} \cdot \frac{\Gamma(y_i+\theta)}{\Gamma(\theta)\Gamma(y_i+1)} \left(\frac{\theta}{\lambda_i+\theta}\right)^{\theta} \left(1-\frac{\theta}{\lambda_i+\theta}\right)^{y_i} & \text{if } y_i > 0 \end{cases} \tag{3.20}$$

In a manner similar to zero-inflated model, Another way to write hurdle model is:

$$\log(\lambda_i) = x_i^T \alpha, \quad \text{logit}(p_i) = z_i^T \beta \tag{3.21}$$

In this α is regression coefficient for the covariate x_i and β is regression coefficient for covariate z_i. Hurdle model does not take into consideration the excessive zeros in the count distribution, which is the fundamental distinction between this model and the typical zero-inflated count models. In a mixture model, the primary distribution and the mixing weight both affect the likelihood that the variable will have a value of zero. Model for a random variable x that is especially zero-inflated

$$\Pr(x = 0) = \pi + (1-\pi) \times p(x = 0)$$
$$\Pr(x = h_i) = (1-\pi) \times p(x = h_i)$$

where π is the mixture weight that decides how much zero inflation there will be. The only thing a zero-inflated model can do is enhance the likelihood that $displaystylePr(x = 0)$, but hurdle models are not constrained by this.

3.4 Random Forest Model

Random Forest is powerful machine learning technique that may used for both regression and classification. Utilizing an ensemble technique, a random forest model is built by amalgamating multiple small decision trees or estimators, each yielding distinct sets of predictions. To enhance prediction accuracy, the random forest model combines the predictions generated by these estimators.

Regression trees, sometimes known as decision trees, are the foundation of RF. The node that houses the relevant variable is known as the root node. Leaf nodes or terminal nodes are the names given to the bottom nodes. Each leaf node of the dependent variable contains a value, while the roots or decision nodes receive input from the independent variables. For regression issues, the tree at this fork can be thought of as the averaging of the explanatory variable with continuous nodes. In the decision tree, a variable may appear multiple times at various nodes. In random forests, the self-adaptive method is used to repeat the training data samples and the observation uses the same sample size after the samples are deleted. Each sample generates a decision tree, and each tree provides a data point. By averaging the estimates of the outcomes of these trees, the projected data point of the dependent variable in a regression issue is established.

3.4.1 Random Forest Working Technique

1. The method randomly selects a subset of features to split on at each step or node. The result ensemble's unpredictability is increased and the correlation between trees is decreased. In classification tasks, the quantity of features deliberated at each node typically equals the square root of the overall feature count, while in regression tasks, this quantity corresponds to one-third of the total features.

   $$m_{\text{try}} = \tfrac{t}{3}$$

2. With replacement by the RF method, a random subset of the training data is selected. A decision tree is trained using this technique, often referred to as bootstrap aggregating or bagging, on each subset. It is referred to as a bootstrap sample. The bootstrap sample is created by selecting n observations at random from Sn and replacing them. Each observation has

an equal chance of being chosen, or 1/n.

3. According to bagging characteristics, when the training data set for each tree is randomly sampled, about one-third of the data samples from the initial sample set will never make it into the training set. For each sample k, approximately one-third of the trees exist but are absent from that data sample. Out of bag data set is the name given to this set of data. Then, sample k is predicted using these trees. The OOB (Out of Bag) error is calculated as follows:

   $$\text{ErrorOOB} = \tfrac{1}{n}\textstyle\sum_{k=1}^{n}(z_k - \hat{z}_k)^2$$

   If m is the total number of samples, z_k stands for sample k's real value, and $\hat{z}_k$ stands for sample k's predictive value. The RF generalisation error may be calculated using the OOB error.

4. The algorithm generates predictions by integrating the output of each decision tree once each tree has been trained. The method uses a majority vote to establish the final class prediction for classification tasks. It uses the average of each tree forecast to solve regression problems.

   $$\tfrac{1}{q}\textstyle\sum_{k=1}^{n} z_k$$

   In this equation, the anticipated value of the kth tree is represented by zi, and q is the total number of trees in RF.

3.5 Support Vector Machines

Vapnik and his associates created the SVM in 1995 at the "AT and T. Bell" research facilities. SVM's initial focus was on classification issues. However, it later found widespread use in a variety of different fields, such as regression issues, the identification of intricate patterns, and time series data prediction. SVM is renowned for being designed to effectively simplify training set results in addition to being successful at ranking. As a result, in recent years, the SVM technique has become well-known for forecasting time series. SVM suggests building a system with a high degree of generalisation using SRM principles. To handle a particular complexity, SVM needs a subset of training data points only to be trained known as support vectors. Even though the fundamental task of SVM in this case is to solve a linearly limited quadratic optimisation problem. As a result, unlike other common stochastic or neural network approaches, SVM always produces unique and optimum results.

The most exceptional aspect of SVM is its independence in analysing the consistency and profundity of a method, regardless of the feature space. The input points are frequently displayed in SVM applications with the aid of a few different feature knock-on kernels, which also generalise well for big measurements. Numerous SVM forecasting techniques have been discovered in earlier investigations.

3.5.1 SVM Working Techniques

Support vector machines (SVMs) have the following operational methodology:

1. As input data, labelled feature vectors are employed, and each feature vector is assigned to one of two classifications. Multi-class classification issues can also be handled by SVM.

2. By calculating the similarity between two feature vectors, the kernel function K(zi, zj) transforms the input characteristics into higher-dimensional space. Most popular kernel operations are:

   Linear Kernel:

   $$K(z_i, z_j) = z_i^T z_j$$

   Polynomial Kernel:

   $$K(z_i, z_j) = (\gamma z_i^T z_j + s)^d$$

Radial Basis Function (RBF) Kernel:

$$K(z_i, z_j) = \exp(-\gamma \|\mathbf{z}_i - \mathbf{z}_j\|^2)$$

where γ, s, and d are the kernel parameters.

