BANGLADESH UNIVERSITY OF ENNGINEERING AND TECHNOLOGY

DEPARTMENT OF INDUSTRIAL AND PRODUCTION ENGINEERING

Course: Data Science

Course Code: IPE 6107

Submitted to:

Assistant Professor Ridwan Al Aziz

Submitted by:

Ashiqur Rahman Khan - 0422082002

Zahidul Islam Sayeem - 0422082017

A Machine Learning Approach for Predicting Malaria Based on Nigeria Malaria Indicator Survey (MIS) of Demographic and Health Surveys 2021

A Machine Learning Approach for Predicting Malaria Based on Nigeria Malaria Indicator Survey (MIS) of Demographic and Health Surveys 2021

Abstract

Malaria is a deadly disease if it is not treated with immediate approach. Nigeria is the hotspot of Malaria in Africa with 31.3% of the total malaria death population. Detection of malaria is another tough task. Demographic and Health Surveys 2021 of Malaria report shows that *the Rapid Malaria test* may miss the chance to give accurate result. It also shows that a *Final result of malaria from blood smear test* gave malaria positive result but the *Result of malaria rapid test* gave negative result. In this case, it becomes hard to give proper treatment to the patient as enough time is lost due to *Final result of malaria from blood smear test* which consumes so much time for giving results. To overcome these limitations predictive analysis like Machine Learning (ML) algorithms can be implemented. In this study, ML algorithms Logistic Regression (LR), Decision Tree (DT) and H2O Automated Machine Learning (AutoML) framework are implemented on Demographic and Health Surveys 2021 of Malaria of Nigeria dataset. Generalized Linear Model (GLM) and Stacked Ensemble algorithms of H2O AutoML showed better result over LR and DT on original dataset and random oversampled dataset respectively. GLM and Stacked Ensemble showed 99.87910% and 99.9267% accuracy respectively. GLM also identified top four important features Presence of species: falciparum (Pf), Presence of species: malariae (Pm), Mother's highest educational level and Presence of species: ovale (Po) which plays a vital role in predicting malaria.

1. Introduction

1.1 Background

Malaria is a disease caused by Plasmodium parasites. Generally, it spreads to people by the bites of infected female Anopheles mosquitoes. Beside mosquito there are other 5 parasites that also cause malaria. *P. falciparum* and *P. vivax* are responsible for the greatest threat. On African continent the deadliest malaria parasite is *P. falciparum* and outside of the sun-Saharan Africa *P. vivax* is the dominant malaria parasite. The initial symptoms of malaria are fever, headache and chills. The symptoms appear after 10 to 15 days of infection by the bite of mosquito. It becomes

very difficult to recognize the disease as malaria and sometimes left untreated. When left untreated and if the malaria is caused by *P. falciparum*, then within 24 hours patient progress to severe illness to death. Half of the population was at risk in 2021 due to malaria. Infants, children under age of 5, pregnant women and patients with HIV/AIDS are considerably at high risk. People with low immunity moving to intense malaria transmission area also at high risk. According to the World malaria report there were 247 million cases in 2021 compared to 245 million cases in 2020 and the estimated death due to malaria 619000 in 2021 and 625000 in 2020. In 2021 the African region was the home of malaria with 95% of all malaria and 96% of death of all malaria cases and death population. Nigeria is leading the percentage of deaths due to malaria with 31.3% of the total malaria death population [1].

1.2 Motivation

Though the test result of having malaria is provided within 2 to 15 minutes, but is the result is negative and still having symptoms of malaria then it needs to provide blood smears every 12 to 24 hours over a period of two to three days [2][3]. We also know that it's very hard to identify, the patient is affected with malaria or not. And if left untreated then the patient may lead to death within 24 hours [1]. The dataset used in this study also shows that in a "Result of malaria rapid test" about 267 tests shows negative result but later on it was found that all those results became positive in a "Final result of malaria from blood smear test" [4]. So, if the test fails to identify, then the patient is at high risk to death. The medical test may fail to identify in initial stage, and may take enough time to figure out the actual result, what if the patient leads to no return stage within that time.

There are millions of cases of malaria and are being recorded every year in a structed way. By analyzing those data predictive models as Machine Learning (ML), Deep Learning (DL), etc. models can be used to predict the probability of a patients having malaria without medical test and preliminary treatments can be started based on the result of those predictive models. Medical tests will have to be done parallelly as predictive models are only for initial stage only.

1.3 Summary: Expected results and insights

Machine Learning, Deep Learning and Artificial Intelligence etc. are the leading technology in this era. There are many well-known efficient algorithms exists nowadays for predicting with high accuracy and precision.

For predicting malaria (positive or negative) based on household data can be a new approach beside biological parameters. Models like Logistic Regression (LR), Decision Tree (DT), Automated Machine Learning (AutoML) framework H2O can be implemented for predicting the outcome of having malaria. Best model can be identified by comparing the performance measures of those models.

AutoML H2O may perform better than other models. Performance of Traditional ML models depends on the expertise of the users. But AutoML is an off-the-shelf ML method [5] it is a self-learning method which can provide best performing model after having trial of as many models as user wants.

2. Literature Review

2.1 Current Knowledge

Decision making technologies like Machine Learning, Artificial Intelligence, Deep learning etc. have opened a new era for health care sector. Nowadays health care experts may use these types of decision-making technologies for analyzing the patient's history with random multiple parameters. This may lead the experts to give decisions within a short time and in health care sector fraction of second is very vital for patients. These technologies in health care will reduce the human error and increase the efficiency to predict the disease. In the case of malaria, it is hard to predict the disease until it shows the symptoms. Even though sometimes it is hard to get the correct result by rapid test. Researchers tried to predict malaria positive or negative from various types of factors like clinical factors, RNA sequencing, image processing from blood sample, biological factors etc. Some recent researches on malaria using predictive models are going to be mentioned in Section 2.2 literature review table.

2.2 Literature Review Table

Table.1: Literature Review Table

SL No.	Title	Author	Method	Predictors	Outcome	Data Source
01	Predicting malarial outbreak using Machine Learning and Deep Learning approach: A review and analysis [6]	Godson Kalipe et al Year: 2018	1. K Nearest Neighbors (KNN) 2. Random Forest (RF) 3. Support Vector Machine (SVM) 4. Extreme Gradient Boosting (XGBoost) 5. Logistic Regression (LR) 6. Artificial Neural Network (ANN) 7. Naïve Bayes (NB) 9. Accuracy 10. Precision 11. Recall 12. Error Rate 13. Matthews Correlation Coefficient (MCC) 14. Specificity 15. False Positive Rate (FPR)	1. Minimum temperature 2. Maximum temperature 3. Humidity level 4. Ratio of the number of malaria cases by the population	Outbreak (Predict malaria)	1. National Vector Borne Disease Control Program 2. Indian meteorological Centre and Cyclone Warning Center
02	Machine learning model for	You Won Lee et al.	1. Synthetic Minority Oversampling Technique (SMOTE)	1. Gender 2. Age 3. Nationality	Predict malaria	1.Center for Disease Control and

	predicting malaria using clinical information [7]	Year: 2020	2. SVM 3. RF 4. Multilayered Perceptron (MLP) 5. AdaBoost (Ada) 6. Gradient boosting (GB) 7. CatBoost (CB) 8. Accuracy 9. F1-Score 10. Precision 11. Recall 12. Cross-validation (CV) 13. Feature Importance by RF 14. Area under curve (AUC)	4. Symptomic body region 5. Symptom		Prevention (CDC) 2. PubMed
03	Prediction of malaria incidence using climate variability and machine learning [8]	Odu Nkiruka et al. Year: 2021	1. K-means clustering 2. XGBoost 3. SVM 4. NB 5. LR 6. Accuracy 7. AUC 8. CV 9. Variance Inflation factor (VIF) 10. Receiver Operating Characteristic Curve (ROC)	1. Precipitation 2. Surface Radiation, 3. Temperature 4. Atmospheric pressure 5. Relative Humidity	Case of Malaria Increase or decrease	1. WHO data repository 2. National Centre for Atmospheric Research (NCAR)

04	Machine Learning based Malaria Prediction using Clinical Findings [9]	Samir S. Yadav et al. Year: 2021	1. NB 2. LR 3. Decision Tree (DT) 4. RF 5. SVM (kernel=gaussian) 6. SVM (kernel=polynom) 7. ANN(MLP) 8. Accuracy 9. Precision 10. Recall 11. F1-Score 12. AUC 13. ROC	1. Address 2. Days 3. Sex 4. Death 5. Diagnostic 6. Hospitalization 7. Month 8. Number of patients 9. Number of visit days 10. Observation 11. Rapid Diagnosis Test 12. Reverence 13. Signs – symptoms 14. Treatment 15. Weeks 16. Years	Malaria or Not Malaria	1. Distinct Places in Senegal, collected in 2016 during the “Grand Magal of Touba” 2. Medical records of the districts Diourbel, Thies and Fatick
05	ICA Learning Approach for Predicting of RNA-Seq Malaria Vector Data Classification Using SVM Kernel	Micheal Olaolu Arowolo Year: 2022	1. Independent component analysis (ICA) 2. SVM-Gaussian kernel 3. SVM-Polynomial 4. SVM-Linear kernel 5. SVM-RBF Kernel 6. Confusion matrix 7. Accuracy 8. Precision	1. Test ID 2. Gene ID 3. Gene 4. Locus 5. Sample_1 6. Sample_2	Status (Ok, not ok)	Western Kenya Mosquito Gene Dataset – Mosquito Anopheles Gambiae

	Algorithms [10]		9. Recall 10. Sensitivity 11. Specificity 12. F-Score			
06	Malaria Detection using Deep Learning [11]	Gautham Shekar et al. Year: 2020	1. Basic Convolutional Neural Network 2. Frozen Convolutional Neural Network 3. Fine-Tuned Convolutional Neural Network 4. Accuracy 5. Classification Error 6. Sensitivity 7. Precision 8. F1-Score 9. F_beta 10. Specificity, 11. FPR 12. MCC	Public database containing 27,558 images, Uninfected 13779 Infected 13779	Infected or uninfected cell	1. National Institute of Health (https://ceb.nlm.nih.gov/repositories/malaria-datasets/) 2. https://www.kaggle.com/datasets/iarunava/cell-images-for-detecting-malaria
07	Leveraging Deep Learning Techniques for Malaria Parasite Detection Using Mobile Application [12]	Mehedi Masud et al. Year: 2020	1. CNN 2. Cyclical Learning Rate (CLR) 3. Stochastic Gradient Descent (SGD) 4. Brad Kenstler's implementation 5. Automatic Learning Rate Finder 6. Accuracy 7. AUC	Public database containing 27,558 images, Uninfected 13779 & Infected 13779	Infected or uninfected cell	1. National Institute of Health (https://ceb.nlm.nih.gov/repositories/malaria-datasets/) 2. https://www.kaggle.com/d

			8. Precision 9. Recall (sensitivity) 10. F1-score 11. MCC			atasets/iaruna va/cell-images-for-detecting-malaria
08	Predicting malaria epidemics in Burkina Faso with machine learning [13]	David Harvey et al. Year: 2021	1. Time Series analysis 2. Poisson Distribution 3. Gaussian Distribution 4. Random Forest Regressor (RFR) 5. Feature Importance 6. One and two-tailed uncertainties 7. Monte Carlo Markov Chain 8. Accuracy 9. Precision 10. Recall (sensitivity)	1. The absolute number of consultations 2. Absolute number of tests required 3. The confirmed number of cases of malaria within a 30km gaussian smoothed region 4. Confirmed number of malaria cases within 100km 5. Rain-fall 6. Surface water	The absolute number of confirmed cases of malaria	Integrated e-Diagnostic Approach & Burkina Faso government classified dataset
09	A Deep Learning Model for Malaria Disease Detection and	Mahendra Kumar Gourisaria et al.	1. Image Augmentation 2. Deep Convolutional Neural Network (DCNN) 3. Accuracy 4. AUC	Public database containing 27,558 images, Uninfected 13779 & Infected 13779	Infected or uninfected cell	1. National Institute of Health (https://ceb.nlm.nih.gov/repositories/ma

	Analysis using Deep Convolutional Neural Networks [14]	Year: 2020	5. Precision 6. Recall (sensitivity) 7. F1-score			laria-datasets/) 2. https://www.kaggle.com/datasets/iarunava/cell-images-for-detecting-malaria
10	Determining suitable machine learning classifier technique for prediction of malaria incidents attributed to climate of Odisha [15]	Pallavi Mohapatra et al. Year: 2021	1. MLP 2. J48 classifier model (C4.5 decision tree method) 3. Root Mean Square Error (RMSE) 4. accuracy 5. kappa 6. ROC	1. Rainfall 2. Relative humidity 3. Surface (2-meter height from ground) 4. Maximum temperature	Malaria incidents	The Directorate of Public Health, Odisha, Special Relief Commissioner, Odisha, ECMWF Reanalysis land data (ERA5-Land), ECMWF Reanalysis land Data (ERA5-Land)
11	A Symptom-Based Machine Learning Model for Malaria	Bilyaminu Muhammad et al.	1. Exploratory Data Analysis 2. CART Algorithm Decision Tree 3. Accuracy 4. Confusion matrix	1. Age 2. Sex 3. Fever days 4. High temperature 5. Headache	Malaria Status	Hospitals of Nigeria

	Diagnosis in Nigeria [16]	Year: 2021		6. Cough 7. Vomit 8. Weakness 9. Sweat 10. Loss of Appetite 11. Skin rash 12. Abdominal Pain 13. Constipation 14. Convulsion 15. Diarrhea 16. Nausea 17. Frequent-Urination 18. Muscle Pains		
12	Diagnosing malaria from some symptoms: a machine learning approach and public health implications [17]	Hilary I. Okagbue et al. Year: 2020	1. LR 2. DT 3. Neural Network (NN) 4. RF 5. KNN 6. AdaBoost 7. AUC 8. Precision 9. Sensitivity 10. F1-score 11. Accuracy 12. log-loss 13. Specificity	1. Age 2. Sex 3. Fever 4. Cold 5. Rigor 6. Fatigue 7. Headache 8. Bitter tongue 9. Vomiting 10. Diarrhea 11. Convulsion 12. Anemia 13. Jaundice	Severe malaria (Positive or negative)	Federal Polytechnic Ilaro Medical centre, Ilaro Ogun state, Nigeria (https://ars.els-cdn.com/content/image/1-s2.0-S2352340919313526-mmc1.csv)

				14. Cocacola urine 15. Hypoglycemia 16. Prostration 17. Hyperpyrexia.		
13	Malaria patients in Nigeria: Data exploration approach [18]	Nureni Olawale Adeboye et al. Year: 2020	1. Chi-square test of independence 2. Contingency table 3. LR 4. Omnibus Tests of Model Coefficients 5. Hosmer and Lemeshow test	1. Age 2. Sex 3. Fever 4. Cold 5. Rigor 6. Fatigue 7. Headache 8. Bitter tongue 9. Vomiting 10. Diarrhea 11. Convulsion 12. Anemia 13. Jaundice 14. Cocacola urine 15. Hypoglycemia 16. Prostration 17. Hyperpyrexia.	Association with: severe malaria (Positive or negative)	Federal Polytechnic Ilaro Medical centre, Ilaro Ogun state, Nigeria (https://ars.els-cdn.com/content/image/1-s2.0-S2352340919313526mmc1.cs)
14	An ICA-ensemble learning approaches for	Micheal Olaolu Arowolo et al.	1. ICA 2. Bootstrap aggregating 3. Adaptive boosting	1. Test ID 2. Gene ID 3. Gene 4. Locus	Status	Western Kenya Mosquito Gene Dataset –

	prediction of RNAseq malaria vector gene expression data classification [19]	Year: 2021	4. Ensemble Subspace Discriminant Classification 5. Ensemble Bagged Tree Classification 6. Confusion matrix 7. Accuracy 8. Precision 9. Recall 10. Sensitivity 11. Specificity 12. F-Score	5. Sample_1 6. Sample_2		Mosquito Anopheles Gambiae
15	Comparative Study on the Prediction of Symptomatic and Climatic based Malaria Parasite Counts Using Machine Learning Models [20]	Opeyemi A. Abisoye Et al. Year: 2018	1. SVM 2. ANN 3. Confusion Matrix 4. Accuracy 5. Recall 6. Specificity 7. FPR 8. False Negative Rate 9. Mean Square Error	1. Headache 2. Fever 3. Dizziness 4. Body pain 5. Vomiting 6. Temperature 7. Relative humidity 8. Rainfall	Malaria	1. hospitals patients laboratory experimental 2. NECOP weather station, FUT Minna
16	Malaria Epidemic Prediction Model by Using Twitter Data and Precipitation	Nduwayezu Maurice Et al. Year: 2019	1. Web crawled 2. Geo Coding 3. Natural Language Processing Toolkits 4. Random oversampling technique	Raw data: 1. Id 2. Tweet text 3. Date 4. Location 5. Language 6. User	Related to malaria or not	Twitter

	Volume in Nigeria [21]		5. Bernoulli Naive Bayes test classifier 6. RF 7. SVM 8. XGBoost 7. Accuracy 8. Precision 9. Recall 12. F1-Score	After geocoding and preprocessing, for training: 1. Id 2. Tweet text		
17	Africa's Malaria Epidemic Predictor: Application of Machine Learning on Malaria Incidence and Climate Data [22]	Muthoni Masinde Et al. Year: 2020	1. Principal Component Analysis (PCA) 2. ANOVA 3. Friedman test 4. Kaiser-Meyer-Olkin Measure of Sampling Adequacy 5. Bartlett's Test of Sphericity 6. Chi-Square test 7. Kendall's coefficient of concordance 8. Decision Tree 9. Deep Learning 10. Fast Large Margin 11. Generalized Linear Model 12. Gradient Boosted Trees 13. Logistic Regression 14. Naive Bayes 15. Random Forest	1. Maternal mortality ratio 2. County name 3. Who region 4. Monthly rainfall 5. Monthly mean 6. Temperature 7. Altitude 8. Longitude	Malaria incidence	1. WHO (http://apps.who.int/gho/data/node.gswcah) 2. World Bank's climate knowledge portal (https://climateknowledgeportal.worldbank.org/download-data)

			16. SVM 17. Accuracy 18. AUC 19. Classification Error 20. F Measure 21. Precision 22. Recall 23. Sensitivity 24. Specificity 25. Training Time			
18	Data-driven malaria prevalence prediction in large densely populated urban holoendemic sub-Saharan West Africa [23]	Biobele J. Brown Et al. Year: 2020	1. Generalized Linear Models (GLM) 2. Ensemble Methods (EM) 3. Support Vector Machines (SVM) 4. Mean absolute error (MAE) 5. Mean square error (MSE) 6. Pearson Correlation coefficient (PCC) measures 7. L1–L2 ratio 8. Regularization strength elastic net parametrization 9. Cross-validation	1. Year 2. Month 3. Total number screened 4. Median age of malaria-negative 5. Median age of malaria-positive 6. IQR age malaria-negative 7. IQR age malaria-positive 8. Mean of blood parasite densities 9. STD of blood parasite densities	Malaria prevalence	Collected by: Department of Pediatrics of the College of Medicine of the University of Ibadan (COMUI), University College Hospital (UCH), Ibadan, Nigeria located in sub-Saharan West Africa

				10. Month total rainfall 11. Proportion of that year total rainfall 12. Month minimum temperature 13. Month maximum temperature 14. Month mean temperature		
19	Machine Learning Techniques for Malaria Incidence and Tuberculosis Prediction [24]	Odu Nkiruka Bridget et al. Year: 2021	1. Correlation Coefficient 2. VIF 3. k-means clustering 4. XGBoost 5. Accuracy 6. AUC 7. ROC 8. Sensitivity 9. Specificity 10. Cross-validation 11. Akaike Information Criterion (AIC) 12. Naïve Baye 13. SVM 14. LR 15. Precision 16. Recall	1. Precipitation 2. Surface radiation 3. Temperature 4. Atmospheric pressure 5. Relative humidity	Malaria incidence	1. WHO data repositor 2. National Centre for Atmospheric Research (NCAR) (https://ncar.ucar.edu/what-we-offer/data-services

			17. F1-Score			
20	Spatial Predictive Model for Malaria in Nigeria [25]	Adebayo Peter Idowu et al. Year: 2009	1. Artificial Neural Network 2. Demonstration of GMAL 1.0 software (a GIS based software) for Malaria Prediction	1. Id 2. Sex 3. Age 4. Month 5. Year 6. Longitudinal position 7. Latitudinal position 8. Address id 9. Location of study area	Spread of malaria	Primary Health care Centre Ife Central.

2.3 Research Gap

From the above literature review of recent researches, it is seen that there is not quite any research on malaria in Sub-Saharan African region of Nigeria on DHA MIS Dataset 2021, Nigeria [4]. Also, no research is done on demographic household data but clinical and climate data. ML and DL algorithms have been implemented by researchers but no AutoML approach is seen.

AutoML is opening a new path of predictive analysis. It is the automation of the entire process of using machine learning to solve real-world problems, from obtaining the raw data to creating a model ready for implementation. It is a solution that utilizes AI to address the increasing difficulty of implementing machine learning [8][9][10].

So, comparison of ML & DL models with AutoML models on DHS MIS Dataset 2021 of Nigeria has a very scope of research.

2.4 Research Objectives and Questions

2.4.1 Research Objectives

- To predict Malaria (Positive or Negative) using Machine Learning Models
- To identify best model for predicting, among Logistic Regression (LR), Decision Tree (DT), Random Forest (RF) and Automated Machine Learning (AutoML) framework H2O
- To identify the most important features that highly affect the outcome

2.4.2 Research Questions and Answers

- What is Automated Machine Learning and how it works?
    - Automated machine learning is the process of automating the tasks of applying machine
    - learning to real-world problems [5]
- How Automated Machine Learning is convenient over traditional Machine Learning?
    - Doesn't need Pre-process and clean the data.
    - Selects and constructs appropriate features. Selects an appropriate model family.
    - Optimizes model hyper parameters.
    - Designs the topology of neural networks (if deep learning is used).
    - Post processes machine learning models.
    - Critically analyzes the results obtained [5].
- Is it possible to predict the malaria using machine learning? Can we find the best ML model by this method?
    - One of the features of AUTO ML is that not only all the various types of models are formulated and visualized, but also the model performances and accuracies can be known. Thus, the best ML model can be found with highest accuracy [5]
- What makes this study unique to existing Researches?
    - Large Scale of patients' historical data, more than 40 predictors
    - Latest Data from DHS MIS Dataset 2021 of Nigeria
    - New ML model approach
    - Multilevel analysis to identify the factors associated with Malaria
- How will be the expected result and outcome of the research?
    - Result: Binary (Positive or Negative)

- Outcome: From this study we will be able to predict the Malaria in its initial stage without wasting time necessary steps can be taken to prevent the risk of severe condition to death.

3. Data Description

Data Source

The data used in this study is from a survey named Malaria Indicator Survey (MIS) of year 2021 of Nigeria. Malaria Indicator Survey (MIS) is developed by The Monitoring and Evaluation Working Group (MERG) of Roll Back Malaria, they coordinate global efforts to combat malaria. They conduct a stand-alone household survey called the Malaria Indicator Survey (MIS) to collect data from a representative sample of respondents at the national and regional/provincial levels. The DHS Program plays a major role in the development and implementation of the MIS, including co-chairing the MERG Survey and Indicator Guidance Task Force, contributing to the development of the MIS package (questionnaires, manuals, and guidelines based on Demographic and Health Surveys materials), and maintaining a website that provides information and data for Malaria Indicator Surveys worldwide. They also provide standardized malaria indicators for nearly 30 countries [26].