3. SVM requires labelled training data, where each value of the data is given a class label yi, where yi is a binary value of $y_i \in \{-1, 1\}$ classification.

4. A hyperplane that maximises the difference between the two training classes is found by SVM. The hyperplane is denoted by the equation $w^T z + \beta = 0$ where w is a vector of weights and β is a bias.

5. Support vectors are the data points that are either on the margin or inside the margin. The main components that establish the decision boundary are these support vectors.

6. A new data point z's class label can be predicted using the decision function, which can be written as:

$$f(z) = \text{sign}(w^T \phi(z) + b)$$

where $\phi(z)$ is the feature vector created by the kernel function's transformation of the input data.

7. SVM uses the following optimisation problem to determine the hyperplane:

$$\min \tfrac{1}{2}\|w\|^2 + C \textstyle\sum_{i=1}^{n} \delta_i$$

$$\text{subject to } y_i(w^T z_i + b) \geq 1 - \delta_i \quad \text{and} \quad \delta_i \geq 0$$

In this case, the inequality constraints make sure that all of the data points are correctly categorised or fall inside the margin, whereas C and δ_i are the regularisation parameter and the slack variables, respectively, that allow for misclassifications.

Chapter 4

Results and Conclusions

4.1 Introduction

The GLM NB model is used since both model showed over-dispersion in Poisson GLM. Now, Perumean-Chaney et al. (2013) use hurdle models and zero inflated to deal with zero inflation. claimed that disregarding zero inflation might lead to Type-II error, while disregarding over-dispersion within zero inflation could lead to Type-I mistake and inaccurate estimation. Therefore, to investigate the over-dispersion and zero-inflation in the response variable of our data, we used ZIP, ZINB, HP, and NBH models. For model selection, AIC and Voung tests are employed. We ran GLM Poisson and GLM NB analyses in R using the glm with family=poisson and glm.nb functions from the MASS package. The zeroinfl and hurdle functions from the pscl package are used for zero-inflated and hurdle models, respectively. We have also discussed the effectiveness of the logistic regression, random forest, and SVM models on a particular dataset will be investigated and compared in this study. We may select the model that will work best for our classification issue and create precise predictions by considering the advantages and disadvantages of each model.

4.2 Model Estimation and Selection

Since many people in Eritrea have never had malaria or seen any malaria deaths, we fitted ZINB, ZIP and Hurdle regression models to malaria data. Models with zero inflation may offer a good fit to the facts. Different link functions, like logit or probit, among others, can be used to model the zero-inflated models and explain the zero counts. Logit link was taken into consideration for this study. For the ZIP, ZINB, HP and NBH models, we have consider the Poisson and NB distributions for positive counts. Zero counts can be individually modelled using above link functions, whereas non-zero counts can be modelled using classical distributions. Zero-inflated is a mixture model, hurdle is a two component model, and zero counts can each be individually represented using the aforementioned link functions, as we have already demonstrated.

Table 4.1: Count and Zero-Inflation Model Coefficients with Significance

	Count Model (Negative Binomial)			
	Estimate	Std. Error	z value	P-value
(Intercept)	−6.9596	0.8538	−8.151	0.0000***
Rainfall	0.0108	0.0023	4.648	0.0000***
Temperature	0.0597	0.0188	3.174	0.0015**
Log(theta)	9.1261	93.6194	0.097	0.9223
	Zero-Inflation Model (Binomial)			
	Estimate	Std. Error	z value	P-value
(Intercept)	74.8633	97.5889	0.767	0.4430
Rainfall	0.6271	0.6817	0.920	0.3580
Temperature	−6.2608	7.7587	−0.807	0.4200

For count model:

Rainfall has positive and statistically significant effect on Malaria Deaths (Estimate = 0.0108, p-value=<0.0000). Temperature also has positive and statistically significant effect on Malaria Deaths (Estimate = 0.0597, p-value= <0.0015). The intercept is negative and statistically significant, indicating a baseline level of Malaria Deaths when both Rainfall and Temperature are zero (Estimate = -6.9596, p-value= <0.0000).

For the zero-inflation model:

None of the predictors (Rainfall and Temperature) in the zero-inflation model are statistically significant (all p-values <0.05). The intercept in zero-inflation model is not statistically significant, suggesting that there is no evidence of excess zeros in the data (Estimate = 74.8633, p-value = 0.4430). Consequently, both Rainfall and Temperature have a positive and statistically significant association with Malaria Deaths, while zero-inflation component of model makes only a small contribution to explaining the extra zeros in the data.

Table 4.2: Negative Binomial Model

	Estimate	Std. Error	z value	P-value
(Intercept)	−6.9530	0.8558	−8.124	0.0000 ***
Rainfall	0.0080	0.0022	3.645	0.0003 ***
Temperature	0.0612	0.0189	3.249	0.0012 **
Significance codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1				
Theta: 331				
Number of Fisher Scoring iterations: 1				
Log likelihood: −514.4711 on 4 Df				

With p-values less than 0.05, all three coefficients Intercept, Rainfall and Temperature are statistically significant, indicating that there is sufficient data to reject the null hypothesis and draw the conclusion that these variables are related to malaria deaths.

The negative binomial regression model shows that both Rainfall and Temperature have a significant association with Malaria Deaths. However, given the warning about convergence issues, it is advisable to investigate further and possibly consider alternative modeling approaches to ensure the reliability of the results.

Table 4.3: Poisson Model

	Estimate	Std. Error	z value	P-value
(Intercept)	−6.9530	0.8557	−8.125	0.001 ***
Rainfall	0.0080	0.0022	3.646	0.0003 ***
Temperature	0.0612	0.0188	3.249	0.0012 **
Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1				
Number of Fisher Scoring iterations: 6 Log likelihood: −514.4683 on 3 Df				

The provided statistical output appears to be from Poisson regression model. It investigates the link between two predictor factors (temperature or rainfall) and the response variable (malaria deaths). Intercept estimated to be approximately -6.9530. It represents the estimated log expected count of Malaria Deaths when both Rainfall and Temperature are zero.

In conclusion, the Poisson regression model suggests that both Rainfall and Temperature are significantly associated with Malaria Deaths. However, it's essential to further investigate the model's assumptions and possible alternative models to ensure the reliability of the results.

Table 4.4: Hurdle Poisson model

	Count model coefficients (truncated Poisson with log link)			
	Estimate	Std. Error	z value	P-value
(Intercept)	−6.9608	4875.2883	−0.001	0.999
Rainfall	−0.0656	62.0125	−0.001	0.999
Temperature	−0.2099	117.1776	−0.002	0.999
	Zero hurdle model coefficients (binomial with logit link)			
	Estimate	Std. Error	z value	P-value)
(Intercept)	−6.9850	0.8631	−8.093	5.81 ***
Rainfall	0.0081	0.0022	3.676	0.0002***
Temperature	0.0623	0.0190	3.276	0.0011**
Significance codes: *** p¡0.01, ** p¡0.05, * p¡0.1				
Number of iterations in BFGS optimization: 17 Log-likelihood: -513.5 on 6 Df				

Truncated Poisson Model:

The truncated Poisson distribution with log link function used to evaluate relationship between response variable and predictor variables (rainfall and temperature). The huge p-values (all near 1.000) suggest that the coefficients do not seem to be statistically significant. The log anticipated count of the response variable when both Rainfall and Temperature are zero is represented by the intercept, which is -6.9608.

Zero Hurdle Model:

The model uses binomial distribution with logit link function to estimate likelihood of non-zero outcome (hurdle) and relationship between response variable and predictor variables (rainfall or temperature) for non-zero outcomes. The estimated log-odds of an outcome being non-zero are displayed in the coefficients for each unit increase in the related predictor variable. Statistically significant coefficients for temperature and rainfall exist. Overall, the findings show that the zero hurdle model outperforms the truncated Poisson model in terms of statistical significance for the predictor variables (Rainfall and Temperature). Given the importance of the coefficients in the zero hurdle model, it is possible that temperature and rainfall influence the likelihood that the response variable will not be zero. To make firm conclusions on the links between the variables, more thorough study and model review would be necessary.

Table 4.5: Negative Binomial Hurdle Model

	Count Model Coefficients (Truncated NegBin with Log Link):			
	Estimate	Std. Error	z value	P-value
(Intercept)	-7.0326	1188.8318	-0.006	0.995
Rainfall	-0.1010	358.4004	0.000	1.000
Temperature	-0.3018	158.1994	-0.002	0.998
Log(θ)	0.0513			
	Zero Hurdle Model Coefficients (Binomial with Logit Link)			
	Estimate	Std. Error	z value	P-value
(Intercept)	-6.9850	0.8631	-8.093	5.81 ***
Rainfall	0.0081	0.0022	3.676	0.0002***
Temperature	0.0623	0.0190	3.276	0.0011**
Significance codes: *** p ¡ 0.01, ** p ¡ 0.05, * p ¡ 0.1				
Theta: count = 1.0526 Number of iterations in BFGS optimization: 27 Log-likelihood: -513.2326 on 7 Df				

The provided statistical output includes coefficients from two models: a count model (truncated negative binomial with a log link) and a zero hurdle model (binomial with a logit link). In the truncated negative binomial model, intercept and predictor coefficients (Rainfall and Temperature) are not statistically significant. The high p-values (which are all very close to 1.000) imply that the coefficients do not deviate substantially from zero. The log-odds of an outcome being non-zero when both Rainfall and Temperature are 0 are represented by the intercept, which is -6.9850.

Overall, the findings show that, in comparison to truncated negative binomial model, zero hurdle model offers statistically significant coefficients for the predictor variables (Rainfall and Temperature). Given the importance of coefficients in the zero hurdle model, it is possible that temperature and rainfall influence the likelihood that the response variable will not be zero.

Table 4.6: Zero-inflated Poisson regression model

	Count model coefficients (Poisson with log link)			
	Estimate	Std. Error	z value	P-value
(Intercept)	-7.0643	0.8590	-8.224	0.0000 ***
Rainfall	0.0108	0.0023	4.602	0.0000 ***
Temperature	0.0621	0.0189	3.290	0.0010 **
	Zero-inflation model coefficients (binomial with logit link):			
	Estimate	Std. Error	z value	P-value
(Intercept)	110.53	266.95	0.414	0.6790
Rainfall	24.48			
Temperature	-154.13			
Significance codes: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$				
Number of iterations in BFGS optimization: 240 Log-likelihood: -509.5 on 6 Df				

* The notation: $p < 0.01$ means p-value is less than 0.01.
** The notation: $p < 0.05$ means p-value is less than 0.05.
*** The notation: $p < 0.1$ means p-value is less than 0.1.

In the Poisson model, the intercept and predictor coefficients for temperature and rainfall are both statistically significant. These variables are linked to the response variable (dependent variable), according to the significance codes (***, **, and *) and low p-values (all less than 0.05). The coefficient for the intercept in the zero-inflation model is not statistically significant, according to the relatively high p-value (0.679), which is indicative of this. It means that there is no statistically significant difference between the intercept's log-odds of a zero result and zero. The Poisson model concludes by indicating that the response variable is statistically significantly predicted by both rainfall and temperature. The zero-inflation model, however, does not offer statistically significant coefficients for the predictor variables, and it appears to be difficult to estimate these coefficients because NaN values are present. Addressing the problems with the zero-inflation model is crucial, and the results should be interpreted in light of the particular data and research question.