3.1 Data Description Table

The dataset consists of 70428 observations and 225 features. Where 43 features have been selected for predictive analysis are given below in Table.2

Table.2 Descriptions and data types of selected features [27]

Features	**Description**	**Data Type**	**Values**	Source
HV009	Number of household members	integer	0 to 90	DHS Malaria Indicator Survey
HV024	SubRegion	categorical	1.Sokoto, 2.Zamfara, 3.Katsina, 4.Jigawa, 5.Yobe, 6.Borno, 7.Adamawa, 8.Gombe, 9.Bauchi, 10.Kano, 11.Kaduna, 12.Kebbi,	

			13.Niger, 14.FCT, 15.Nasarawa, 16.Plateau, 17.Taraba, 18.Benue, 19.Kogi, 20.Kwara, 21.Oyo, 22.Osun, 23.Ekiti, 24.Ondo, 25.Edo, 26.Anambra, 27.Enugu, 28.Ebonyi, 29.Cross River, 30.Akwa Ibom, 31.Abia, 32.Imo, 33.Rivers, 34.Bayelsa, 35.Delta, 36.Lagos, 37.Ogun
HV025	Type of place of residence	categorical	1.Urban 2.Rural
HV045C	Native language of respondent	categorical	1.English 2.Hausa 3.Yoruba 4.Igbo 5.Fulfulde 6.Other
HV201	Source of drinking water	categorical	10.PIPED WATER 11.Piped into dwelling 12.Piped to yard/plot 13.Piped to neighbor 14.Public tap/standpipe, 20.TUBE WELL WATER, 21.Tube well or borehole 30. DUG WELL (OPEN/PROTECTED) 31.Protected well 32.Unprotected well, 40.SURFACE FROM SPRING, 41.Protected spring, 42.Unprotected spring, 43.River/dam/lake/ponds/stream/canal /irrigation channel, 51.Rainwater, 61.Tanker truck, 62.Cart with small tank, 71.Bottled water, 72.Sachet water, 96.Other
HV202	Source of non-drinking water	categorical	10.PIPED WATER, 11.Piped into dwelling, 12.Piped to yard/plot, 13.Piped to neighbor, 14.Public tap/standpipe, 20.TUBE WELL

			WATER, 21.Tube well or borehole, 30.DUG WELL (OPEN/PROTECTED), 31.Protected well, 32.Unprotected well, 40.SURFACE FROM SPRING, 41.Protected spring, 42.Unprotected spring, 43.River/dam/lake/ponds/stream/canal /irrigation channel, 51.Rainwater, 61.Tanker truck, 62.Cart with small tank, 96.Other
HV204	Time to get to water source (minutes)	integer	0 to 900, 996 (On premises), 998 (Don't know)
HV205	Type of toilet facility	categorical	10.FLUSH TOILET, 11.Flush to piped sewer system, 12.Flush to septic tank, 13.Flush to pit latrine, 14.Flush to somewhere else, 15.Flush, don't know where, 20.PIT TOILET LATRINE, 21.Ventilated Improved Pit latrine (VIP), 22.Pit latrine with slab, 23.Pit latrine without slab/open pit, 30.NO FACILITY, 31.No facility/bush/field, 41.Composting toilet, 42.Bucket toilet, 43.Hanging toilet/latrine, 96.Other
HV225	Share toilet with other households	categorical	0.No & 1.Yes
HV227	Has mosquito bed net for sleeping	categorical	0.No & 1.Yes
HV228	Children under 5 slept under mosquito bed net	categorical	0.No, 1.All children, 2.Some children, 3.No net in household
HV235	Location of source for water	categorical	1.In own dwelling, 2.In own yard/plot, 3.Elsewhere

HV238A	Location of toilet facility	categorical	1.In own dwelling, 2.In own yard/plot, 3.Elsewhere
HV244	Owns land usable for agriculture	categorical	0.No 1.Yes
HV246	Owns livestock, herds or farm animals	categorical	0.No 1.Yes
HV246G	Owns pigs	integer	0 (None), 1 to 94, 95 (95 or more), 98 (Unknown)
HV270	Wealth index combined	categorical	1.Poorest, 2.Poorer, 3.Middle, 4.Richer, 5.Richest
SHREGION	Region	categorical	1.North Central, 2.North East, 3.North West, 4.South East, 5.South South, 6.South West
HC1A	Child's age in days	integer	0 to 2500
HC27	Sex of the Child	categorical	1.Male 2.Female
HC53	Hemoglobin level (g/dl - 1 decimal)	float	10 to 990, 994 (Not present), 995 (Refused), 996 (Other)
HC57	Anemia level	categorical	1.Severe, 2.Moderate, 3.Mild, 4.Not anemic
HC61	Mother's highest educational level	categorical	0.No education, 1.Primary, 2.Secondary, 3.Higher, 8.Don't know
HML3	Net observed by interviewer	categorical	0.Not seen 1.Yes seen
HML10	Insecticide-Treated Net (ITN)	categorical	0.No 1.Yes
HML22	Obtained net from campaign, antenatal or immuni	categorical	0.No, 1.Yes mass distribution campaign, 2.Yes antenatal care, 3.Yes immunization visit
HML23	Place where net was obtained	categorical	10.Government health facility, 11.Government health facility, 20.Private health facility, 21.Private health facility, 30.Other sources, 31.Pharmacy, 32.Shop/market

			33.CHW, 34.Religious institution, 35.School, 96.Other, 98.Don't know
SH130	Reason net was not used	categorical	1.No mosquitoes, 2.No malaria, 3.Too hot, 4.Don't like smell, 5.Feel 'closed in', 6.Net too old/torn, 7.Net too dirty, 8.Net not available last night (washing), 9.Usual users did not sleep here last night, 10.Net not needed last night, 11.Bed bugs, 96.Other, 98.Don't know
HML32	Final result of malaria from blood smear test	categorical	0.Negative, 1.Positive, 6.Test undetermined, 7.Sample not found in lab database
HML32A	Presence of species: falciparum (Pf)	categorical	0.No 1.Yes
HML32B	Presence of species: malariae (Pm)	categorical	0.No 1.Yes
HML32C	Presence of species: ovale (Po)	categorical	0.No 1.Yes
HML32D	Presence of species: vivax (Pv)	categorical	0.No 1.Yes
HML35	Result of malaria rapid test	categorical	0.Negative, 1.Positive, 3.Not present, 4.Refused, 6.Other
HML37A	Suffer from illness/symptom: extreme weakness	categorical	0.No 1.Yes
HML37B	Suffer from illness/symptom: heart problems	categorical	0.No 1.Yes
HML37F	Suffer from illness/symptom: abnormal bleeding	categorical	0.No 1.Yes
HML37G	Suffer from illness/symptom: jaundice or yellow	categorical	0.No 1.Yes
HML37H	Suffer from illness/symptom: dark urine	categorical	0.No 1.Yes

HML37I	Suffer from illness/symptom: vomiting	categorical	0.No 1.Yes
HML37J	Suffer from illness/symptom: pallor	categorical	0.No 1.Yes
HML37K	Suffer from illness/symptom: refusal to eat	categorical	0.No 1.Yes
HML37L	Suffer from illness/symptom: very cold hands	categorical	0.No 1.Yes

3.2 Feature Selection

Feature Selection is very important to get a good performance from a ML model. This study is unique by its dataset and its verity features. This dataset is focused on household, social and economic information. Some features are selected from previous researches are given below

Table.3 Features Selected as per previous researches

SL No.	**Feature**	**Source (Related Literature)**
01	Gender	Machine learning model for predicting malaria using clinical information [7]
02	Age	
03	Location/ Region (Address)	Machine Learning based Malaria Prediction using Clinical Findings [9], Africa's Malaria Epidemic Predictor:
04	Rapid Diagnosis Test	

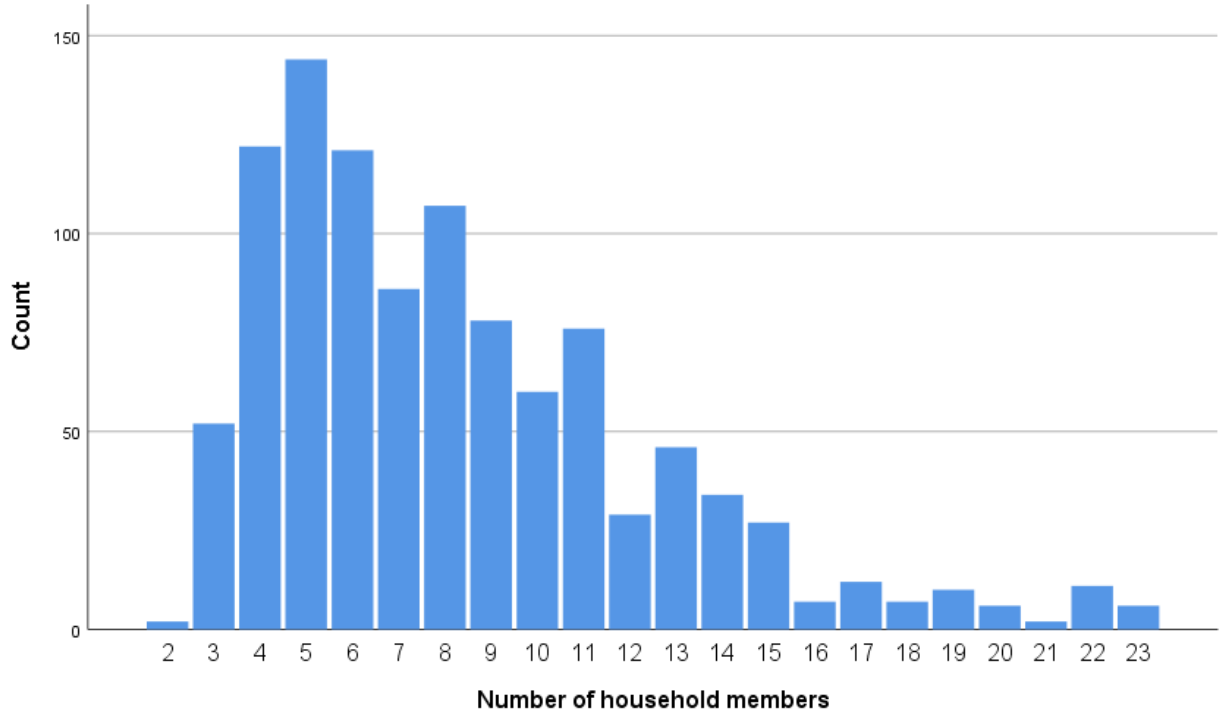
		Application of Machine Learning on Malaria Incidence and Climate Data [22]
05	Water	Predicting malaria epidemics in Burkina Faso with machine learning [13]
06	Language	Malaria Epidemic Prediction Model by Using Twitter Data and Precipitation Volume in Nigeria [21]

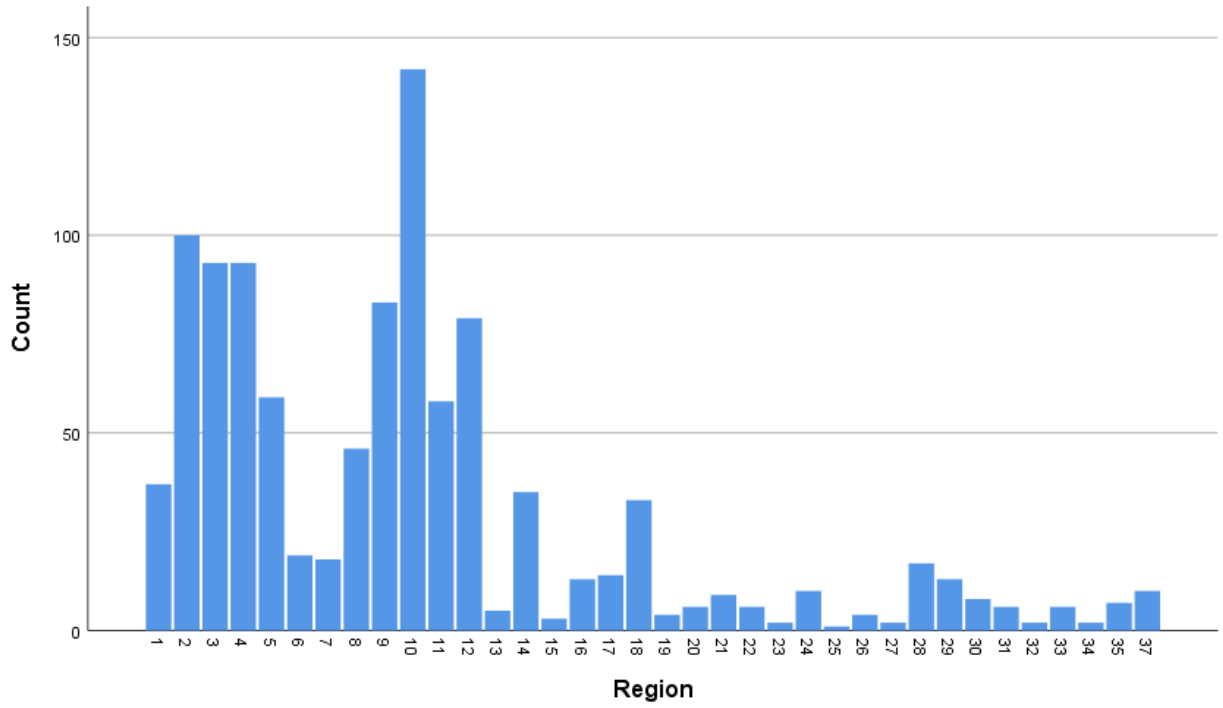
Except above feature of Table.3, rest of the features that are shown in Table.2 are selected as per data availability and considering socio-economic and household factors of patients. Features mentioned on Table.3 showed strong relation with the outcome feature in the mentioned researches.

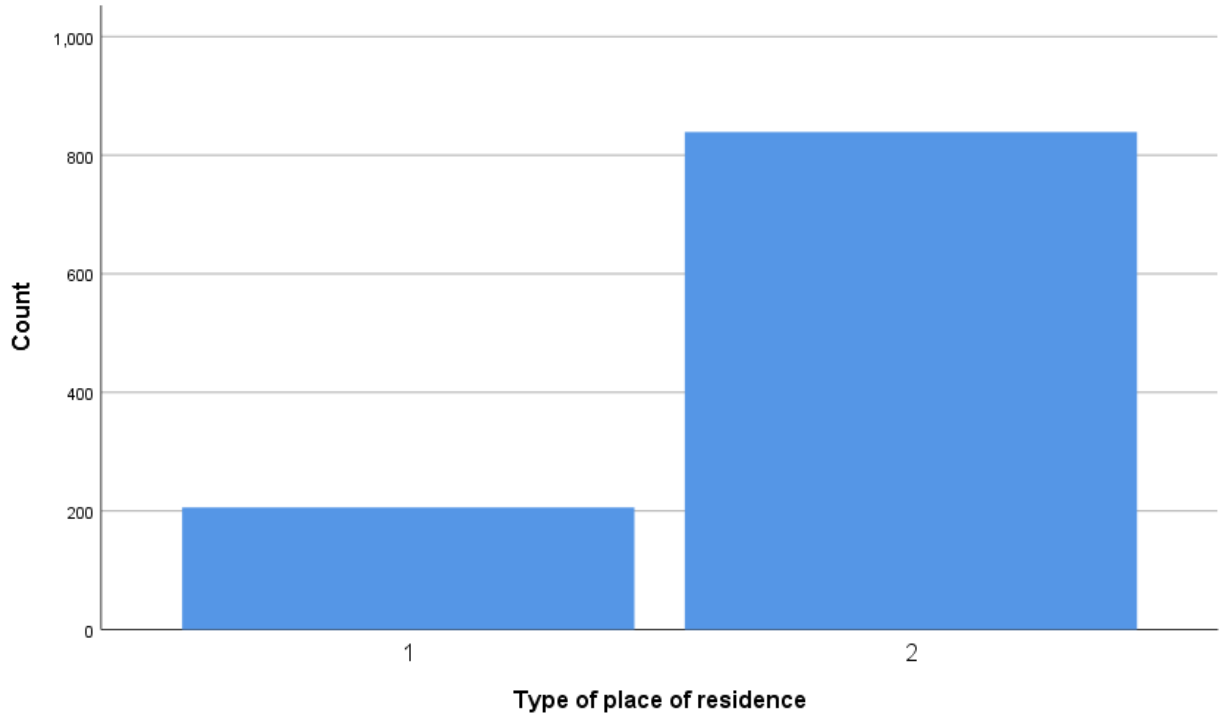
For further analysis all features mentioned on Table.2 are selected and three features “SH130 - Reason net was not used”, “HV202 - Source of non-drinking water” and “HML23 - Place where net was obtained” are eliminated due to data unavailability.

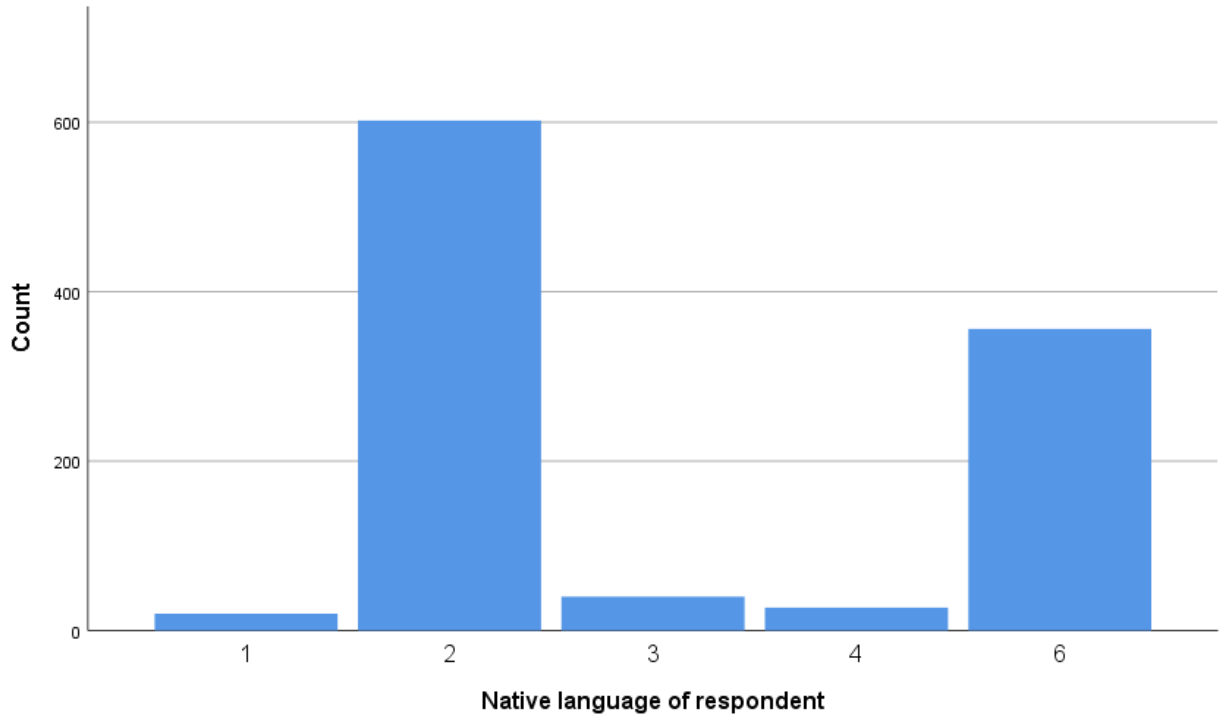
3.3 Data Distribution

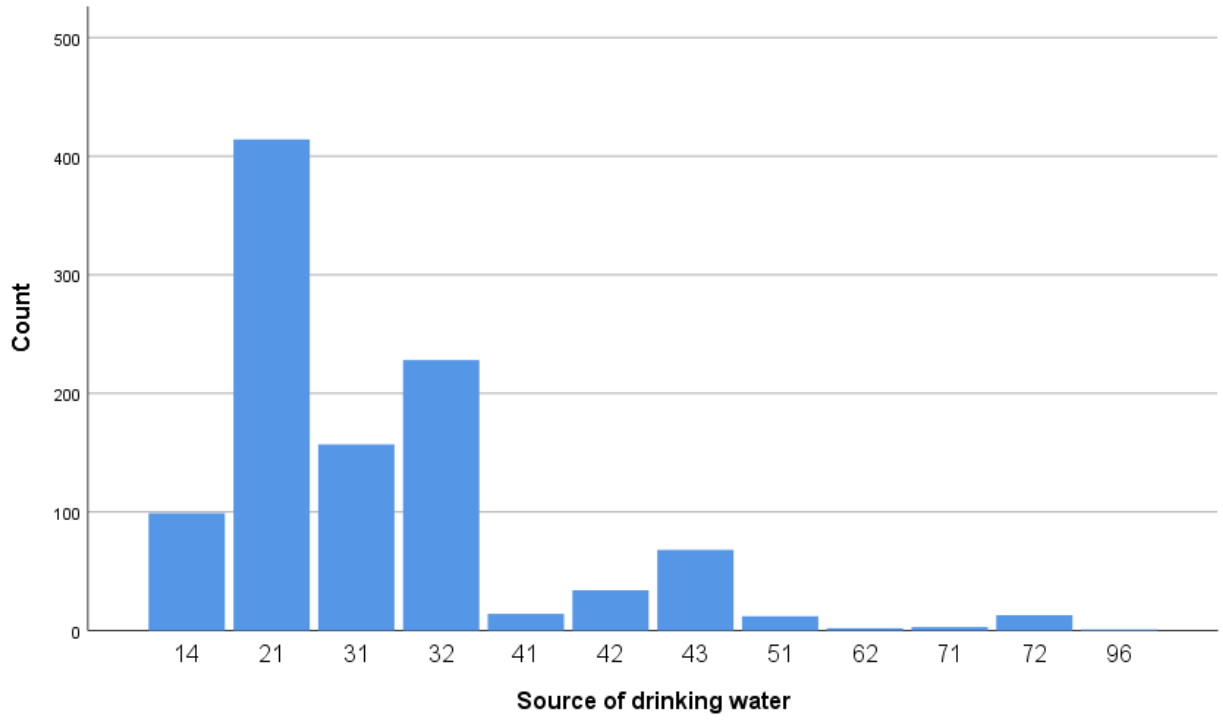
The numerical values represented in the figures refer to the column “Values” in the Table.2

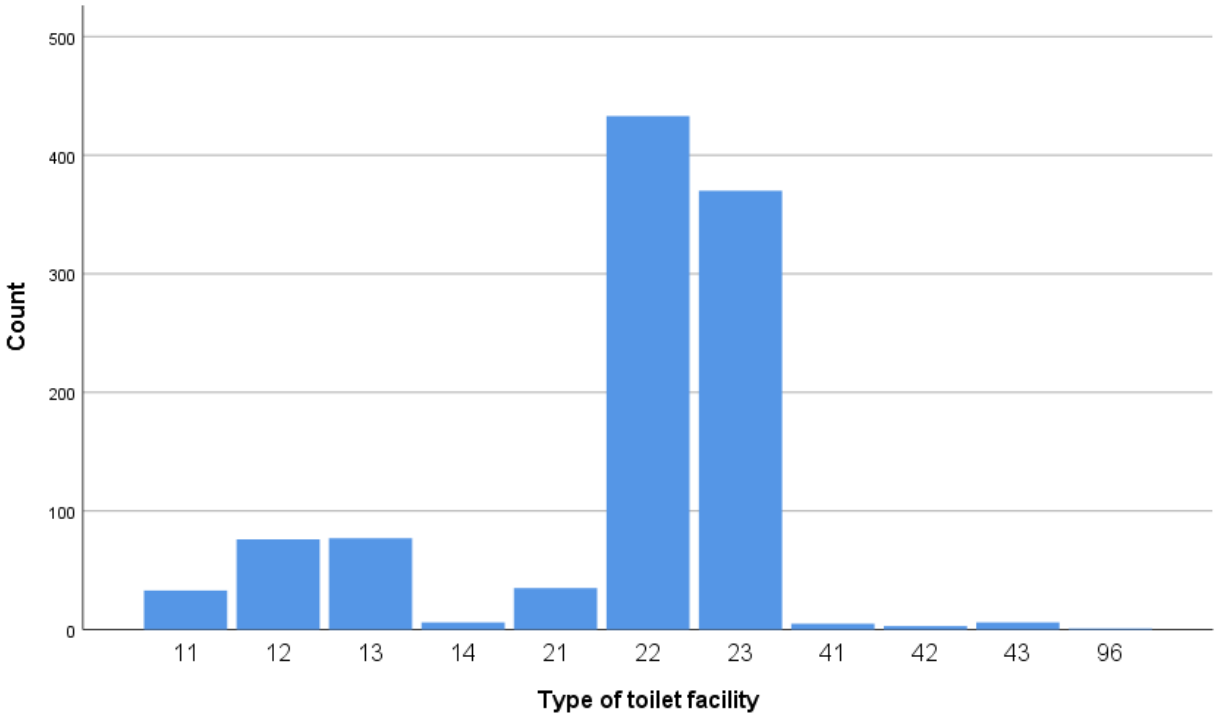


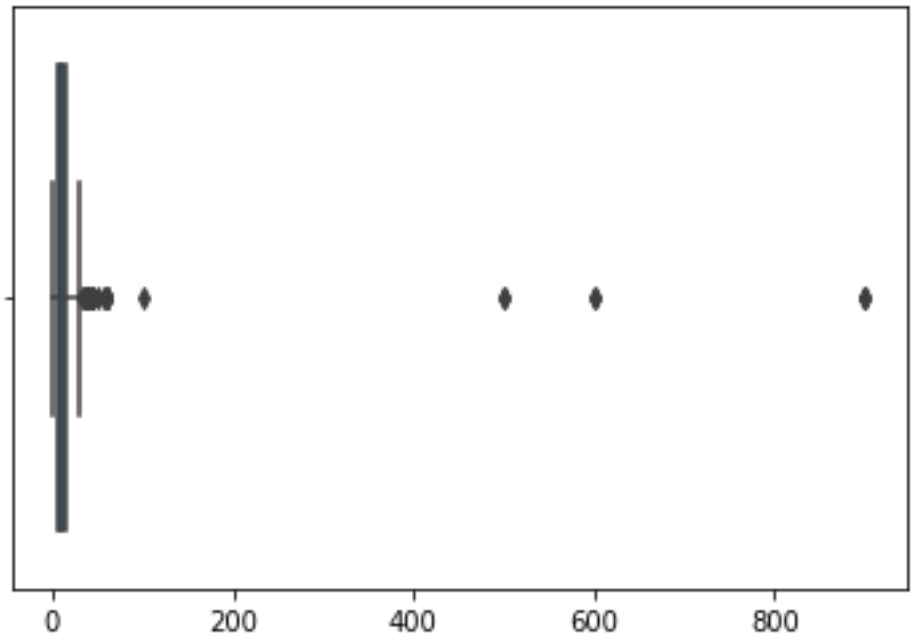

150
100
50
0
Count
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
37
Region