Table 4.7: Estimation of coefficients using different count data models for malaria deaths in Eritrea

Models	P	NB	ZIP	ZINB	HurdleP	HurdleNB
Covariates for positive counts:						
(Intercept)	-6.9529***	-6.9530***				
Rainfall	0.0079***	0.0079**				
Temperature	0.0612**	0.0061**				
(Intercept)			-7.0643***	-6.9596***	-6.9608	-7.0326
Rainfall			0.0100***	0.0100***	-0.0650	-0.1000
Temperature			0.0620**	0.0590**	-0.2090	-0.3010
Covariates for zero counts:						
(Intercept)			110.53	74.8633	-6.985***	-6.985**
Rainfall			24.4800	0.6271	0.008***	0.008**
Temperature			-154.1300	-6.2608	0.062**	0.062**
AIC	1034.9000	1036.9000	1030.9000	1032.3000	1039.0000	1040.4000
Log Likelihood	-514.4680	-514.4710	-509.5000	-509.2000	-513.5000	-513.2320
Num. of Obs.	6264	6264	6264	6264	6264	6264

***p<0.001, **p<0.01, *p<0.05

This table summarises the coefficients and levels of significance of the predictor variables for each model. It reveals that while Temperature is significant in some models, Rainfall consistently has a statistically significant positive effect on positive counts across all models. In hurdle models and zero inflated models, the Intercept and a few predictor factors have an impact on the presence of excess zeros. AIC and log-likelihood values can be used to compare the goodness of fit of various models.

4.2.1 AIC and BIC

The Akaike information criterion (AIC) and the Bayesian information criterion (BIC) are statistical measurements used in model selection and evaluation. To determine the best model for a particular dataset, both criteria strive to strike a compromise between the trade-off between model complexity and goodness-of-fit. Hirotugu Akaike created the AIC (Akaike Information Criterion), which is based on ideas from information theory. It is frequently applied to maximum likelihood estimation. AIC measures a model's quality by taking into account both the model's complexity and how well it matches the data. The formula for AIC is:

$$\text{AIC} = 2k - 2\ln(L)$$

The letter k denotes the degree of freedom (number of model parameters). The likelihood function's natural logarithm, or ln(L), serves as a gauge of how well the model fits on data. The AIC's objective is to reduce the value. Lower AIC values suggest a more suitable model because they show a better balance between model fit and complexity.

BIC (Bayesian Information Criterion): Gideon Schwarz proposed the BIC, commonly referred to as the Schwarz Information Criterion. It is comparable to AIC but uses a Bayesian methodology. Like AIC, BIC takes into account both model fit and complexity, but it penalises model complexity more severely. The BIC formula

Table 4.8: Model comparison using AIC

Models	AIC
P	1034.9
NB	1036.9
ZIP	1030.9
ZINB	1032.3
HP	1039.0
NBH	1040.4

is:

$$\text{BIC} = k \cdot \ln(n) - 2\ln(L)$$

The number of parameters (degrees of freedom) for the model is k. The total number of data points is n. The likelihood function's natural logarithm, or ln(L), serves as a gauge of how well the model fits on data. The aim is to reduce the BIC value, much like AIC. However, BIC has a stronger propensity than AIC to favour simpler models. Consequently, compared to AIC, BIC frequently chooses a more frugal model.

Table 4.9: Model comparison using BIC

Models	BIC
P	1055.1
NB	1063.9
ZIP	1071.3
ZINB	1079.5
HP	1079.4
NBH	1087.6

4.2.2 Voung test

Vuong test mostly used to compare two non-nested models, which are models that are not directly comparable in terms of likelihoods and degrees of freedom. The definition of a non-nested model is one where one cannot be achieved by putting constraints on the other. The test statistic has asymptotically standard normal distribution under the null hypothesis that the models cannot be distinguished.

For example, Comparing zero-inflated count models to their non-zero-inflated equivalents, for instance (e.g., ZIP versus regular Poisson or ZINB versus regular negative-binomial). The test statistic is asymptotically standard normal when the null hypothesis is true, which states that the models cannot be discriminated.

Table 4.10: Comparing models using the Vuong test

Model Comparison	Vuong Test Statistic	P-value	Preferable Model
P vs. NB	4.9220	0.0000	P
P vs. ZIP	-3.1640	0.0007	ZIP
P vs. ZINB	-3.2855	0.0005	**ZINB**
P vs. HP	-4.7962	0.0000	HP
P vs. NBH	-4.7962	0.0000	NBH
NB vs. ZIP	-3.1652	0.0007	ZIP
NB vs. ZINB	-3.2867	0.0005	**ZINB**
NB vs. HP	-4.7967	0.0000	HP
NB vs. NBH	-4.7967	0.0000	NBH
ZIP vs. ZINB	-1.1045	0.1347	**ZINB**
ZIP vs. HP	2.7049	0.0034	ZIP
ZIP vs. NBH	2.7049	0.0034	ZIP
ZINB vs. HP	2.8406	0.0023	**ZINB**
ZINB vs. NBH	2.8406	0.0023	**ZINB**
HP vs. NBH	-5.0263	0.0000	NBH

***p¡0.001, **p¡0.01, *p¡0.05

Since AIC doesn't indicate which models are truly superior, we are aware that it cannot help us find the best model on its own. Since not all of the aforementioned models were nested within one another, the null hypothesis that models were indistinguishable was utilised to compare the aforementioned models further using Vuong tests. The first comparison was made between the Poisson model and the NB model, with a Vuong test , indicating that NB model was more preferred. NB model was then compared with the next model. ZINB was determined to be the best model after number of tests and model comparisons. ZIP could be viewed as a second choice compared to ZINB.

4.3 Machine Learning Models

By using machine learning models like Logistic Regression, Support Vector Machines (SVM), and Random Forest, we may successfully handle a variety of data-driven challenges. It is appropriate for binary classification tasks to use logistic regression to forecast the probability of a binary result. SVM, on the other hand, has the advantage of being useful for both binary and multiclass classification since it can determine the optimum hyperplane to split data points in a high-dimensional space. A flexible ensemble technique called Random Forest uses the advantages of several decision trees to boost prediction accuracy and handle complex data linkages. By utilising these models, we can obtain outstanding performance across a variety of applications, such as finance and healthcare, image identification, and natural language processing.

4.3.1 Logistic Model

Table 4.11: Logistic Model

	Estimate	Std. Error	z value	P-value
(Intercept)	−6.4631	0.9403	−6.873	0.001 ***
Rainfall	0.0071	0.0025	2.868	0.0041 **
Temperature	0.0512	0.0209	2.455	0.0141 *
Significance codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1				
Number of Fisher Scoring iterations: 7				
AIC: 832.42				

The log-odds of a response variable can be predicted using this model based on the values of the predictor variables (temperature and rainfall). Due to their small p-values, rainfall and temperature are both statistically significant predictors of the response variable. The intercept is likewise substantial, indicating that there is a significant non-zero baseline log-odds even with all predictor variables at zero.

4.3.2 Support Vector Machine

Table 4.12: Summary of SVM Model

Parameter	Value
SVM-Type	C-classification
SVM-Kernel	radial
cost	1
Number of Support Vectors	422
Number of Classes	2
Levels	0 1

As indicated by the SVM-Type being set to ”C-classification,” this SVM model is intended for binary classification jobs, where the objective is to divide data points into two classes. Overall, the data outlines the fundamental traits of the SVM model you specified. With a cost value of 1, it is a binary classification model utilising the radial kernel. The model is using two classes with the labels ”0” and ”1” and was trained using 422 support vectors. Additional context and possibly more analysis would be needed for the model's specific application and performance. The accuracy of the model is 0.9835, which means that nearly 98.35% of its predictions were accurate. The accuracy's 95% confidence interval is reported as (0.9795, 0.9868), which implies that the model's actual accuracy should fall within this range with 95% certainty. This is the accuracy attained by a straightforward model that consistently predicts the most prevalent class. The no information rate in this instance is also 0.9835, indicating that the model's accuracy is comparable to that of a basic model that forecasts the most typical class.This p-value indicates that there is insufficient evidence to draw conclusion that accuracy of the model differs significantly from the no information rate. The model's predictions and the actual results did not agree beyond chance, as indicated by the zero kappa coefficient. Sensitivity (true positive rate) is 1, which

Table 4.13: Confusion Matrix for Pre SVM Train

	Metric	Value
1	Accuracy	0.9835
2	95% CI	(0.9795, 0.9868)
3	No Information Rate	0.9835
4	P-Value [Acc > NIR]	0.5293
5	Kappa	0
6	McNemar's Test P-Value	0.0001
7	Sensitivity	1
8	Specificity	0
9	Pos Pred Value	0.9835
10	Neg Pred Value	NaN
11	Prevalence	0.9835
12	Detection Rate	0.9835
13	Detection Prevalence	1
14	Balanced Accuracy	0.5

means that all positive cases were properly predicted by the model. Calculated as the average of sensitivity and specificity, the balanced accuracy is 0.5. This number indicates that the model's ability to discriminate between the two classes is weak. In summary, even though the accuracy appears to be high, the lack of specificity and balanced accuracy raise the possibility that the model is having trouble correctly predicting the negative class, which has an effect on how well it performs as a whole. The accuracy of the model is 0.9847, which means that

Table 4.14: Confusion Matrix for Pre SVM Test

	Metric	Value
1	Accuracy	0.9847
2	95% CI	(0.9764, 0.9906)
3	No Information Rate	0.9847
4	P-Value [Acc > NIR]	0.5591
5	Kappa	0
6	McNemar's Test P-Value	0.0000
7	Sensitivity	1
8	Specificity	0
9	Pos Pred Value	0.9847
10	Neg Pred Value	NaN
11	Prevalence	0.9847
12	Detection Rate	0.9847
13	Detection Prevalence	1
14	Balanced Accuracy	0.5

around 98.47% of its predictions were accurate. The accuracy range with a 95% confidence level is reported as (0.9764, 0.9906), demonstrating that the model's true accuracy is probably contained within this range. The accuracy having a higher p-value than the no information rate is 0.5591. This p-value indicates that there is insufficient evidence to draw conclusion that the accuracy of the model differs significantly from the no information rate. The model's predictions and the actual results did not agree beyond chance, as indicated by the zero kappa coefficient. The performance of the two models differs significantly, as indicated by the p-value of 0.0000. Sensitivity (true positive rate) is 1, indicating that all positive cases were accurately predicted by the model. Specificity (true negative rate) is 0, which means that no negative cases were accurately predicted by the model. The model correctly predicted 98.47% of the positive cases, which is known as the positive predictive value (precision) of 0.9847. Given that the model failed to identify any negative situations, the negative predictive value is "NaN," which is most likely the cause. Calculated as the average of sensitivity and specificity, the balanced accuracy is 0.5. This number indicates that the model's ability to discriminate between the two classes is subpar. In conclusion, the model appears to have a high level of accuracy, but, like the prior interpretation, it struggles with specificity and balanced accuracy. It's important to remember that while high sensitivity suggests the model is correctly detecting positive cases, the lack of sensitivity does not.

4.3.3 Random Forest Model

Table 4.15: Confusion Matrix and Statistics (Train)

Reference	Prediction	0	1
	0	4879	2
	1	0	80
Accuracy		0.9996	
95% CI		(0.9985, 1)	
No Information Rate		0.9835	
P-Value [Acc < NIR]		0.0001	
Kappa		0.9874	
Mcnemar's Test P-Value		0.4795	
Sensitivity		1.0000	
Specificity		0.9756	
Pos Pred Value		0.9996	
Neg Pred Value		1.0000	
Prevalence		0.9835	
Detection Rate		0.9835	
Detection Prevalence		0.9839	
Balanced Accuracy		0.9878	

The table appears to be a confusion matrix combined with other categorization model performance measures. This table probably shows how the model did on a training set of data.