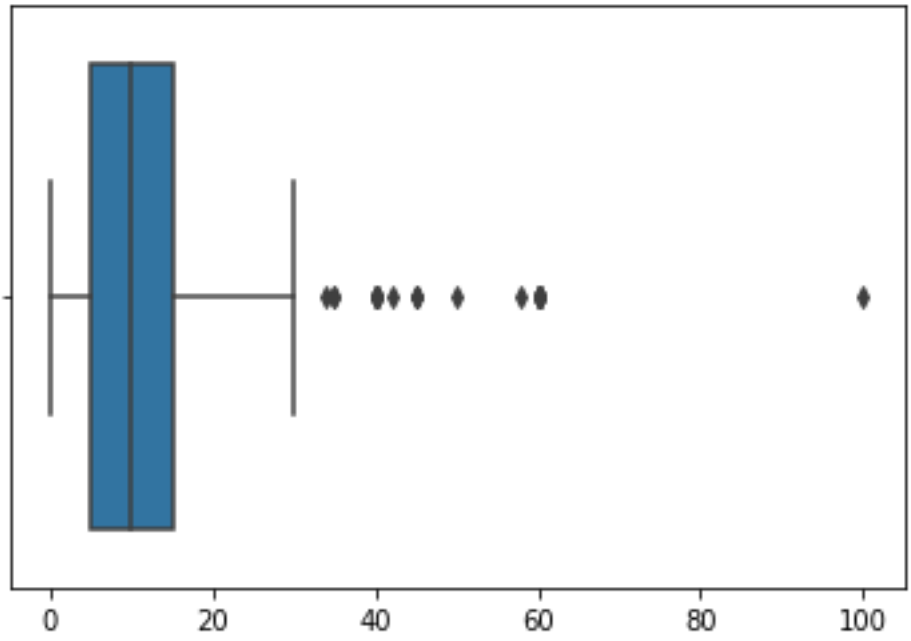

1,000
800
600
400
200
0
Count
1
2
Type of place of residence


Count
600
400
200
0
1
2
3
4
6
Native language of respondent

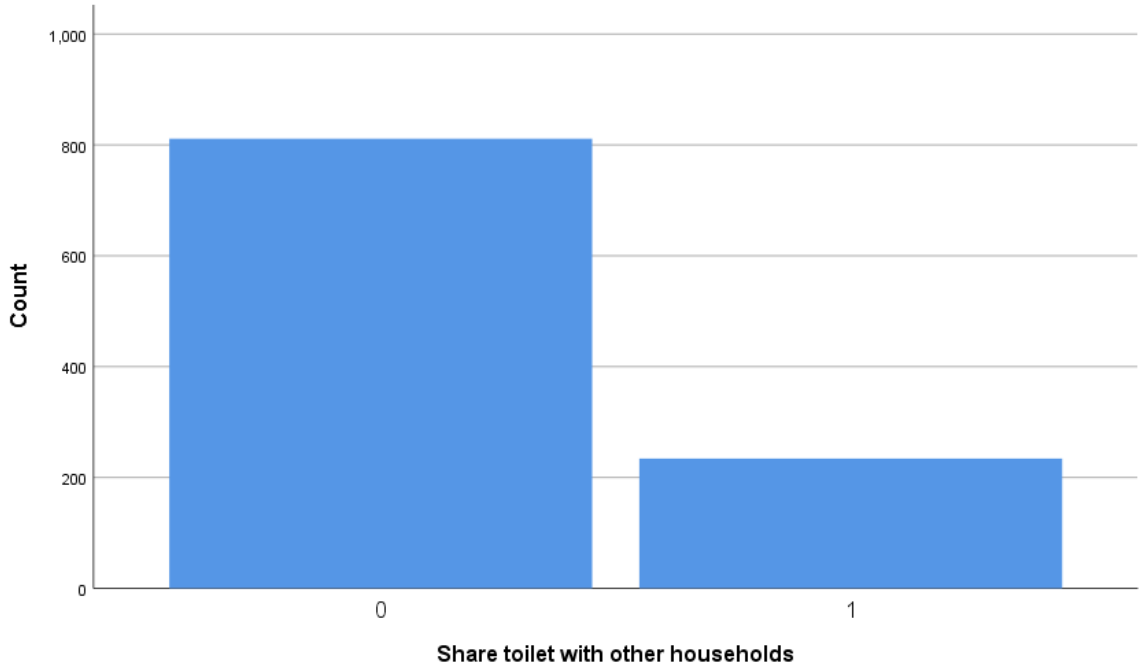

Count
500
400
300
200
100
0
14
21
31
32
41
42
43
51
62
71
72
96
Source of drinking water

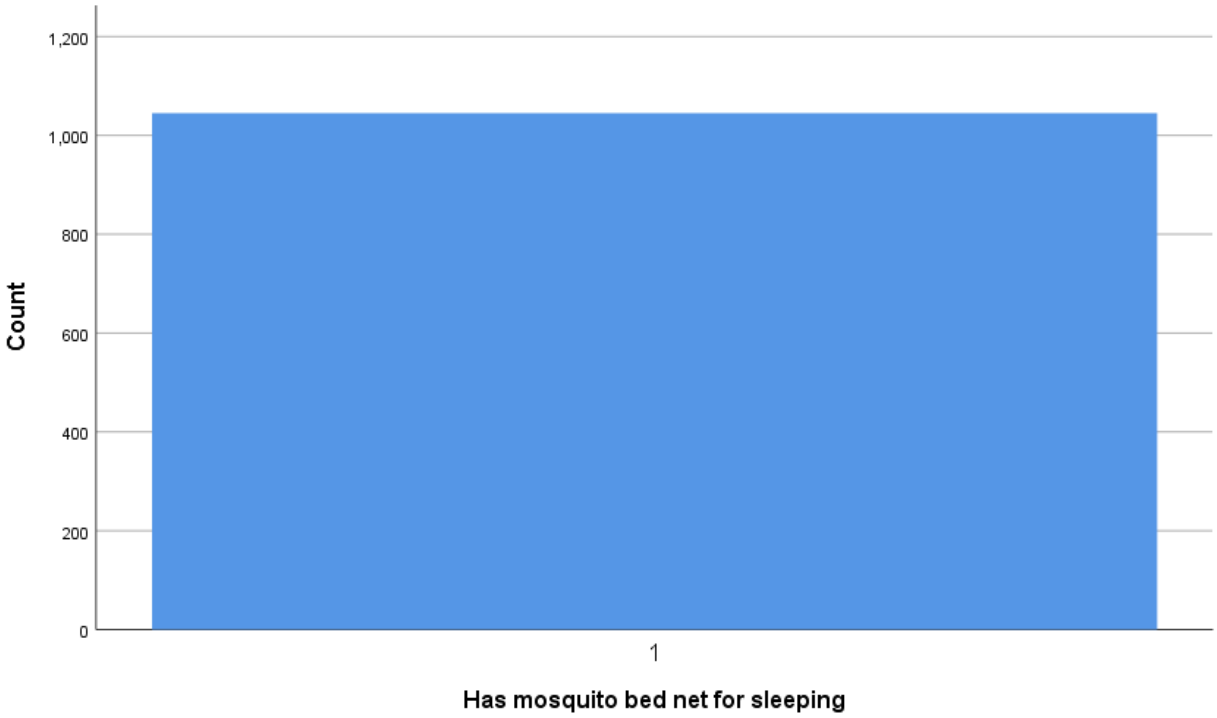

500
400
300
200
100
0
Count
11
12
13
14
21
22
23
41
42
43
96
Type of toilet facility

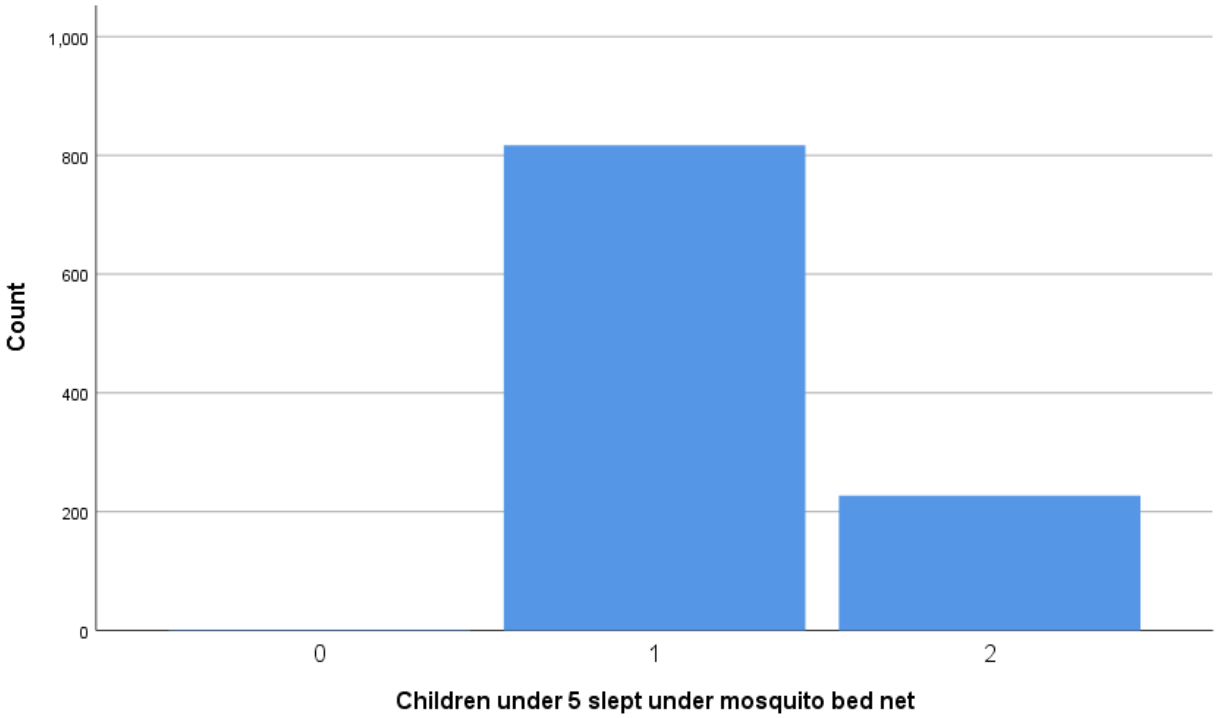

0
200
400
600
800

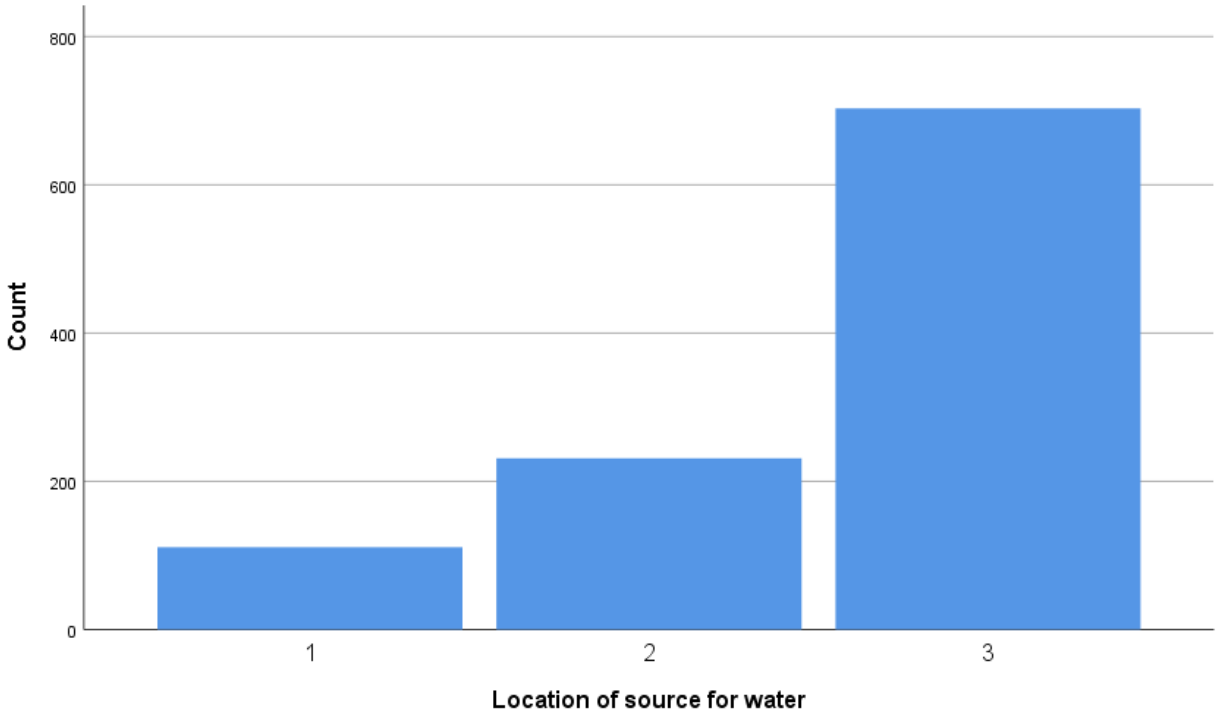


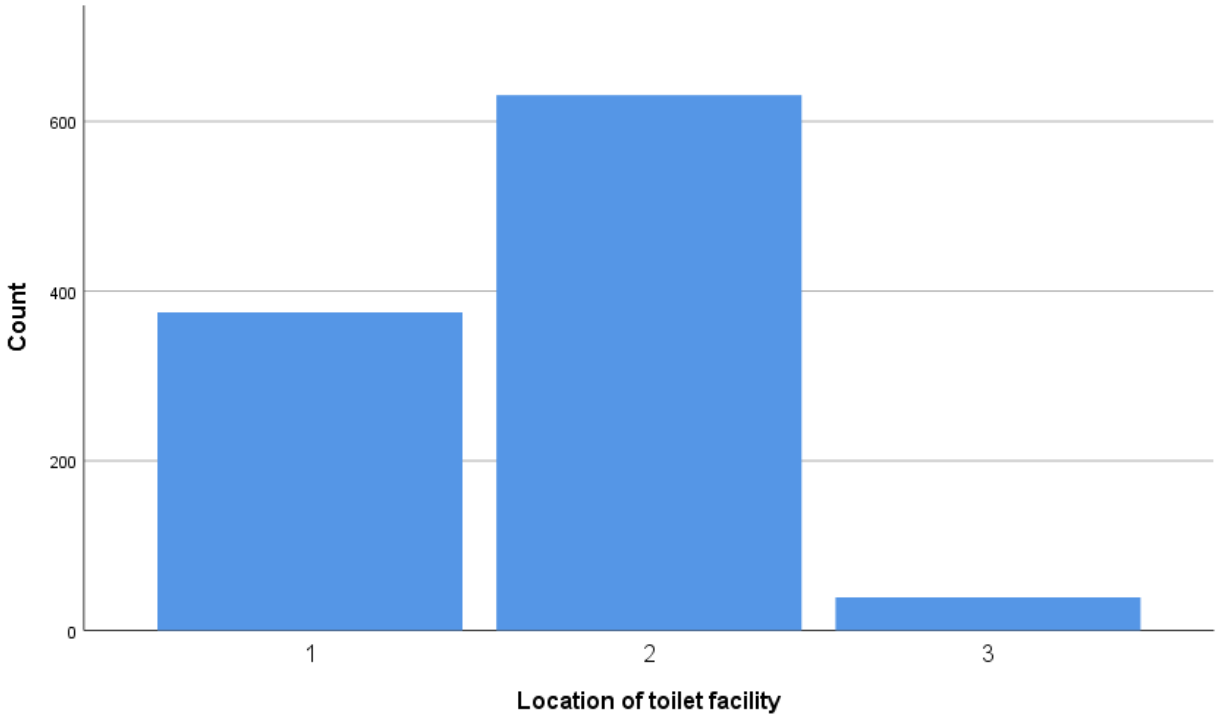
Time to get to water source (minutes)

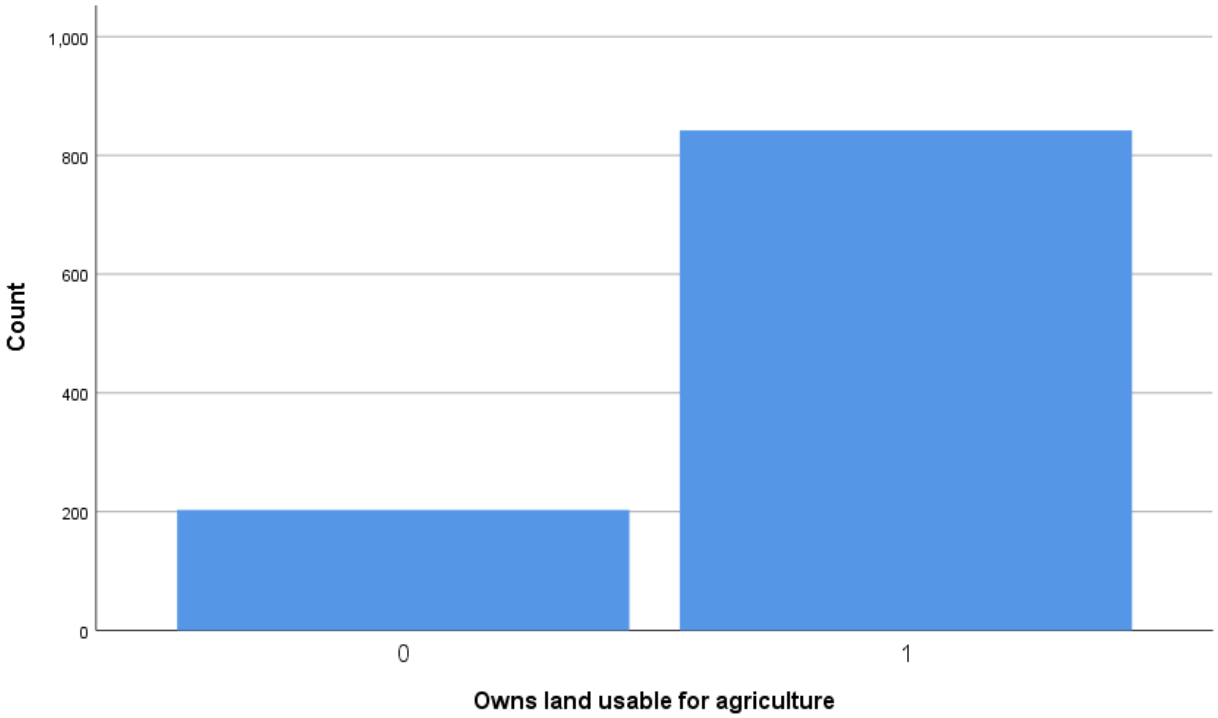


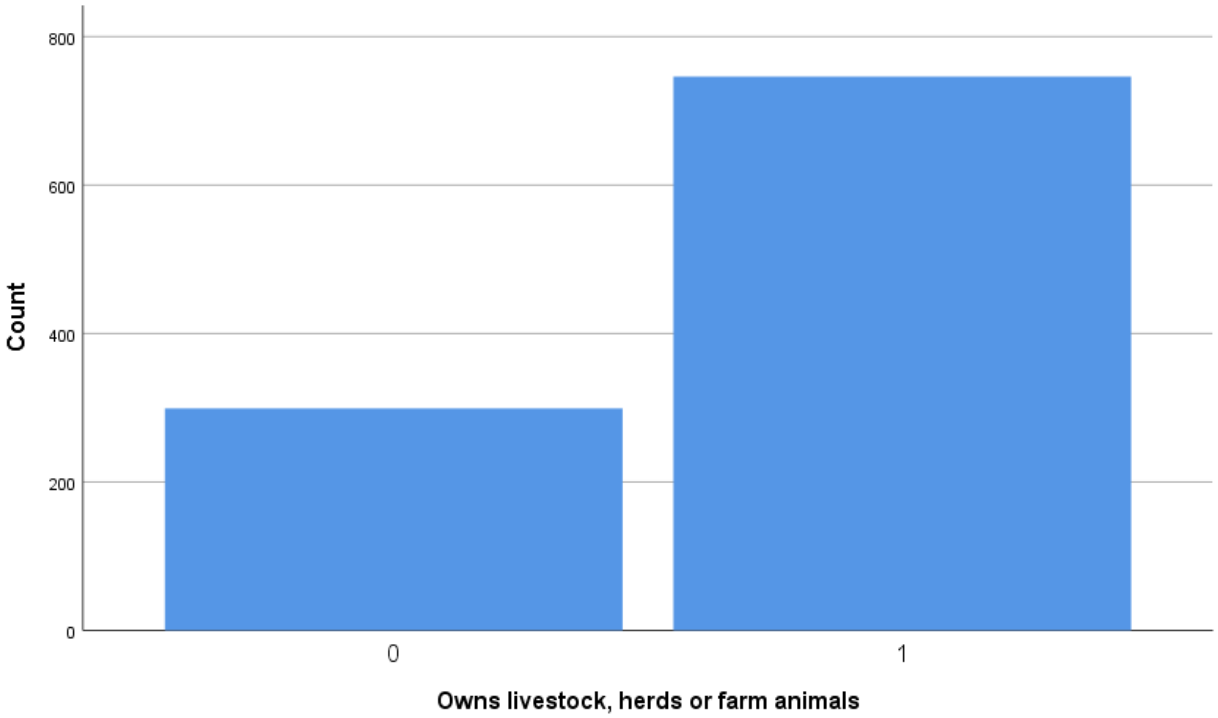

1,200
1,000
800
600
400
200
0
Count
1
Has mosquito bed net for sleeping

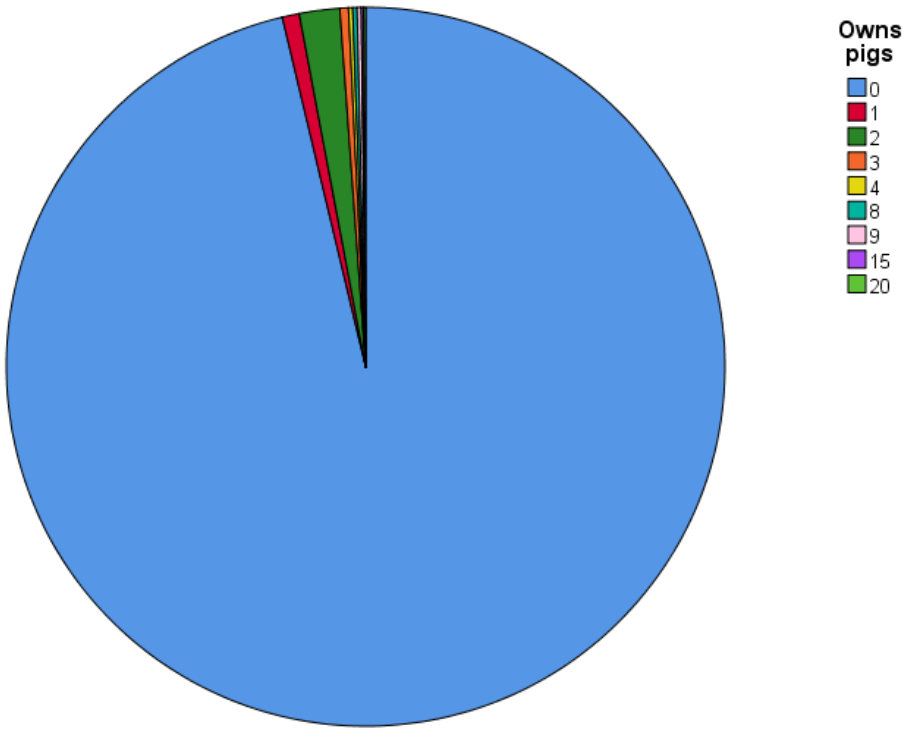

1,000
800
600
400
200
0
Count
0
1
2
Children under 5 slept under mosquito bed net

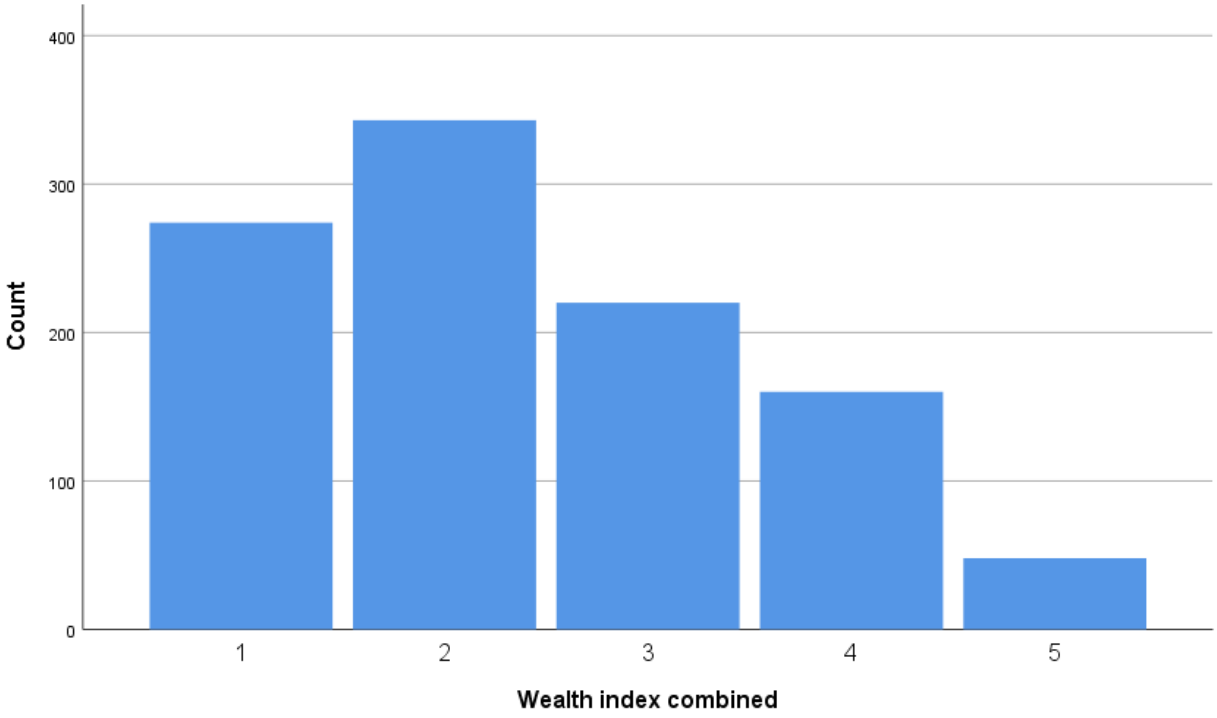

800
600
400
200
0
Count
1
2
3
Location of source for water

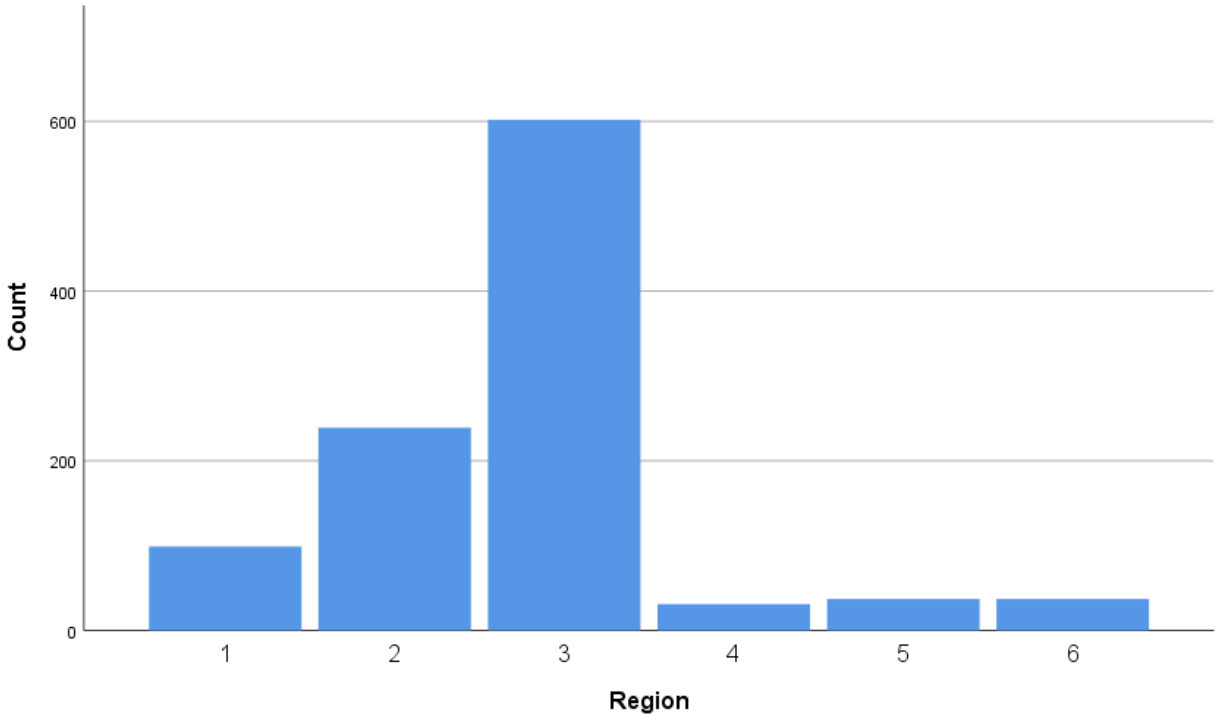

600
400
200
0
Count
1
2
3
Location of toilet facility

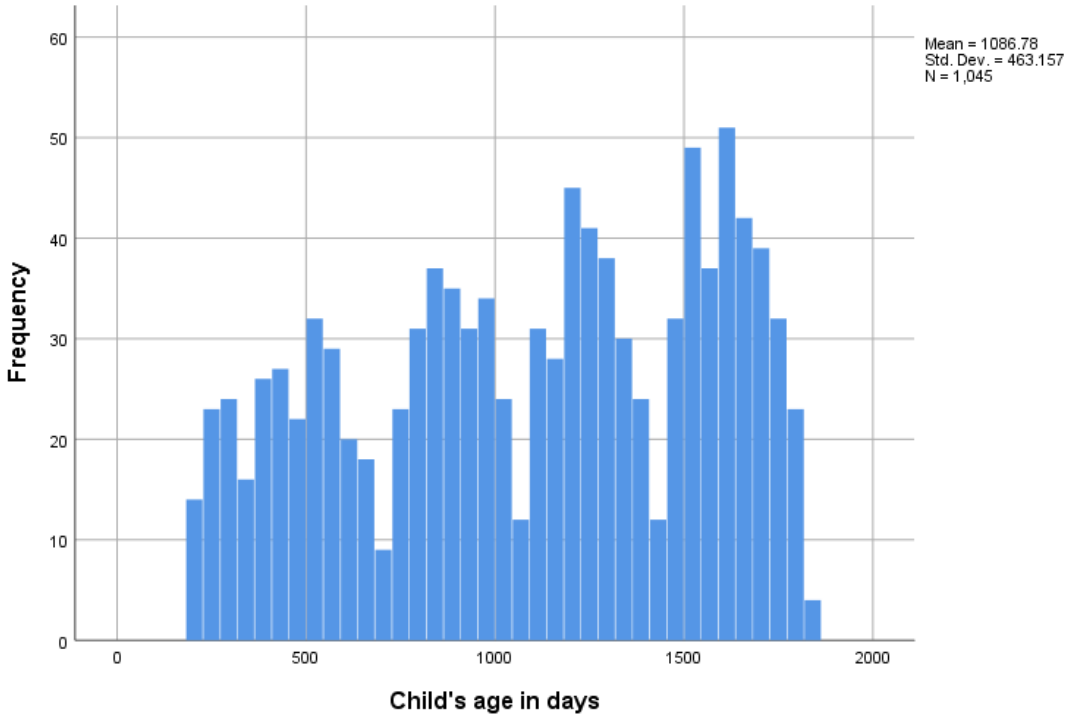

1,000
800
600
400
200
0
Count
0
1
Owns land usable for agriculture

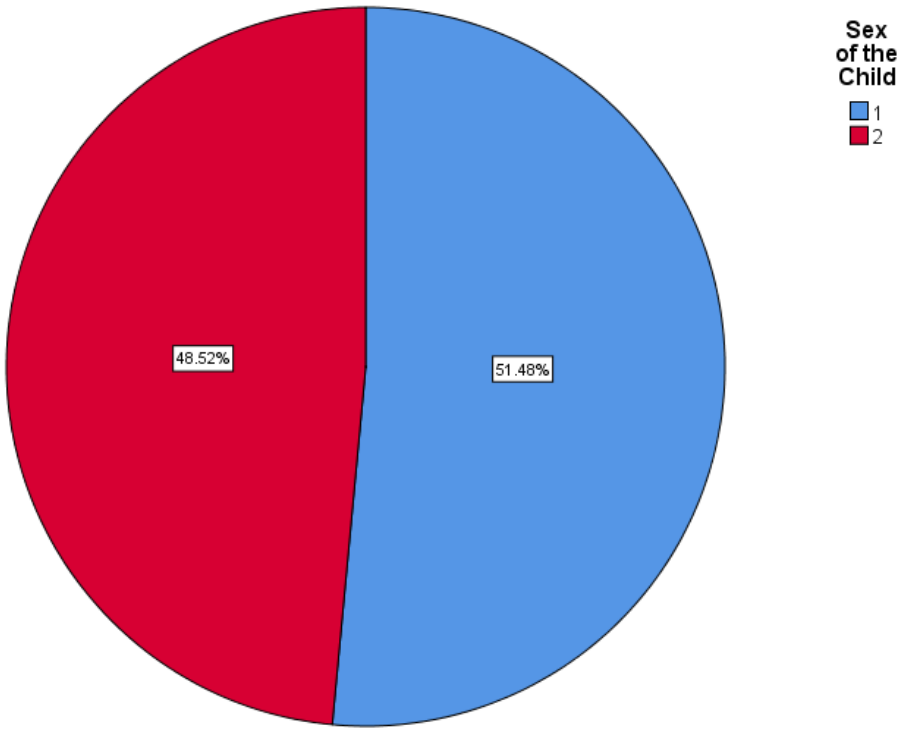

800
600
400
200
0
Count
0
1
Owns livestock, herds or farm animals

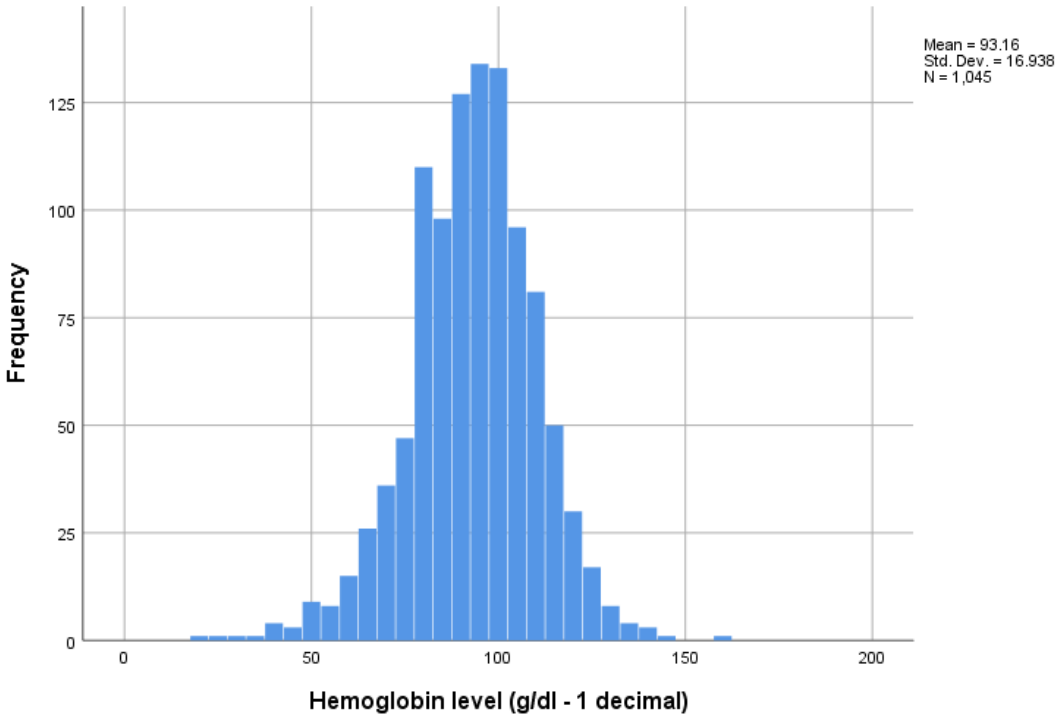

Owns pigs
0
1
2
3
4
8
9
15
20

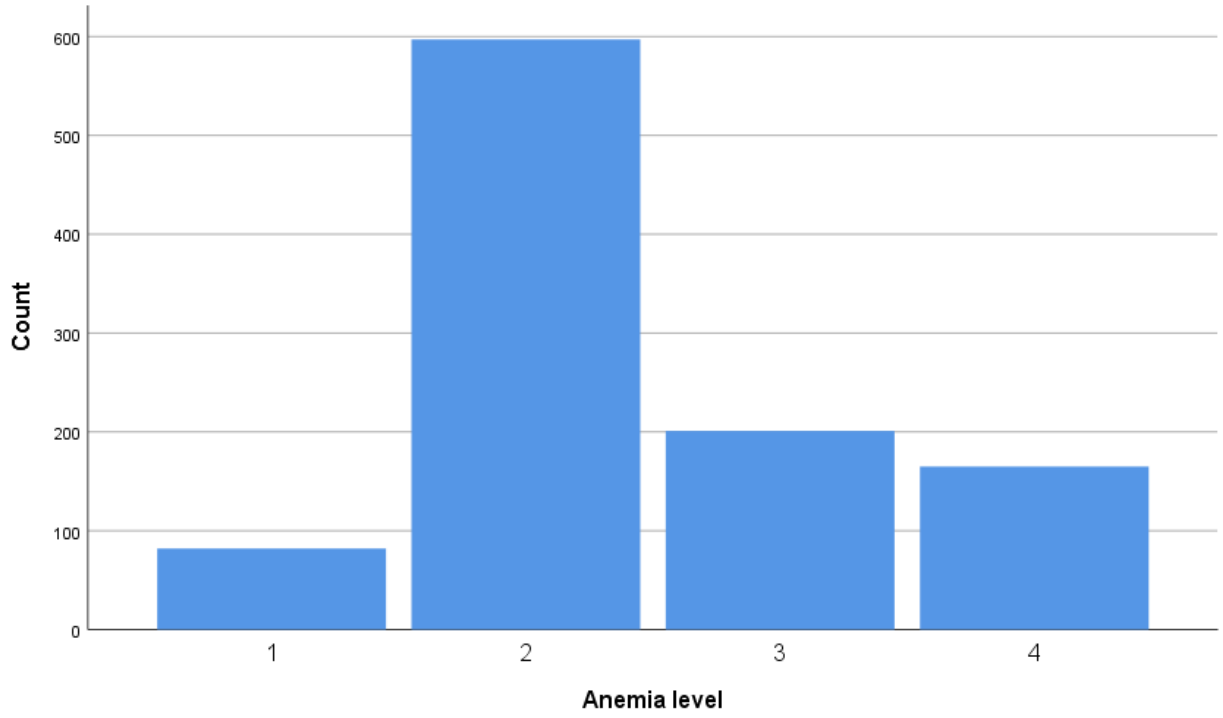

400
300
200
100
0
Count
1
2
3
4
5
Wealth index combined

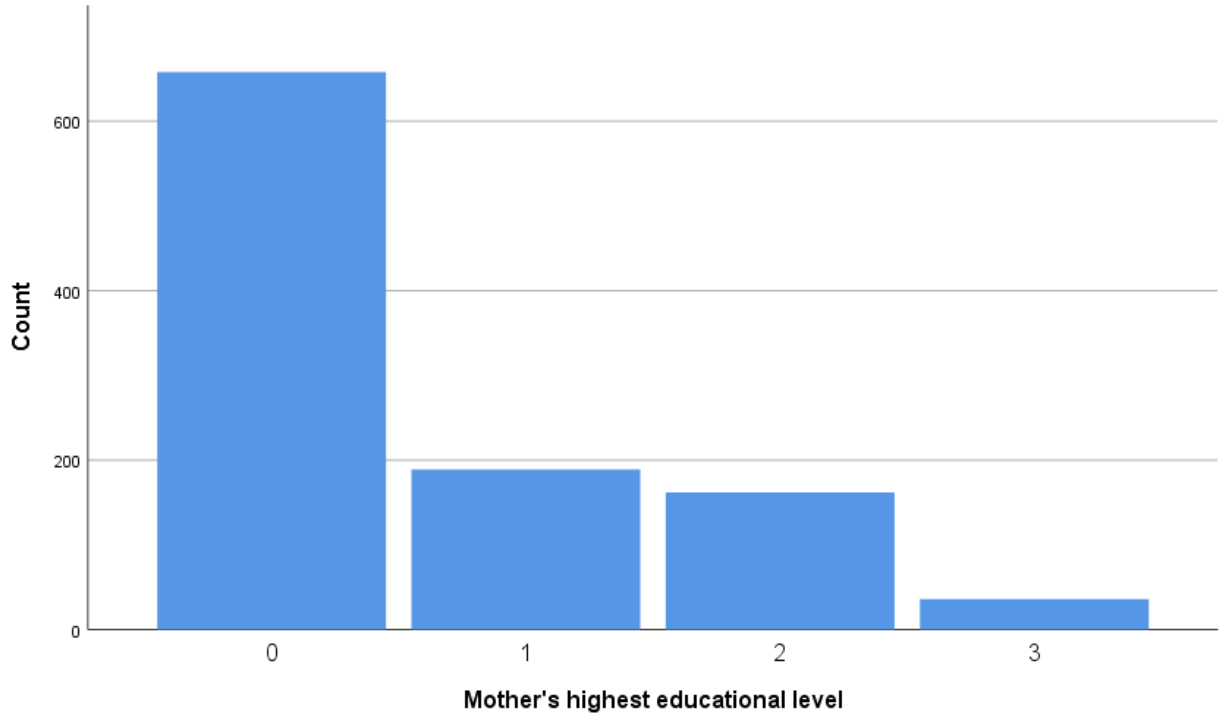

Count
600
400
200
0
1
2
3
4
5
6
Region

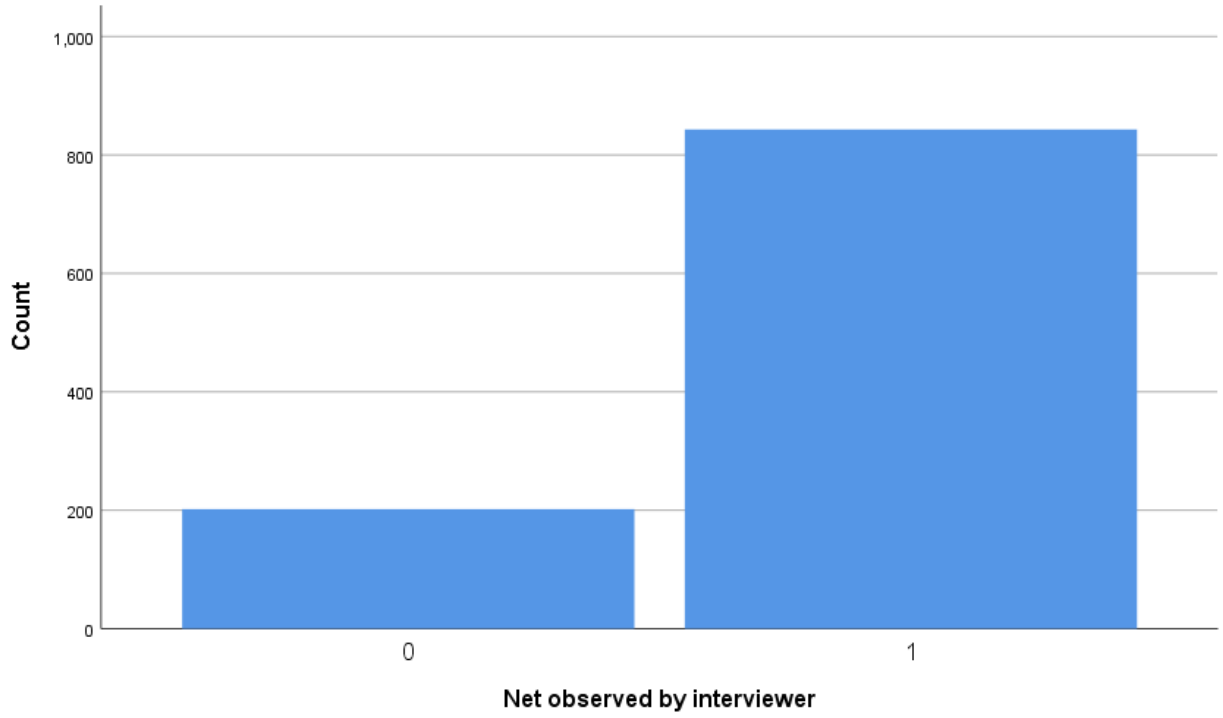

Frequency
60
50
40
30
20
10
0
0
500
1000
1500
2000
Child's age in days
Mean = 1086.78
Std. Dev. = 463.157
N = 1,045