The model's accuracy, which is quite good at 0.9996, means that around 99.96% of its predictions were accurate.

The accuracy's 95 percent confidence interval is given as (0.9985, 1), indicating that the model's actual accuracy is probably going to fall within this range.

The accuracy that would be attained by a straightforward model that consistently predicted the most prevalent class is provided as 0.9835, which is referred to as the no information rate.

The p-value associated with the accuracy being greater than no information rate extremely small < 0.0001, indicating a highly significant difference between the model's accuracy and the no information rate.

The model's predictions and the actual results show a high degree of agreement, as indicated by the kappa coefficient of 0.9874. The p-value for Mcnemar's test is 0.4795, which suggests no significant difference in performance compared to another model.

All positive cases were properly predicted as positive, as shown by the sensitivity (true positive rate) of 1. The model appears to be effective at correctly predicting negative situations, as indicated by its specificity (true negative rate), which is 0.9756.

The positive predictive value (precision) is 0.9996, indicating that among the instances the model predicted as positive, nearly all were actually positive.

The negative predictive value is 1.0000, implying that all instances predicted as negative were truly negative.

The prevalence of the positive class in the data set is 0.9835, which is also the accuracy and the no information rate.

The detection rate is the same as the prevalence, indicating that the model correctly detects all positive cases.

The detection prevalence is very close to the prevalence, suggesting that almost all actual positive cases are detected by the model.

The balanced accuracy is 0.9878, calculated as the average of sensitivity and specificity. This value indicates a high level of performance in distinguishing between the two classes.

In conclusion, the model performs remarkably well on this training dataset, exhibiting high accuracy, sensitivity, specificity, and precision. The good agreement and performance in identifying both positive and negative situations is also indicated by the high kappa value and balanced accuracy.

Table 4.16: Confusion Matrix and Statistics (Test)

Reference	Prediction	0	1
	0	1282	20
	1	1	0
Accuracy		0.9839	
95% CI		(0.9755, 0.99)	
No Information Rate		0.9847	
P-Value [Acc > NIR]		0.6444	
Kappa		-0.0015	
Mcnemar's Test P-Value		0.0001	
Sensitivity		0.9992	
Specificity		0.0000	
Pos Pred Value		0.9846	
Neg Pred Value		0.0000	
Prevalence		0.9847	
Detection Rate		0.9839	
Detection Prevalence		0.9992	
Balanced Accuracy		0.4996	

Approximately 98.39% of occurrences are accurately classified by the model. This suggests that many predictions turn out to be accurate.

The accuracy's 95% confidence interval lies between 97.55 and 99.0 percent. The range of values within which the genuine accuracy is expected to fall is provided by this interval.

The probability of correctly guessing the majority class for all cases (in this example, 0) is approximately 98.47%. The accuracy of the model is marginally higher than this baseline.

The p-value is 0.6444 when comparing the model's accuracy to the no information rate. The accuracy of the model is not statistically different from the no information rate when the p-value is larger.

The agreement between actual and anticipated classifications is gauged by the Kappa coefficient, which is roughly -0.0015. A negative value denotes a lack of agreement between the model's predictions and actual results beyond what could be predicted by chance.

The Mcnemar's test, which uses statistics to compare paired proportions, has a p-value of 0.0001. This test can assist identify whether the errors in the model are considerably unequally distributed among classes. The model's sensitivity (true positive rate), which measures how well it can identify positive class (class 1) instances, is 99.92%; however, its specificity (true negative rate), which measures how well it can identify negative class (class 0) instances, is 0.00%.

The positive predictive value (precision) is 98.46%. This represents the proportion of predicted positive instances that are actually positive.

The negative predictive value is 0.00%, indicating that the model's ability to correctly predict negative instances is very poor.

The prevalence of the positive class is 98.47%, which is the proportion of actual instances belonging to the positive class.

The detection rate (true positive rate) is 98.39%, similar to sensitivity, showing the model's ability to detect positive instances.

The detection prevalence is 99.92%, indicating how often the model predicts positive instances.

The average of the sensitivity and specificity is used to compute the balanced accuracy, which is 49.96%. It suggests that the model's capacity to identify the positive class has a significant impact on its overall performance.

Overall, the model exhibits good sensitivity, but it has trouble with specificity and predicting classes that will be excluded. Limited agreement between expected and actual results is shown by the Kappa coefficient and balanced accuracy. The Mcnemar's test reveals unequal error distributions between classes.

4.4 Results and Discussions

In conclusion, this study has used a variety of statistical models, such as the Poisson (P), Negative Binomial (NB), Zero-Inflated Poisson (ZIP), Zero-Inflated Negative Binomial (ZINB), Hurdle Poisson (HurdleP), and Hurdle Negative Binomial (HurdleNB) models, to provide important insights into the relationship between covariates and count data. According to their coefficients in the various models, the covariates ”Rainfall” and ”Temperature” have had varying effects on the count data. The magnitudes and directions of these effects vary noticeably between the models, indicating that the model used may have an effect on how covariate effects are interpreted. The goodness-of-fit of the various models has been compared using the AIC values and log likelihoods.

In the context of this research, we explored the performance of various models in handling zero-inflated data, including the Zero-Inflated Negative Binomial (ZINB) and Zero-Inflated Poisson (ZIP) models. Notably, our analysis revealed that these models encountered difficulties in effectively capturing the intricate characteristics of the zero-inflated data, leading to suboptimal fits.