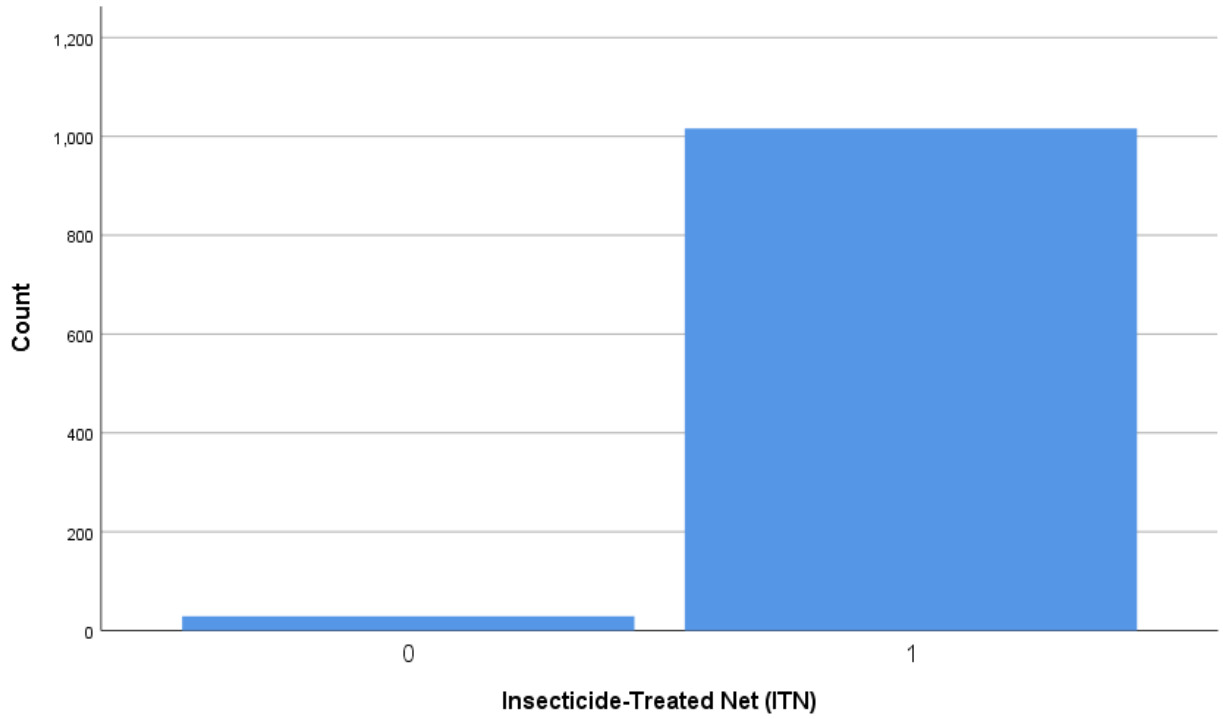

Sex of the Child
1
2
48.52%
51.48%

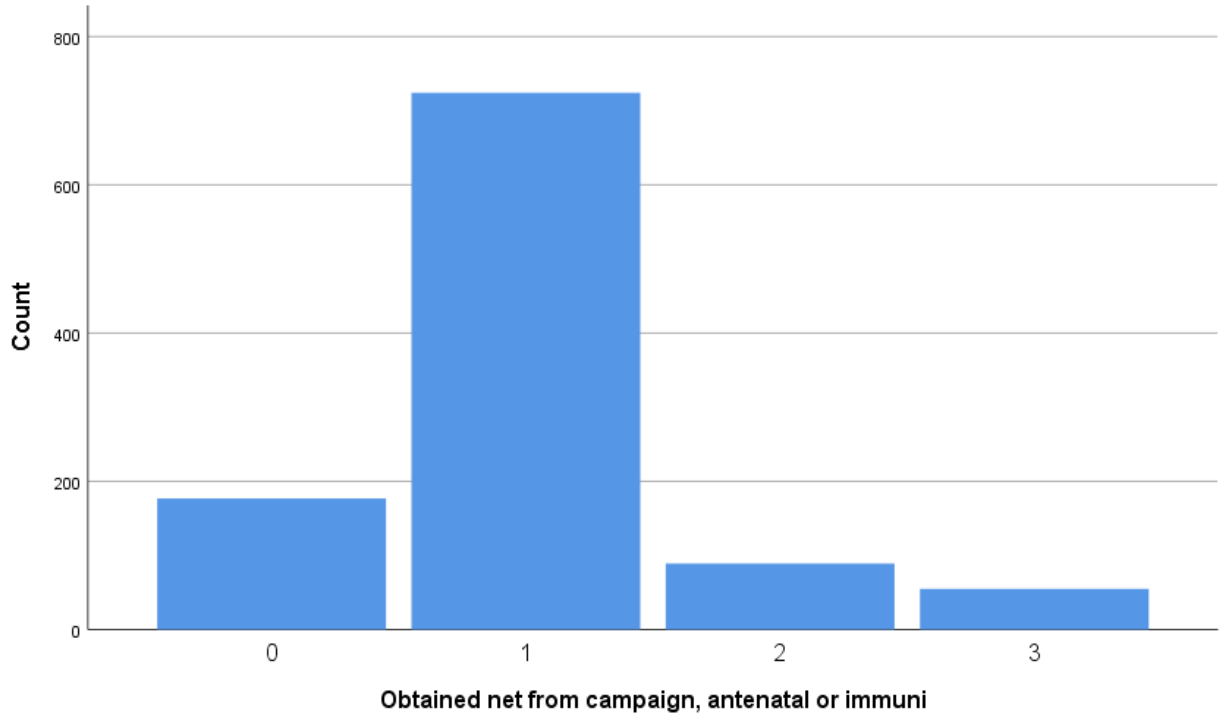

Mean = 93.16
Std. Dev. = 16.938
N = 1,045
Frequency
125
100
75
50
25
0
0
50
100
150
200
Hemoglobin level (g/dl - 1 decimal)

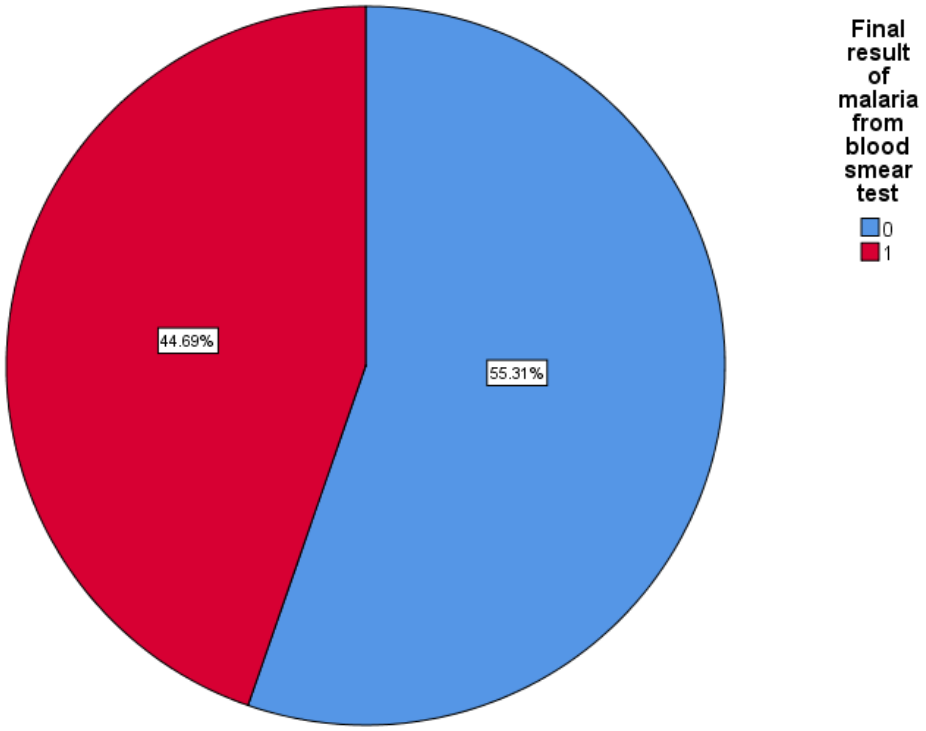

600
500
400
300
200
100
0
Count
1
2
3
4
Anemia level

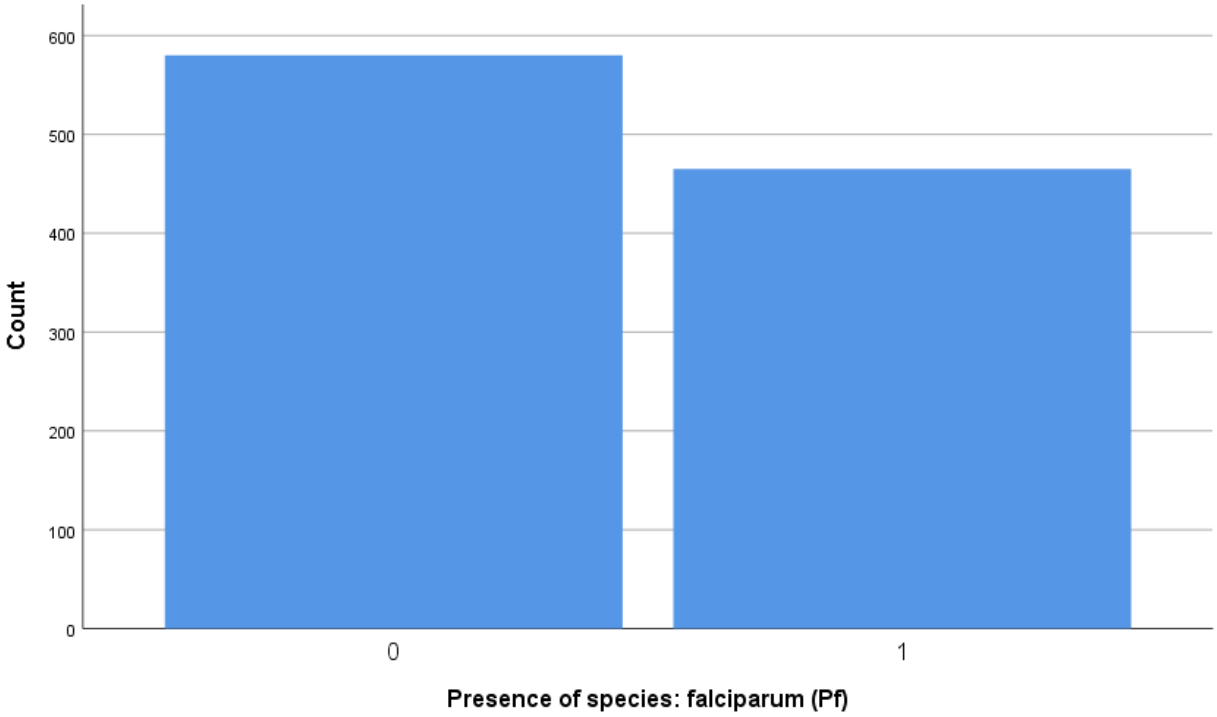

600
400
200
0
Count
0
1
2
3
Mother's highest educational level

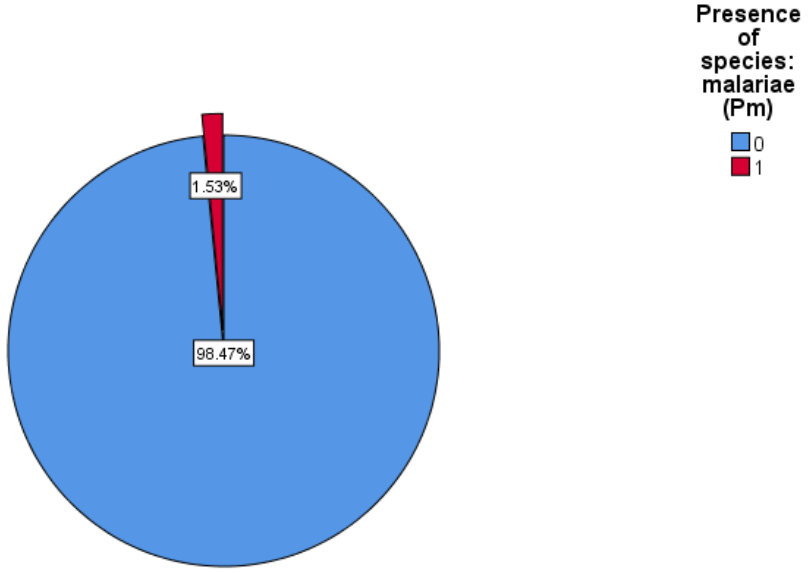

1,000
800
600
400
200
0
Count
0
1
Net observed by interviewer

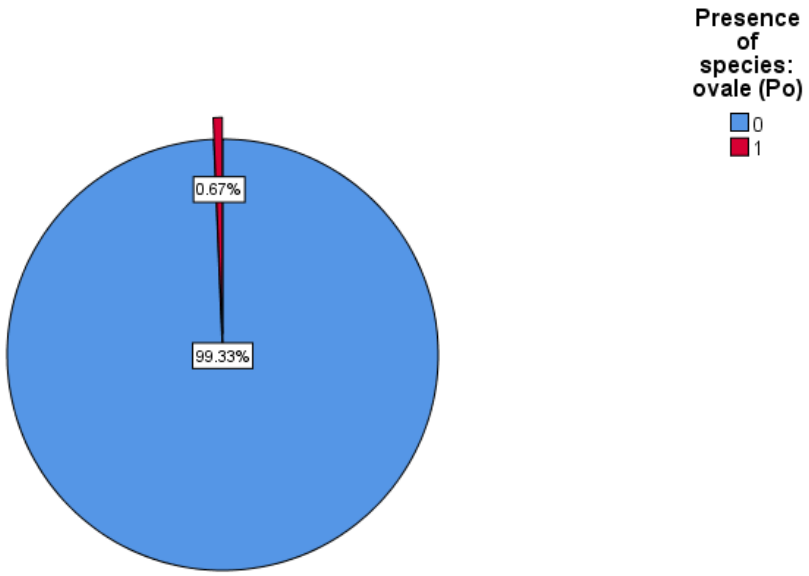

1,200
1,000
800
600
400
200
0
Count
0
1
Insecticide-Treated Net (ITN)

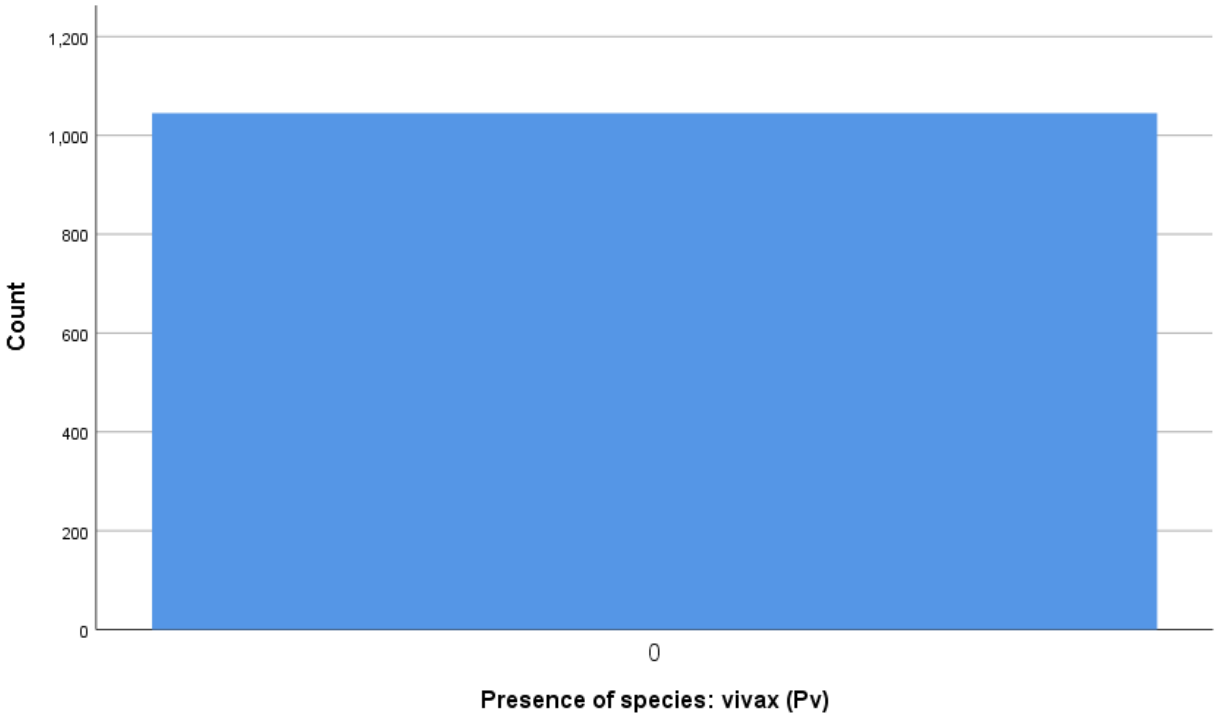

800
600
400
200
0
Count
0
1
2
3
Obtained net from campaign, antenatal or immuni

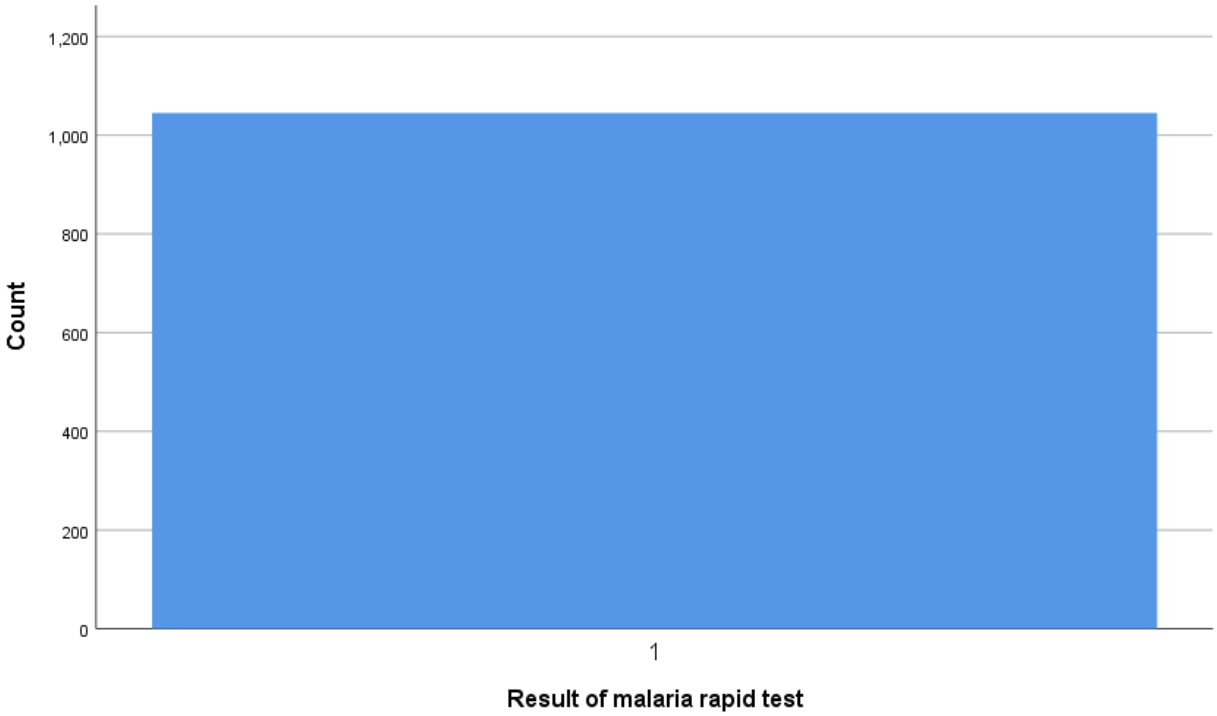

44.69%
55.31%
Final result of malaria from blood smear test
0
1

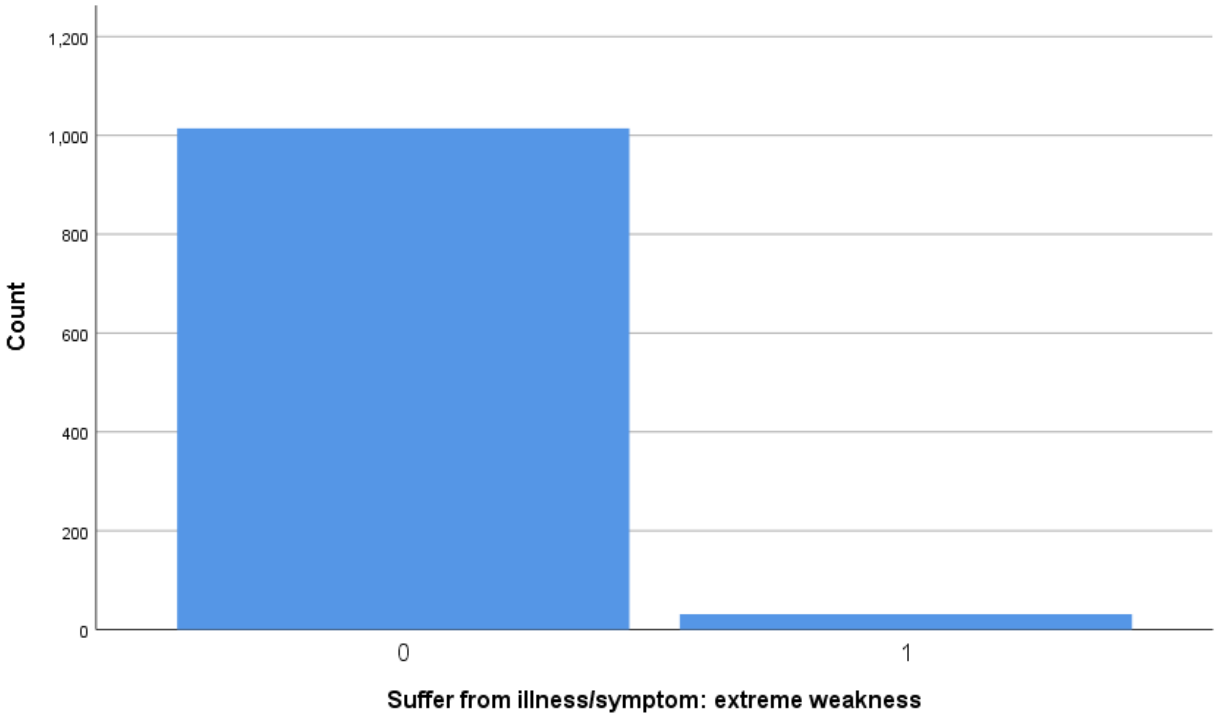

600
500
400
300
200
100
0
Count
0
1
Presence of species: falciparum (Pf)

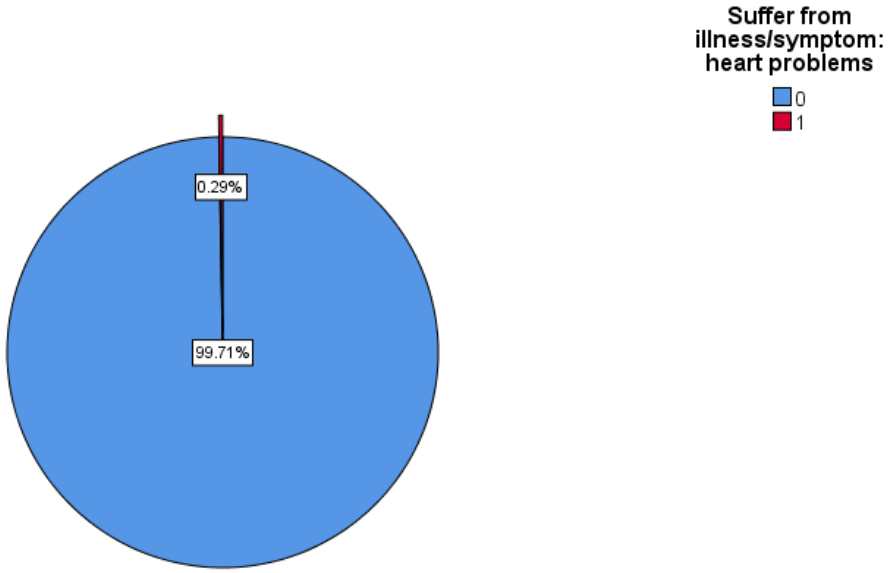

Presence of species: malariae (Pm)
0
1
1.53%
98.47%

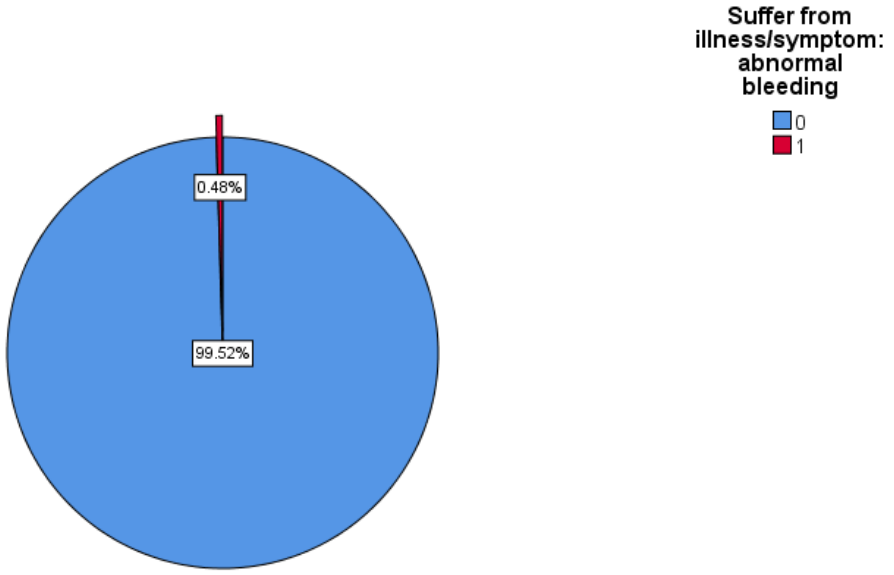

Presence of species: ovale (Po)
0
1
0.67%
99.33%

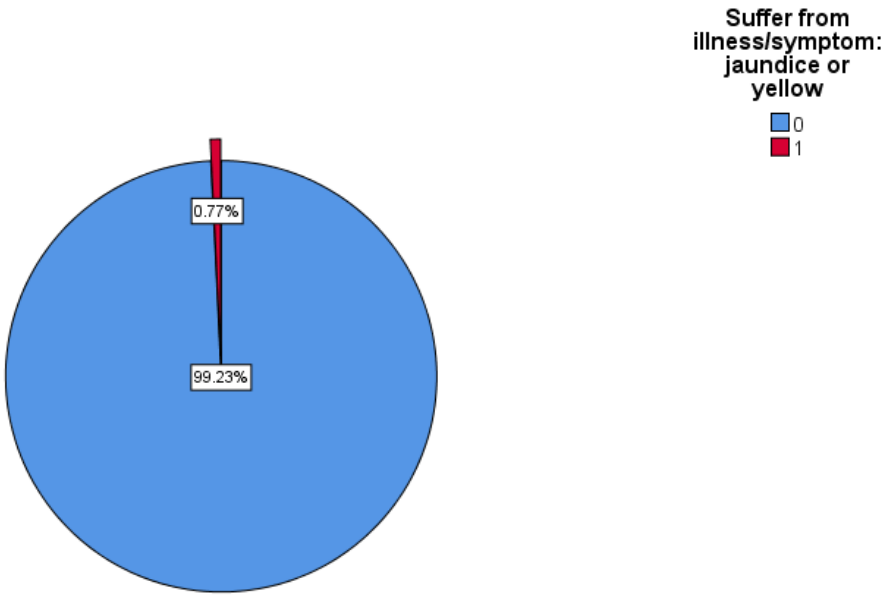

1,200
1,000
800
600
400
200
0
Count
0
Presence of species: vivax (Pv)

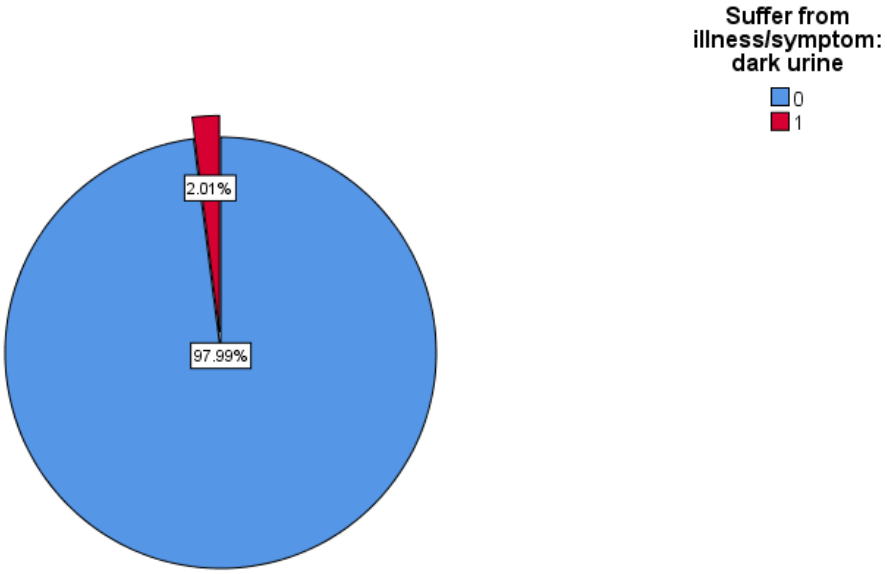

1,200
1,000
800
600
400
200
0
Count
1
Result of malaria rapid test

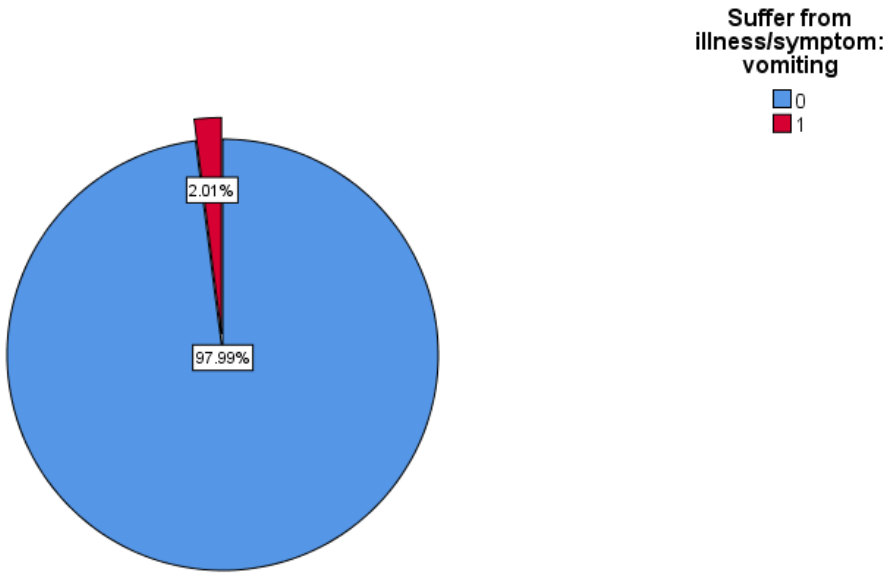

1,200
1,000
800
600
400
200
0
Count
0
1
Suffer from illness/symptom: extreme weakness

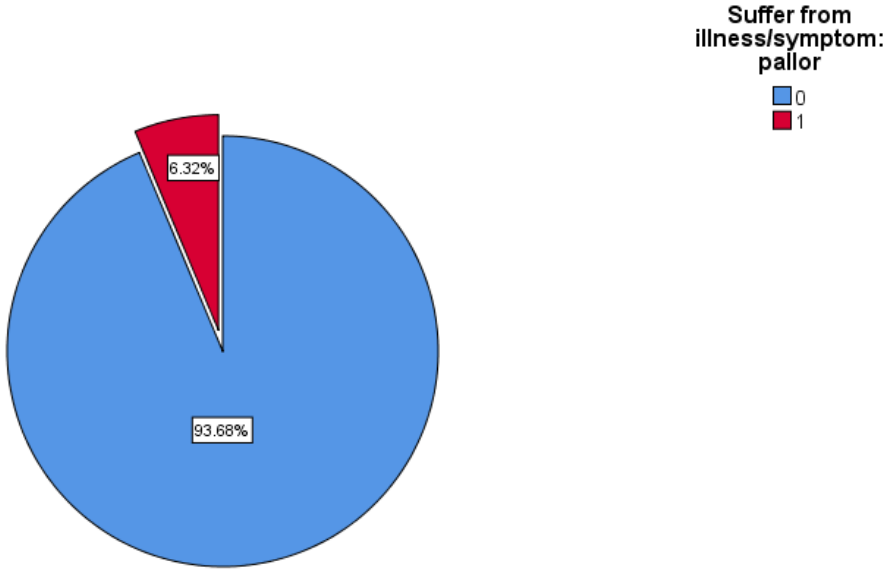

Suffer from illness/symptom: heart problems
0
1
0.29%
99.71%

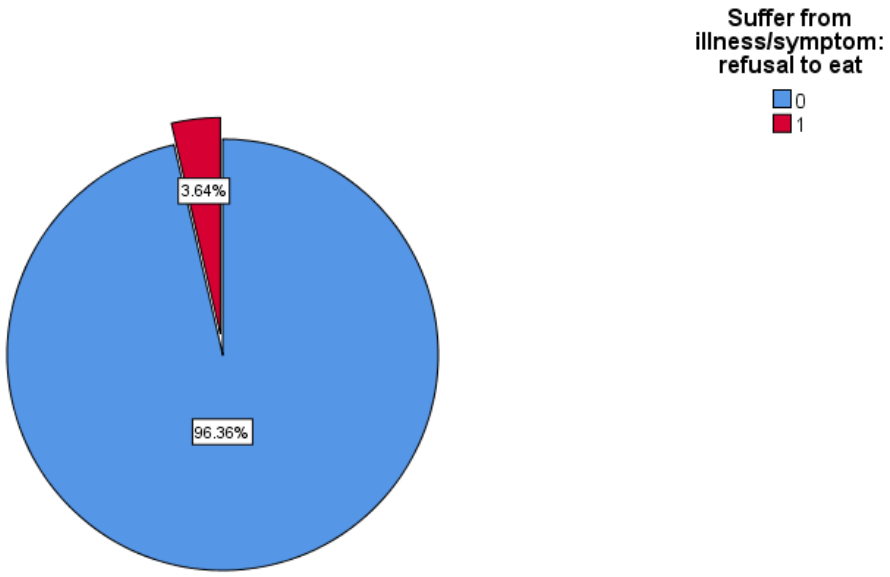

Suffer from illness/symptom: abnormal bleeding
0
1
0.48%
99.52%

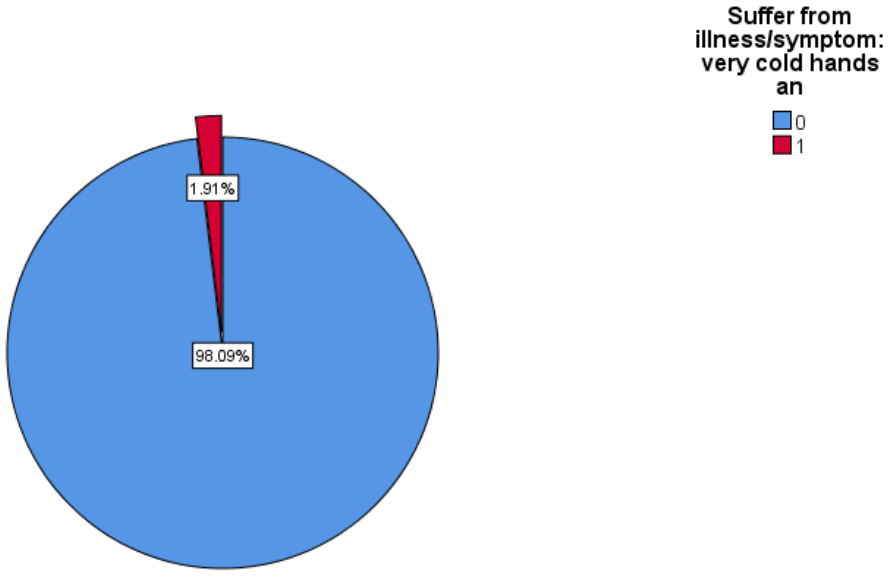

Suffer from illness/symptom: jaundice or yellow
0
1
0.77%
99.23%


Suffer from illness/symptom: dark urine
0
1
2.01%
97.99%


Suffer from illness/symptom: vomiting
0
1
2.01%
97.99%


Suffer from illness/symptom: pallor
0
1
6.32%
93.68%


Suffer from illness/symptom: refusal to eat
0
1
3.64%
96.36%


Suffer from illness/symptom: very cold hands an
0
1
1.91%
98.09%

4 Data Processing

4.1 Dealing with Missing Data

In this study a root level survey data is used, and this survey is conducted by human. And missing value is obvious due to human error.

As our focus is to predict the Malaria positive/negative, so first of all whole dataset of selected features was filtered based on the available data of feature *"Final result of malaria from blood smear test" (df[(df['hml32'] == 0) | (df['hml32'] == 1)]).* Then three features "SH130 - Reason net was not used", "HV202 - Source of non-drinking water" and "HML23 - Place where net was obtained" were eliminated due to data unavailability (very small amount of data were available after filtering).

After the first filtering five features showed the existence of missing values within them. So, gradually five filtering were done to remove missing values.

Filters for removing missing values:

a. df[(df['hml37l'] == 0) | (df['hml37l'] == 1)]

b. df[(df['hml22'] == 0) | (df['hml22'] == 1) | (df['hml22'] == 2) | (df['hml22'] == 3)]

c. df[(df['hv225'] == 0) | (df['hv225'] == 1)]

d. df[(df['hc61'] == 0) | (df['hc61'] == 1) | (df['hc61'] == 2) | (df['hc61'] == 3)]

e. df[(df['hv235'] == 1) | (df['hv235'] == 2) | (df['hv235'] == 3)]

After removing missing values and eliminating some features the shape of final dataset was (1045,40) [row = 1045 and columns = 40].

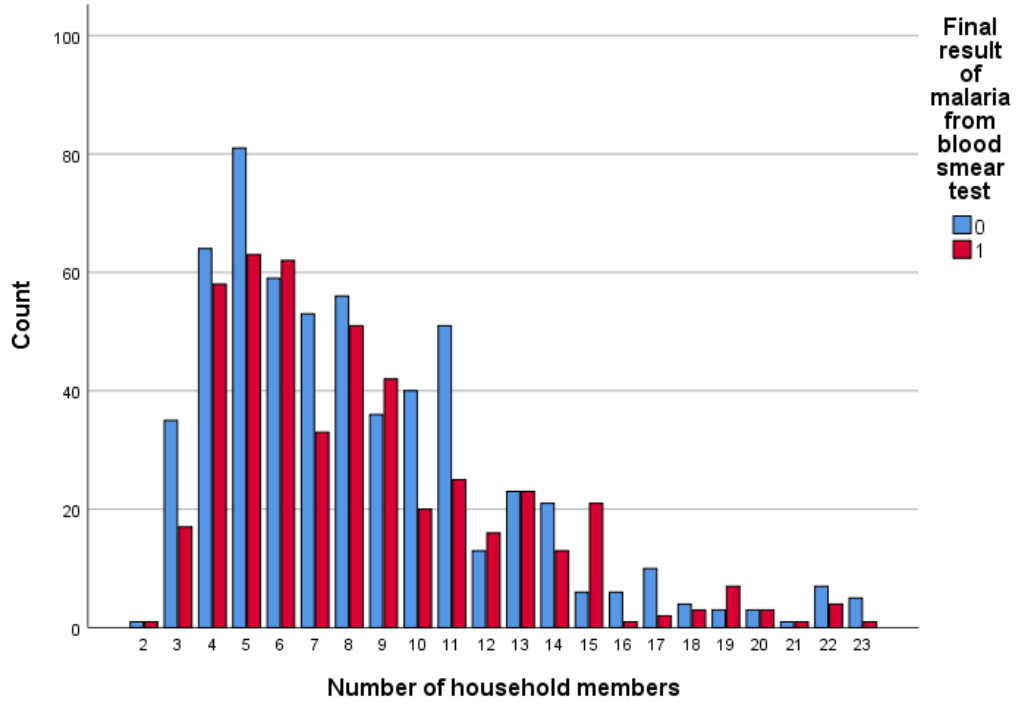
4.2 Feature Calculation

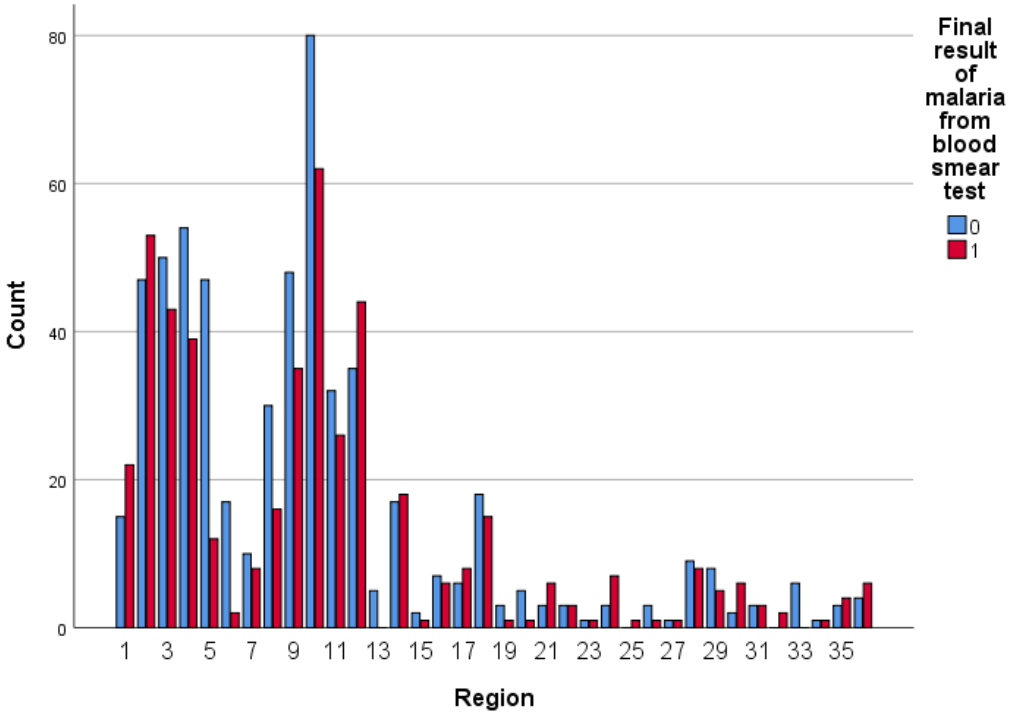
Feature calculation was not need due to the independent* and unique characteristics of the selected features. Label encoding and data type (numerical & categorical) were defined to reduce computational cost.

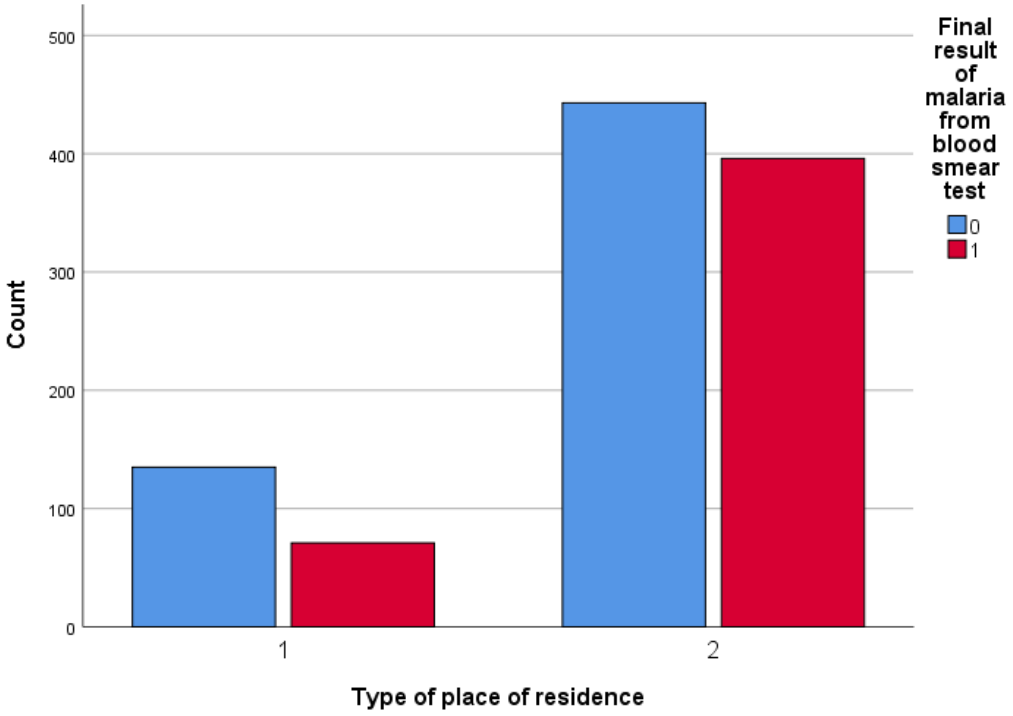
* Here the word “independent” doesn’t mean statistically independent

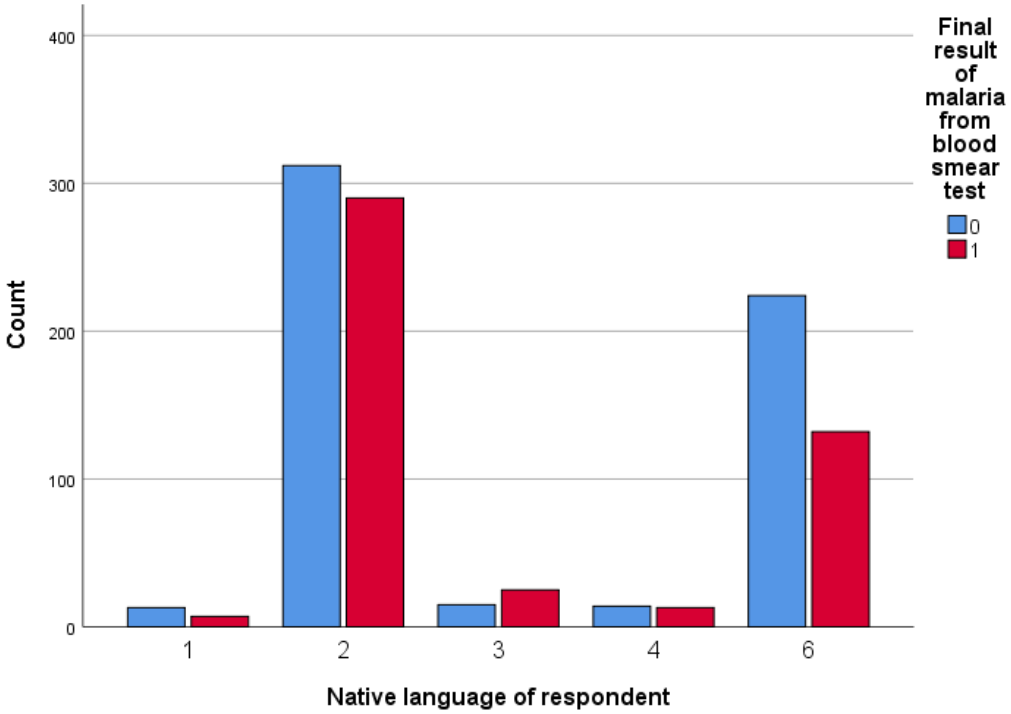
4.3 Relation between Predictors and Outcome

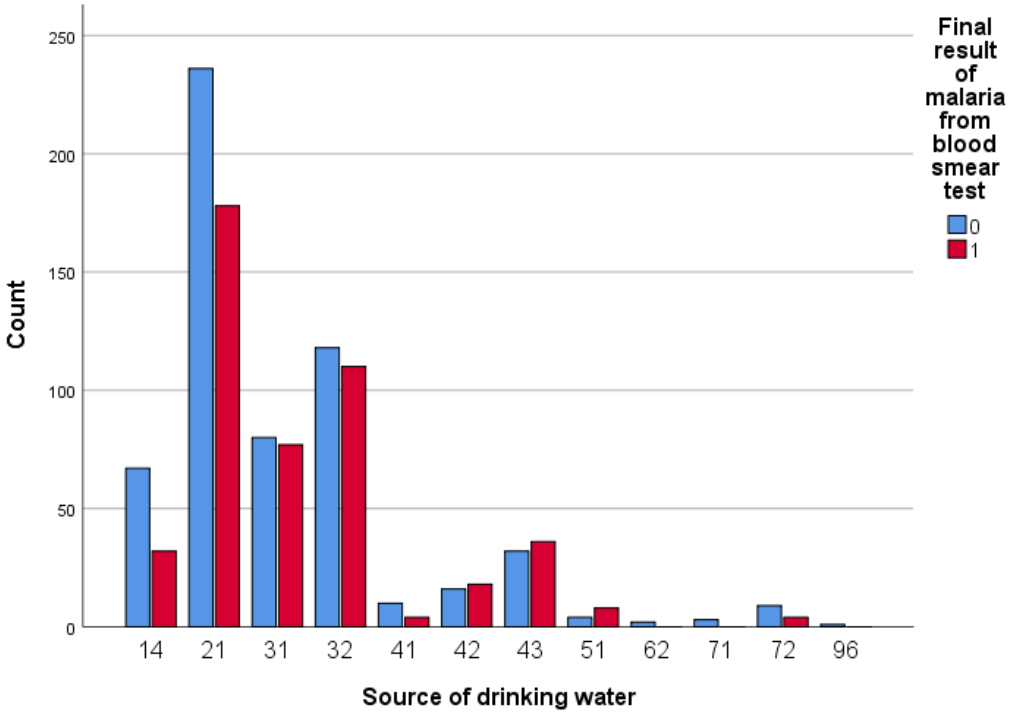
The numerical values represented in the figures refer to the column “Values” in the Table.2