In contrast, when employing the Negative Binomial (NB), Negative Binomial Hurdle (NBH), Poisson (P), and Poisson Hurdle (HP) models, consistent and robust performance across diverse aspects of the data emerged. This observation underscores the critical importance of thoughtfully selecting an appropriate modeling approach when dealing with zero-inflated data. The success demonstrated by the NB, NBH, P, and HP models in this study suggests their potential as potent tools for accurately and reliably modeling zero-inflated data.

This research not only enhances the understanding of model selection within the realm of zero-inflated data analysis but also offers invaluable insights for both researchers and practitioners who engage with similar data types. It’s also important to note that sophisticated machine learning methods, such support vector machines and random forest models, have shown adept at capturing the subtleties of zero-inflated data.

It’s vital to remember that this study has some restrictions. The conclusions made here are dependent on the model assumptions, and additional study might look into different approaches or deal with potential causes of bias. The scope of this study is also limited by the data that are available and the chosen covariates. The results of this study highlight the need of taking into account different statistical models and their underlying premises in order to develop a more thorough knowledge of count data and its interactions with variables.

Adding the outcomes of machine learning models into your conclusion is an excellent approach to give a thorough summary of your thesis. The Logistic Model, which was used to analyse the correlations between variables and count data, was one of the important machine learning models used in this work. By clarifying the connections between covariates and outcomes in the context of logistic regression, these findings from the logistic model provide a dimension of interpretability to the analysis of count data. The significance of "Rainfall" and "Temperature" highlights the importance of these variables in affecting the count data, and their respective p-values give a measure of confidence in the impacts that were seen. A comprehensive grasp of the data, its properties, and the variables that cause its fluctuations is aided by these machine learning insights.

Confusion matrices were used to assess the effectiveness of the SVM model during both the pre-SVM training and testing phases. In both the training and testing phases, the SVM model showed high accuracy; however, the fact that it was unable to predict the negative class (specificity = 0), raises concerns about how well it will perform in classifying this specific result. These measurements emphasise how crucial it is to evaluate model efficiency in a balanced manner, taking into account both real positives and true negatives. A more comprehensive understanding of the underlying patterns in the data and the influence of covariates can be obtained by integrating SVM with statistical models.

This study used a Random Forest Model in addition to the Logistic Model and Support Vector Machine to analyse the associations between count data in greater detail. The Random Forest Model demonstrated amazing accuracy throughout the training phase and showed the capacity to accurately capture both positive and negative outcomes with high sensitivity and specificity. However, the testing phase poses particular difficulties, with a reduced balanced accuracy as a result of the model's ineffective prediction of the negative class. The results of the Random

Forest Model highlight the need to take model robustness and generalisation into account when using machine learning methods. Although the model performs well during training, its performance on untried data emphasises the difficulties of real-world application.

Chapter 5

Recommendations

In this thesis, the models that we built, include the predictors regardless whether they are significant or not. In the future, as some zones are completely free of malaria incidence we may analyse the data and evaluate how malaria mortality of each zone are affected by rainfall and temperature.

1. For variable selection in ZIP and Poisson models, we may also employ lasso (least absolute shrinkage and selection operator)-based methods.In addition to identifying all statistically significant predictors of the response variable, the variable selection processes also allow us to rule out any potential collinearity issues between the predictors.

2. To see if we can forecast future malaria fatalities due to malaria or not, we may also utilise zero-inflated poisson auto-regressive (ZIPA) or zero-inflated negative binomial auto-regressive (ZINBA) for time series analysis.

3. A nonlinear relationship between variables and count data would be worth investigating. Some variables could have complex connections that conventional models find difficult to account for. It could be interesting to investigate nonlinear models like generalised additive models (GAMs).

4. Include regional and temporal aspects in your analysis that could have an impact on malaria occurrences. Time-series models and spatial statistics may

be able to provide light on patterns and trends.

5. You can continue to deepen your comprehension of the connections between variables and malaria occurrences by heeding these advice and looking into new research directions, and you can make significant contributions to the field

References

Abiodun, G. J., Makinde, O. S., Adeola, A. M., Njabo, K. Y., Witbooi, P. J., Djidjou-Demasse, R., and Botai, J. O. (2019). A dynamical and zero-inflated negative binomial regression modelling of malaria incidence in limpopo province, south africa. *International Journal of Environmental Research and Public Health*, 16.

Abiodun, G. J., Njabo, K. Y., Witbooi, P. J., Adeola, A. M., Fuller, T. L., Okosun, K. O., Makinde, O. S., and Botai, J. O. (2018). Exploring the influence of daily climate variables on malaria transmission and abundance of anopheles arabiensis over nkomazi local municipality, mpumalanga province, south africa. *Journal of Environmental and Public Health*, 2018.

Akpan, G. E., Adepoju, K., and Oladosu, O. R. (2019). Potential distribution of dominant malaria vector species in tropical region under climate change scenarios. *PLoS ONE*, 14.

Biau, G. (2010). Analysis of a random forests model. *J. Mach. Learn. Res.*, 13:1063–1095.

Buri, M. and Hothorn, T. (2020). Model-based random forests for ordinal regression. *The International Journal of Biostatistics*, 16.

Caminade, C., Kovats, S., Rocklov, J., Tompkins, A. M., Morse, A. P., Colón-González, F. J., Stenlund, H., Martens, P., and Lloyd, S. J. (2014). Impact of climate change on global malaria distribution. *Proceedings of the National Academy of Sciences*, 111:3286 – 3291.

Chuang, T.-W., Soble, A., Ntshalintshali, N. E., Mkhonta, N., Seyama, E., Mthethwa, S., Pindolia, D. K., and Kunene, S. (2017). Assessment of climate-driven variations in malaria incidence in swaziland: toward malaria elimination. *Malaria Journal*, 16.

Craig, M. H., Snow, R. W., and le Sueur, D. (1999). A climate-based distribution model of malaria transmission in sub-saharan africa. *Parasitology today*, 15 3:105–11.

Dlamini, N., Zulu, Z., Kunene, S., Geoffroy, E., Ntshalintshali, N. E., Owiti, P., Sikhondze, W., Makadzange, K., and Zachariah, R. (2018). From diagnosis to case investigation for malaria elimination in swaziland: is reporting and response timely? *Public health action*, 8 Suppl 1:S8–S12.

Feng, C. X. (2021). A comparison of zero-inflated and hurdle models for modeling zero-inflated count data. *Journal of Statistical Distributions and Applications*, 8.

Greene, W. H. (1994). Accounting for excess zeros and sample selection in poisson and negative binomial regression models. *ERN: Discrete Regression & Qualitative Choice Models (Single) (Topic)*.

Hay, S. I., Rogers, D. J., Randolph, S. E., Stern, D. I., Cox, J. S. H., Shanks, G. D., and Snow, R. W. (2002). Hot topic or hot air? climate change and malaria resurgence in east african highlands. *Trends in parasitology*, 18 12:530–4.

Hu, M.-C., Pavlicova, M., and Nunes, E. V. (2011). Zero-inflated and hurdle models of count data with extra zeros: Examples from an hiv-risk reduction intervention trial. *The American Journal of Drug and Alcohol Abuse*, 37:367 – 375.

Jury, M. and Kanemba, A. (2007). A climate-based model for malaria prediction in southeastern africa. *South African journal of science*, 103(1-2):57–62.

Khan, A., Ullah, S., and Nitz, J. (2011). Statistical modelling of falls count data with excess zeros. *Injury Prevention*, 17:266 – 270.

Kim, D.-W., Deo, R. C., Park, S.-J., Lee, J.-S., and Lee, W.-S. (2018). Weekly heat wave death prediction model using zero-inflated regression approach. *Theoretical and Applied Climatology*, 137:823–838.

Lambert, D. (1992). Zero-inflacted poisson regression, with an application to defects in manufacturing. *Quality Engineering*, 37:563–564.

le Cessie, S. and van Houwelingen, J. C. (1994). Logistic regression for correlated binary data. *Applied statistics*, 43:95–108.

Liaqat, M., Kamal, S., Fischer, F., and Zia, N. (2020). Influence of covariates on involved lymph nodes in primary breast cancer patients: Mixture distribution zero-inflated modeling methodological framework.

Loevinsohn, M. E. (1994). Climatic warming and increased malaria incidence in rwanda. *The Lancet*, 343:714–718.

Makinde, O. S., Abiodun, G. J., and Ojo, O. T. (2020). Modelling of malaria incidence in akure, nigeria: negative binomial approach. *GeoJournal*, pages 1–10.

Mbokazi, F., Coetzee, M., Brooke, B. D., Govere, J. N., Reid, A. J., Owiti, P., Kosgei, R. J., Zhou, S., Magagula, R., Kok, G., Namboze, J. M., Tweya, H., and Mabuza, A. M. (2018). Changing distribution and abundance of the malaria vector anopheles merus in mpumalanga province, south africa. *Public health action*, 8 Suppl 1:S39–S43.

Mihreteab, S., Lubinda, J., Zhao, B., Rodríguez-Morales, A. J., Karamehic-Muratovic, A., Goitom, A., Shad, M. Y., and Haque, U. (2020). Retrospective data analyses of social and environmental determinants of malaria control for elimination prospects in eritrea. *Parasites & Vectors*, 13.

Mopuri, R., Kakarla, S. G., Mutheneni, S. R., Kadiri, M. R., and Kumaraswamy, S. (2020). Climate based malaria forecasting system for andhra pradesh, india. *Journal of Parasitic Diseases*, 44:497 – 510.

Mullahy, J. (1986). Specification and testing of some modified count data models. *Journal of Econometrics*, 33:341–365.

Purnama, D. I. (2021). Comparison of zero inflated poisson (zip) regression, zero inflated negative binomial regression (zinb) and binomial negative hurdle regression (hnb) to model daily cigarette consumption data for adult population in indonesia.

Tan, A. K. G. and Yen, S. T. (2017). Cigarette consumption by individuals in malaysia: a zero-inflated ordered probability approach. *Journal of Public Health*, 25:87–94.

Thomson, M. C., Ukawuba, I., Hershey, C. L., Bennett, A., Ceccato, P., Lyon, B., and Dinku, T. (2017). Using rainfall and temperature data in the evaluation of national malaria control programs in africa. *The American Journal of Tropical Medicine and Hygiene*, 97:32 – 45.

Tüzen, M. F. and Erbaş, S. (2018). A comparison of count data models with an application to daily cigarette consumption of young persons. *Communications in Statistics - Theory and Methods*, 47:5825 – 5844.

Vuong, Q. H. (1989). Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica*, 57:307–333.

Wardrop, N. A., Barnett, A. G., Atkinson, J.-A., and Clements, A. C. A. (2013). Plasmodium vivax malaria incidence over time and its association with temperature and rainfall in four counties of yunnan province, china. *Malaria Journal*, 12:452 – 452.

Wei, C. K., Lu, N., Yang, R., Tang, Y., Lü, Q., and Jiang, J.-Y. (2020). [epidemic situation of malaria in yunnan province from 2014 to 2019]. *Zhongguo xue xi chong bing fang zhi za zhi = Chinese journal of schistosomiasis control*, 32 5:483–488.

Zeileis, A., Kleiber, C., and Jackman, S. D. (2008). Regression models for count data in r. *Journal of Statistical Software*, 27:1–25.

Zhang, Y., Ma, X., Zhang, Y., and Yang, J. (2013). Support vector machine of the coal mine machinery equipment fault diagnosis. *2013 IEEE International Conference on Information and Automation (ICIA)*, pages 1141–1146.