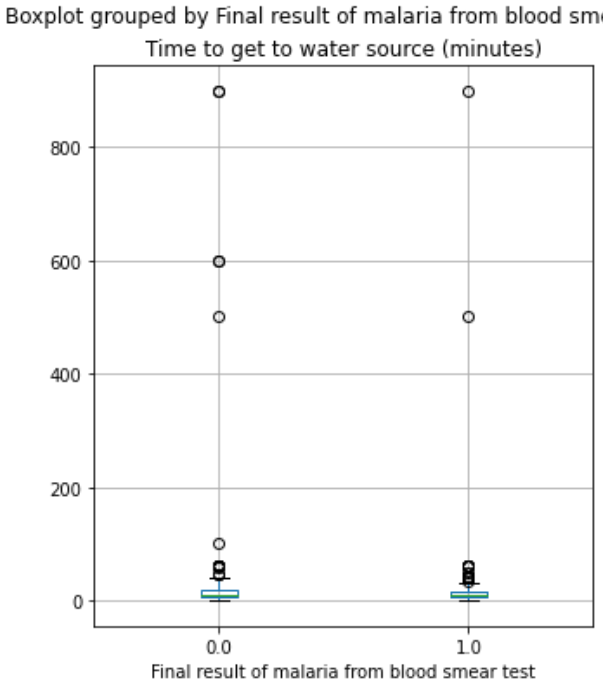


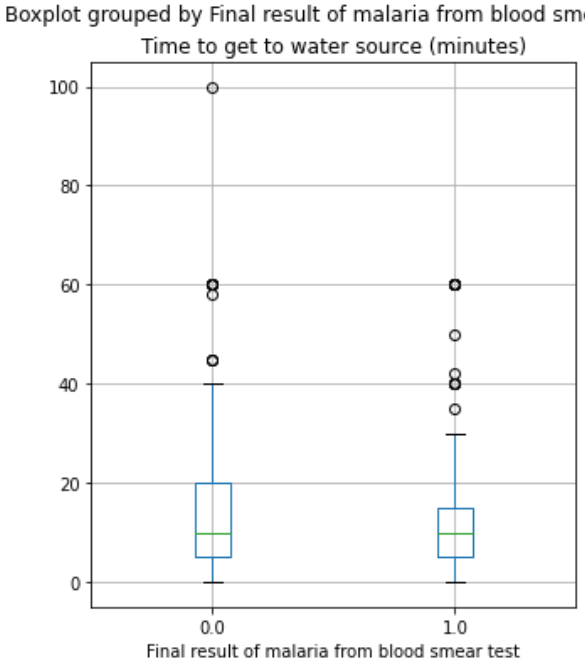

Count
80
60
40
20
0
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31 33 35
Region
Final result of malaria from blood smear test
0
1

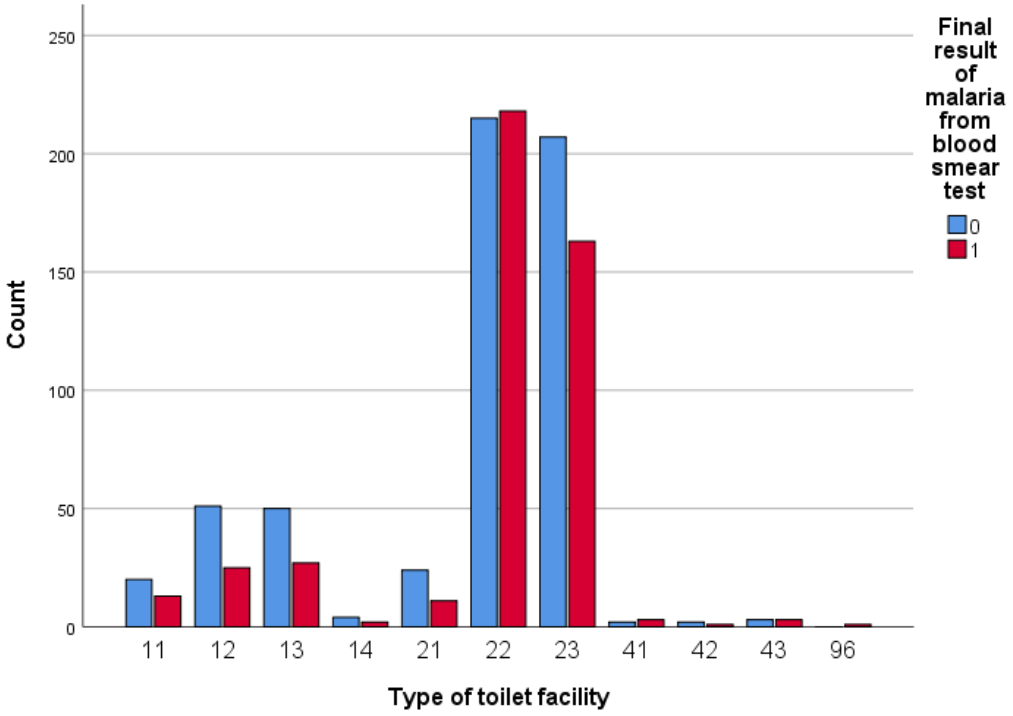

Count
500
400
300
200
100
0
1
2
Type of place of residence
Final result of malaria from blood smear test
0
1

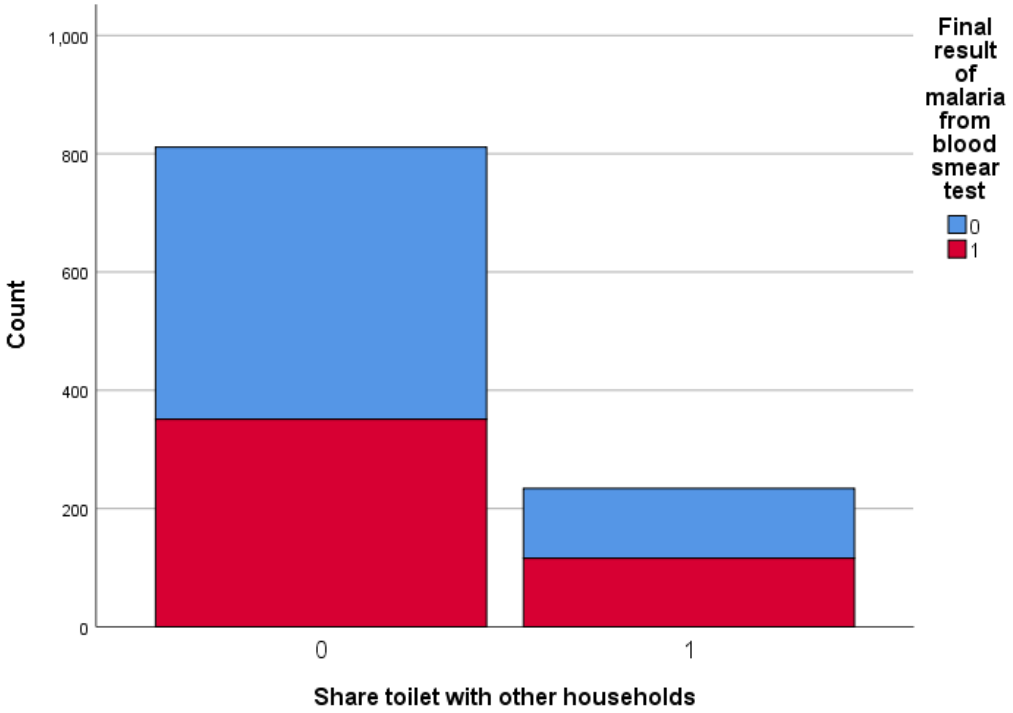

Final result of malaria from blood smear test
0
1
Count
400
300
200
100
0
1
2
3
4
6
Native language of respondent

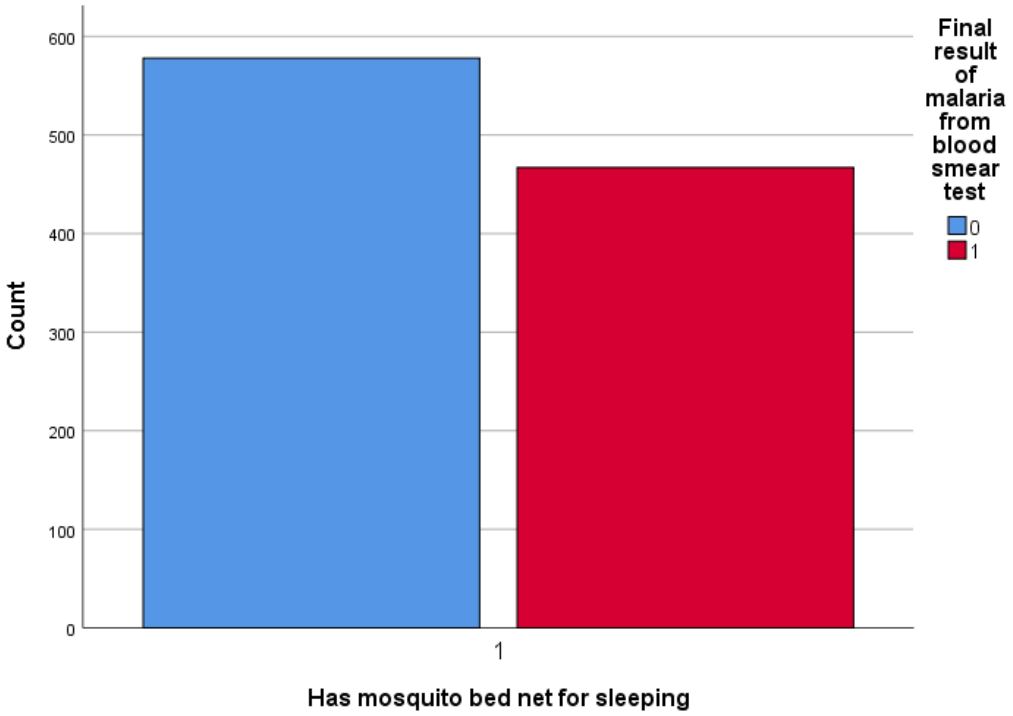

Final result of malaria from blood smear test
0
1
Count
250
200
150
100
50
0
14
21
31
32
41
42
43
51
62
71
72
96
Source of drinking water

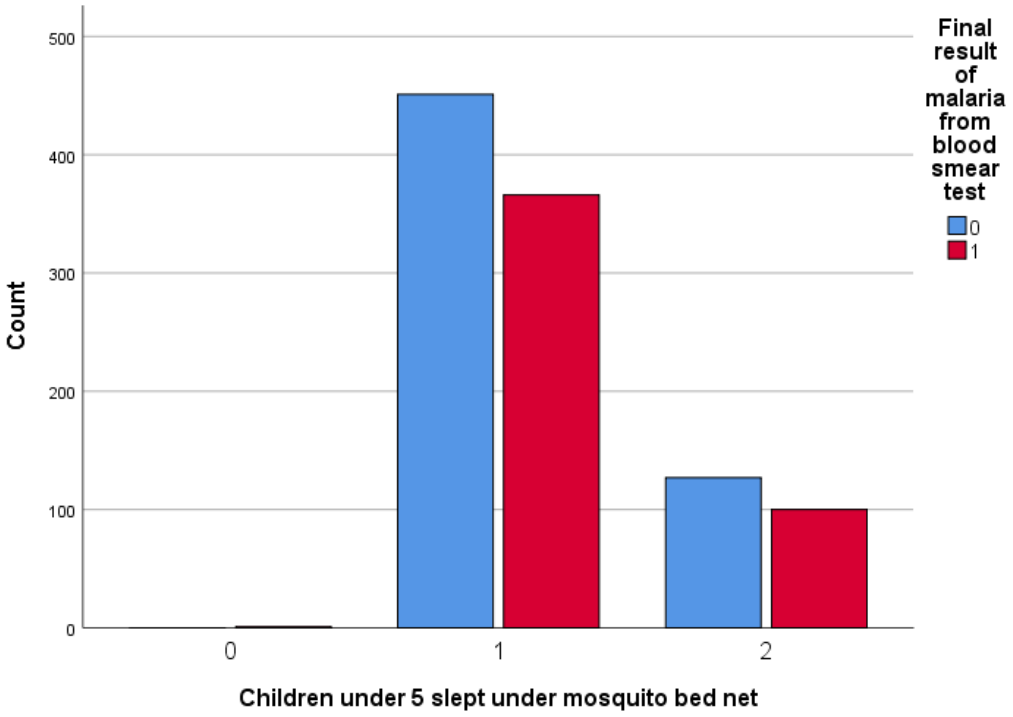

Boxplot grouped by Final result of malaria from blood smear test
Time to get to water source (minutes)
800
600
400
200
0
0.0
1.0
Final result of malaria from blood smear test

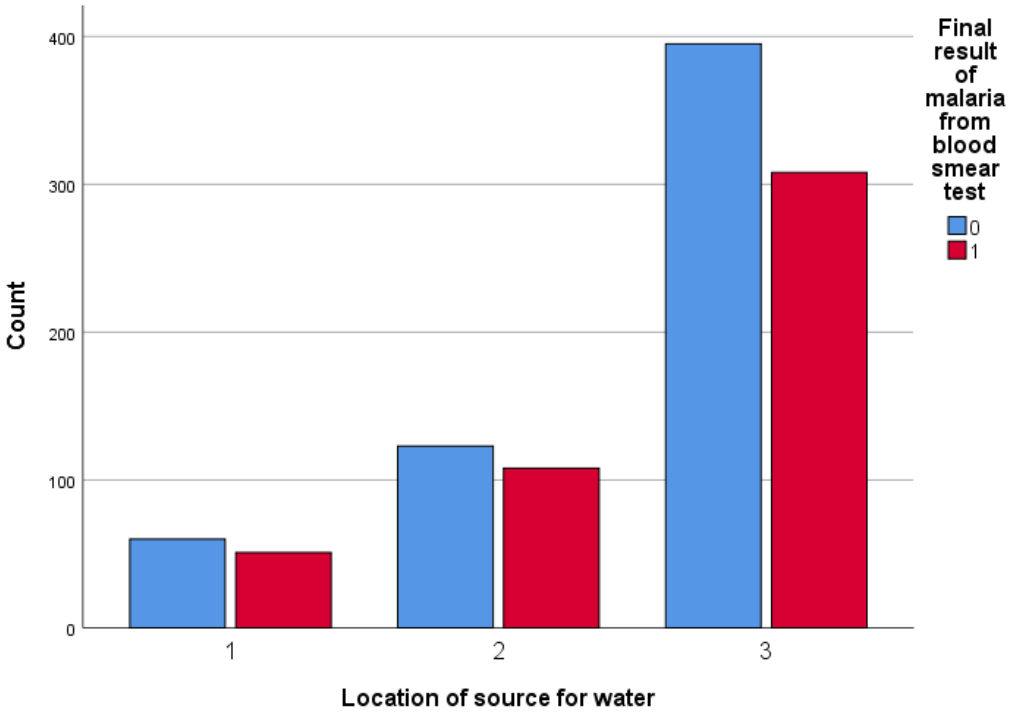

Boxplot grouped by Final result of malaria from blood smear test
Time to get to water source (minutes)
100
80
60
40
20
0
0.0
1.0
Final result of malaria from blood smear test

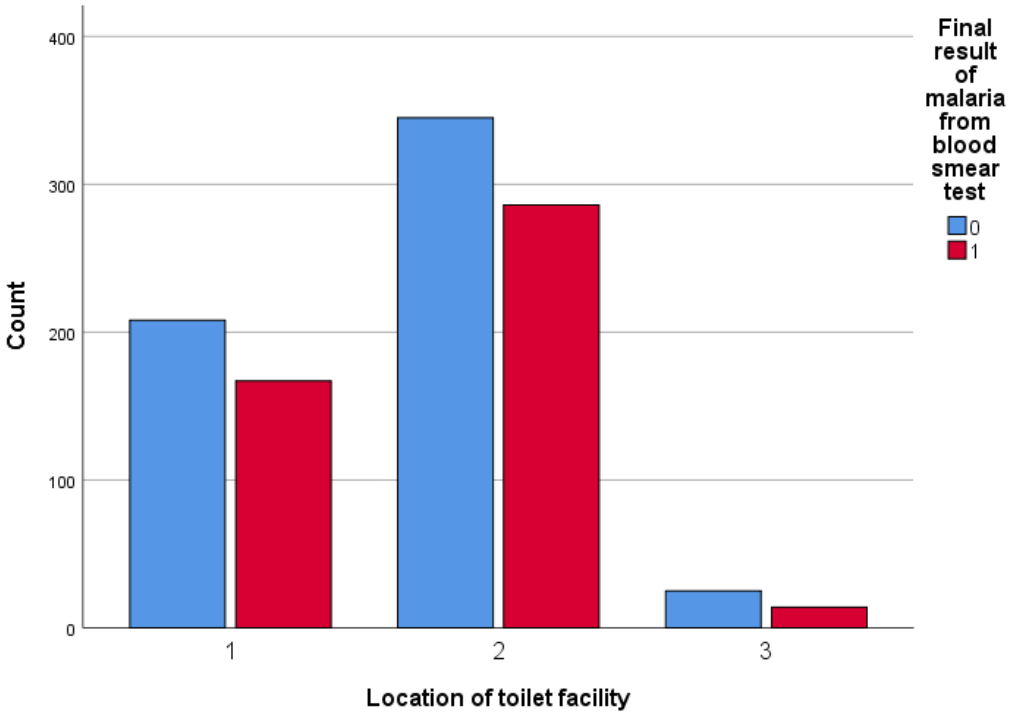

Count
250
200
150
100
50
0
11
12
13
14
21
22
23
41
42
43
96
Type of toilet facility
Final result of malaria from blood smear test
0
1

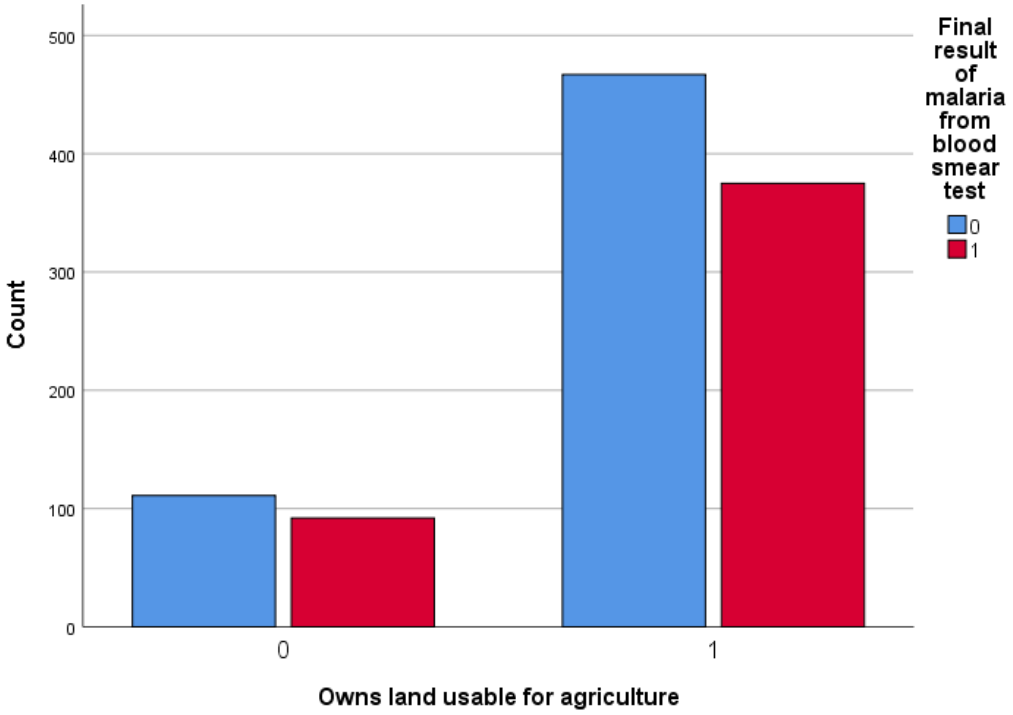

Count
1,000
800
600
400
200
0
0
1
Share toilet with other households
Final result of malaria from blood smear test
0
1

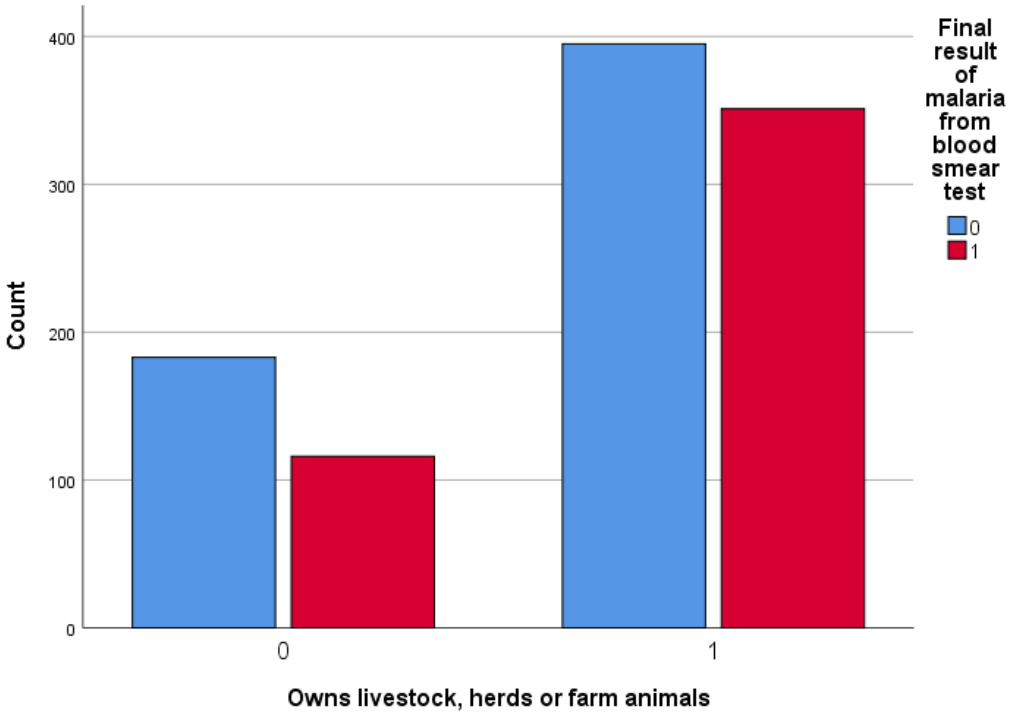

Count
600
500
400
300
200
100
0
1
Has mosquito bed net for sleeping
Final result of malaria from blood smear test
0
1

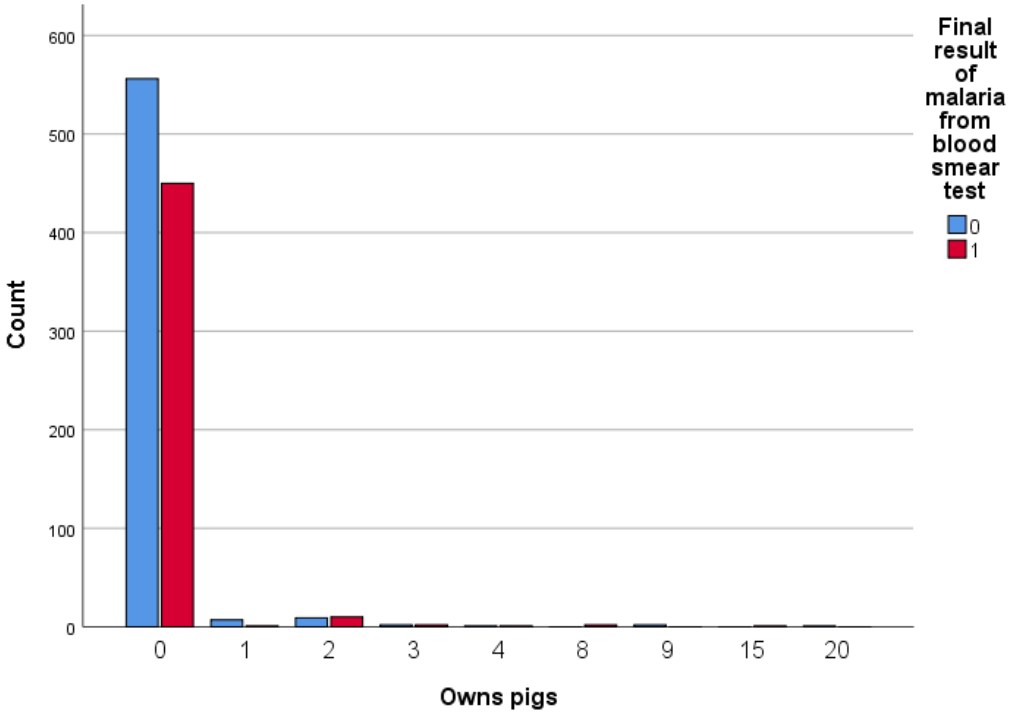

Count
500
400
300
200
100
0
0
1
2
Children under 5 slept under mosquito bed net
Final result of malaria from blood smear test
0
1

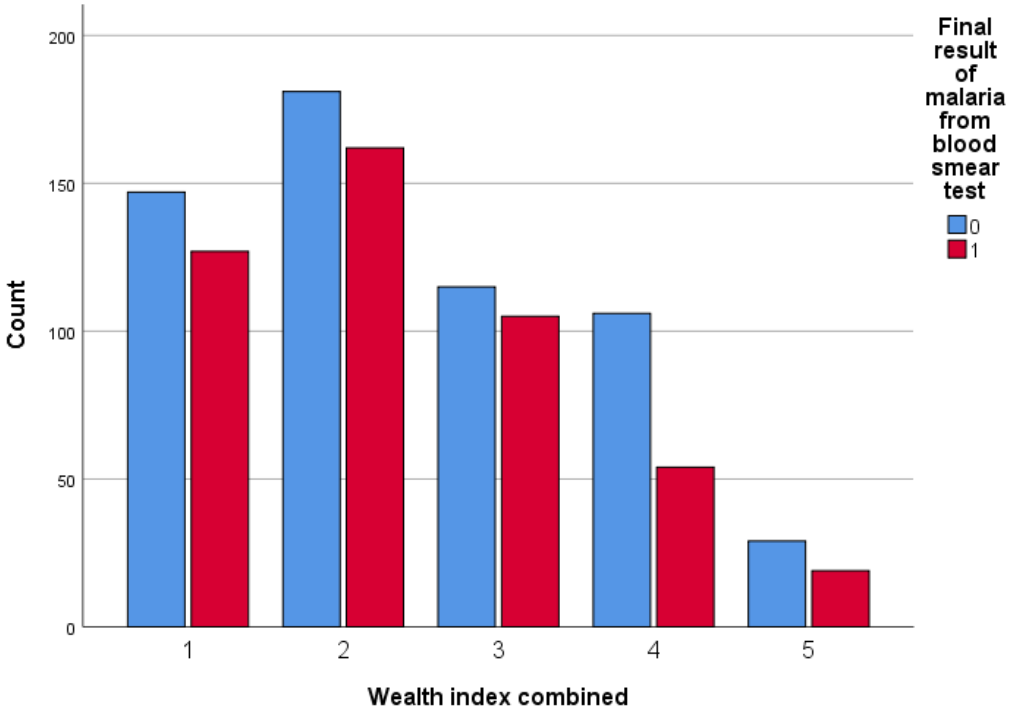

Count
400
300
200
100
0
1
2
3
Location of source for water
Final result of malaria from blood smear test
0
1

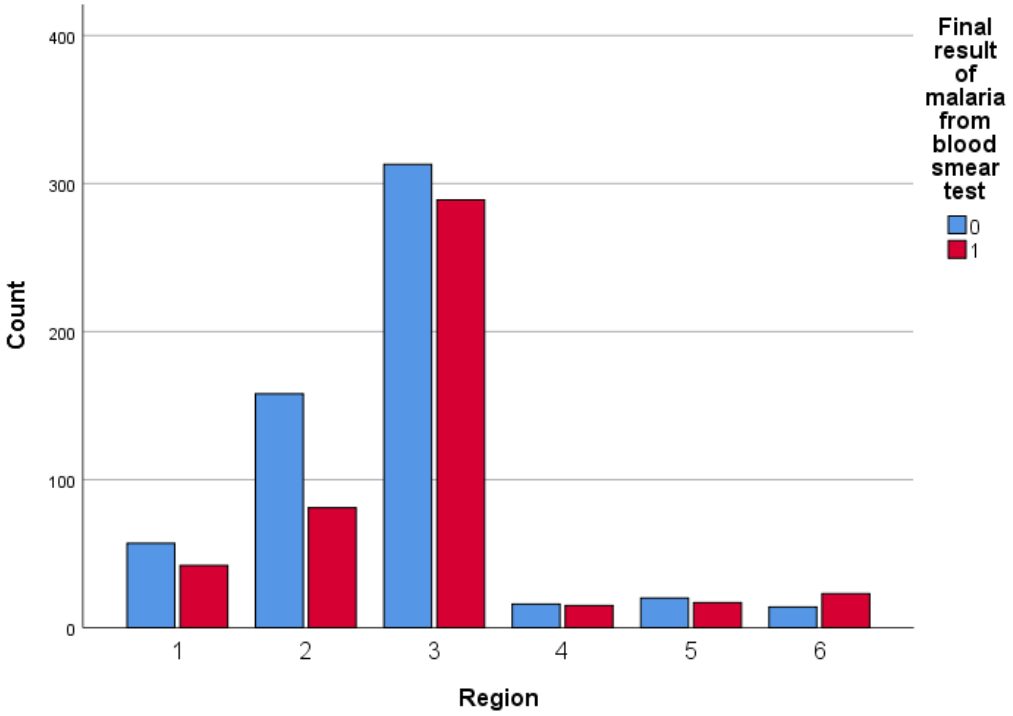

Count
400
300
200
100
0
1
2
3
Location of toilet facility
Final result of malaria from blood smear test
0
1

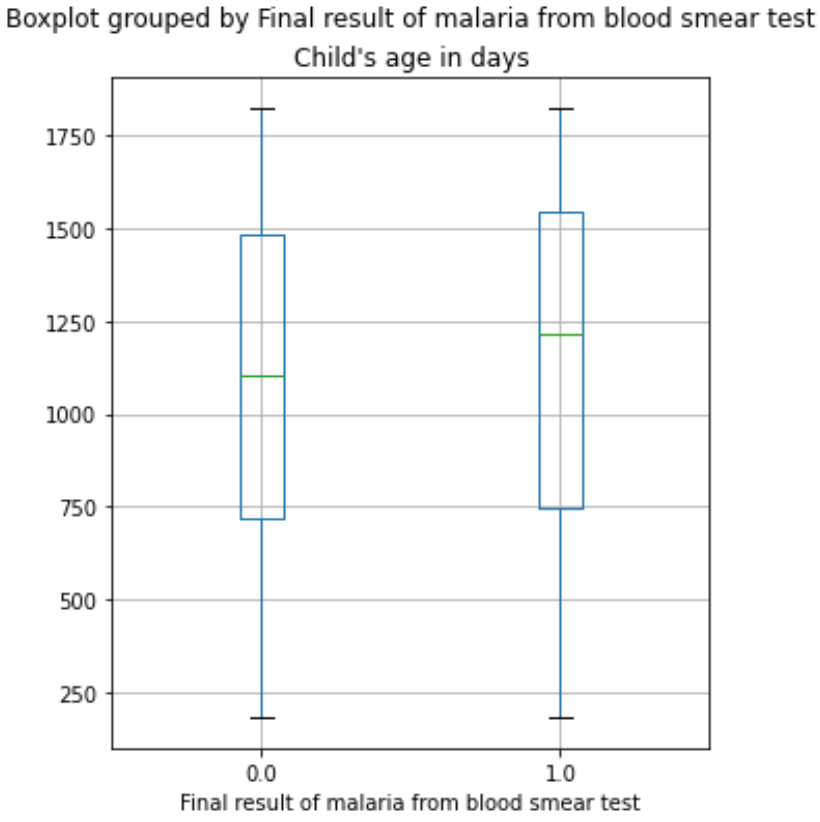

Count
500
400
300
200
100
0
0
1
Owns land usable for agriculture
Final result of malaria from blood smear test
0
1

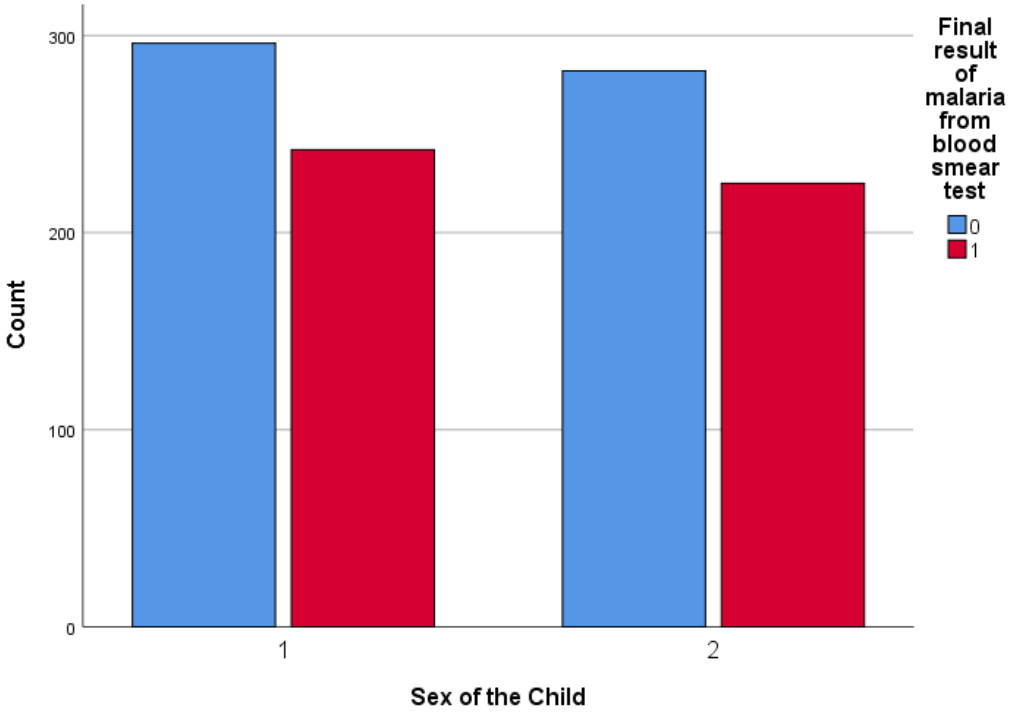

Count
400
300
200
100
0
0
1
Owns livestock, herds or farm animals
Final result of malaria from blood smear test
0
1

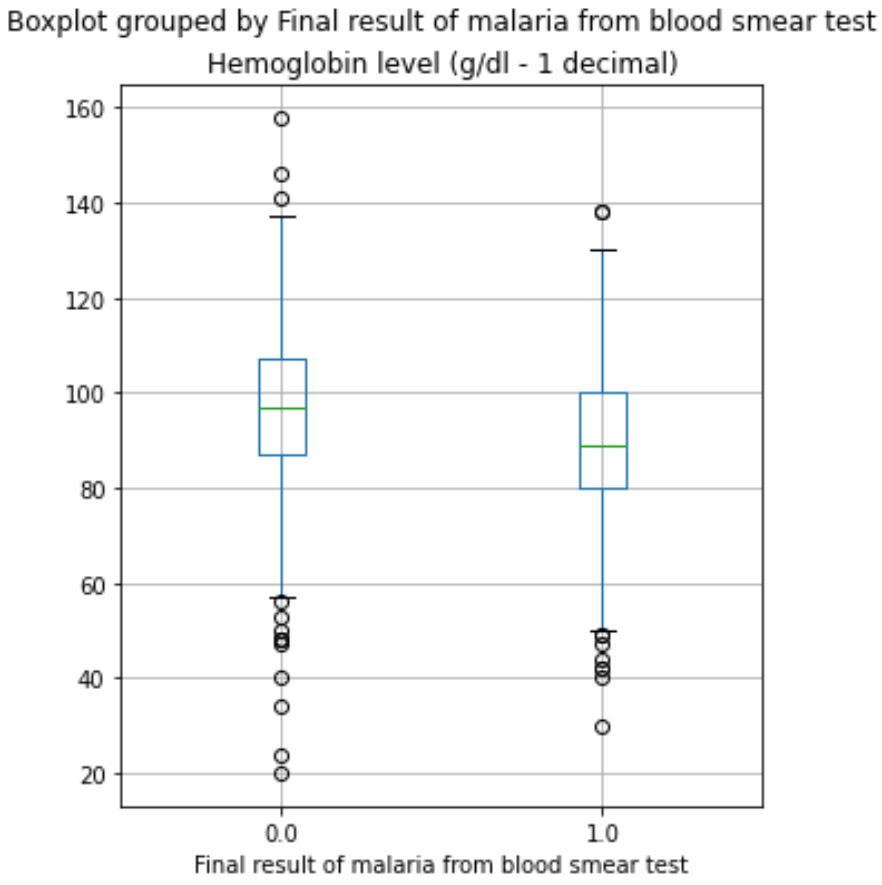

Count
600
500
400
300
200
100
0
0
1
2
3
4
8
9
15
20
Owns pigs
Final result of malaria from blood smear test
0
1

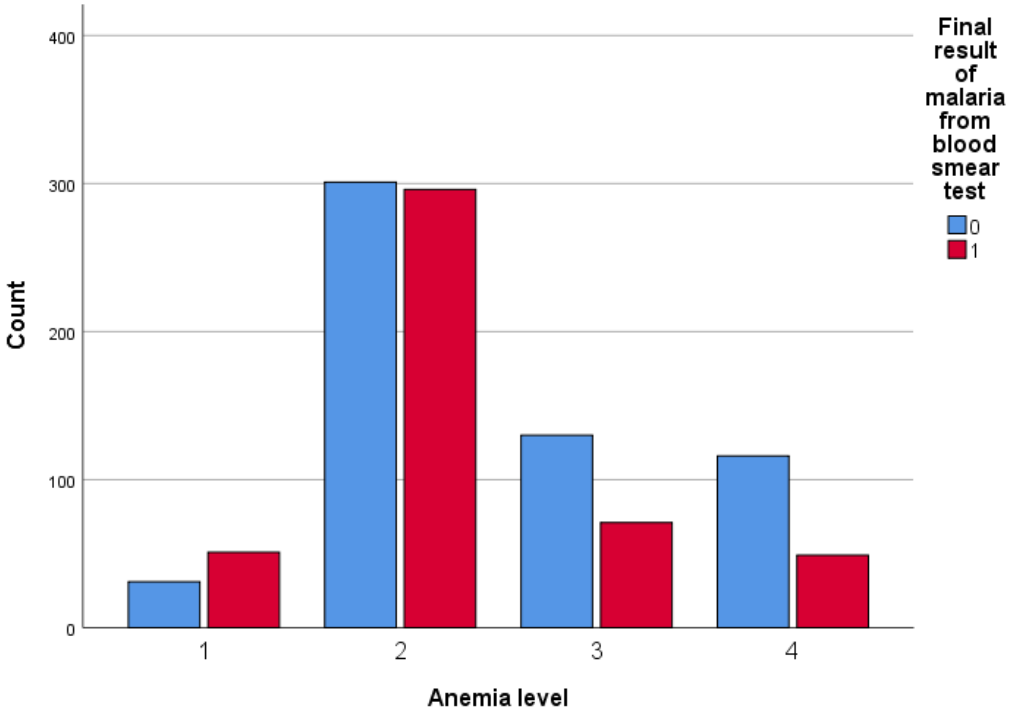

Count
200
150
100
50
0
1
2
3
4
5
Wealth index combined
Final result of malaria from blood smear test
0
1

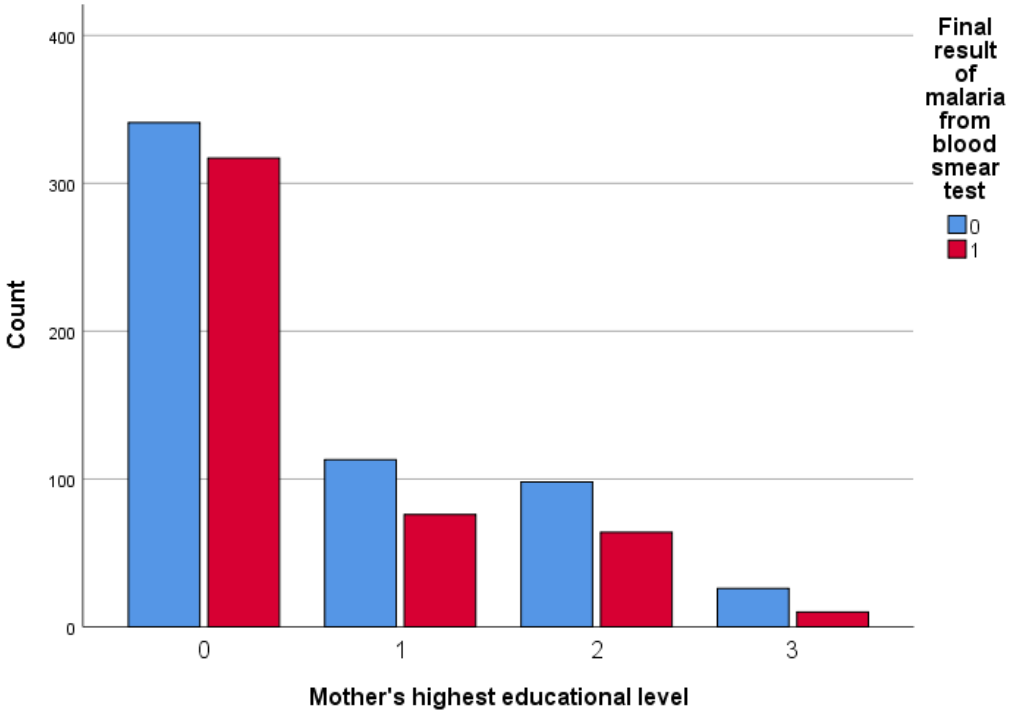

Count
400
300
200
100
0
1
2
3
4
5
6
Region
Final result of malaria from blood smear test
0
1

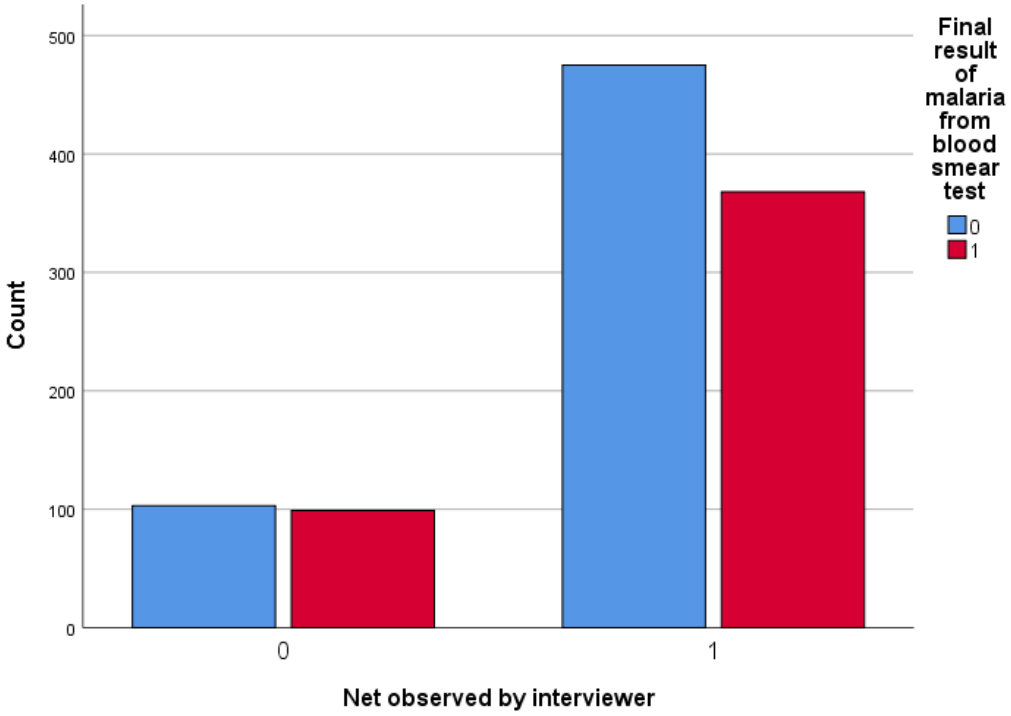

Boxplot grouped by Final result of malaria from blood smear test
Child's age in days
1750
1500
1250
1000
750
500
250
0.0
1.0
Final result of malaria from blood smear test

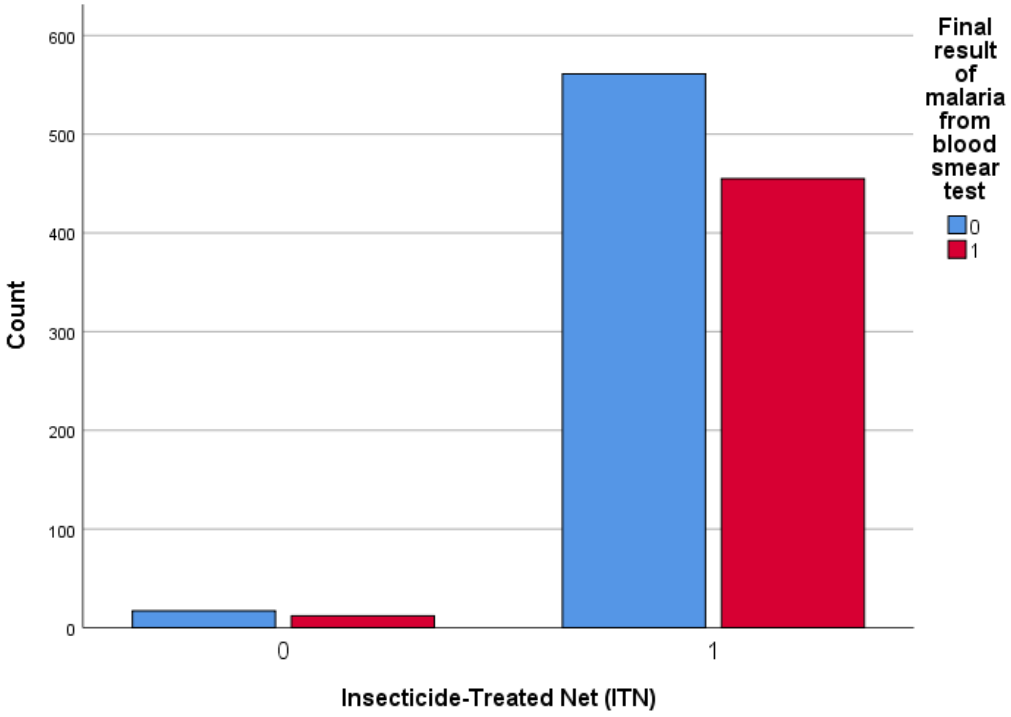

Count
300
200
100
0
1
2
Sex of the Child
Final result of malaria from blood smear test
0
1

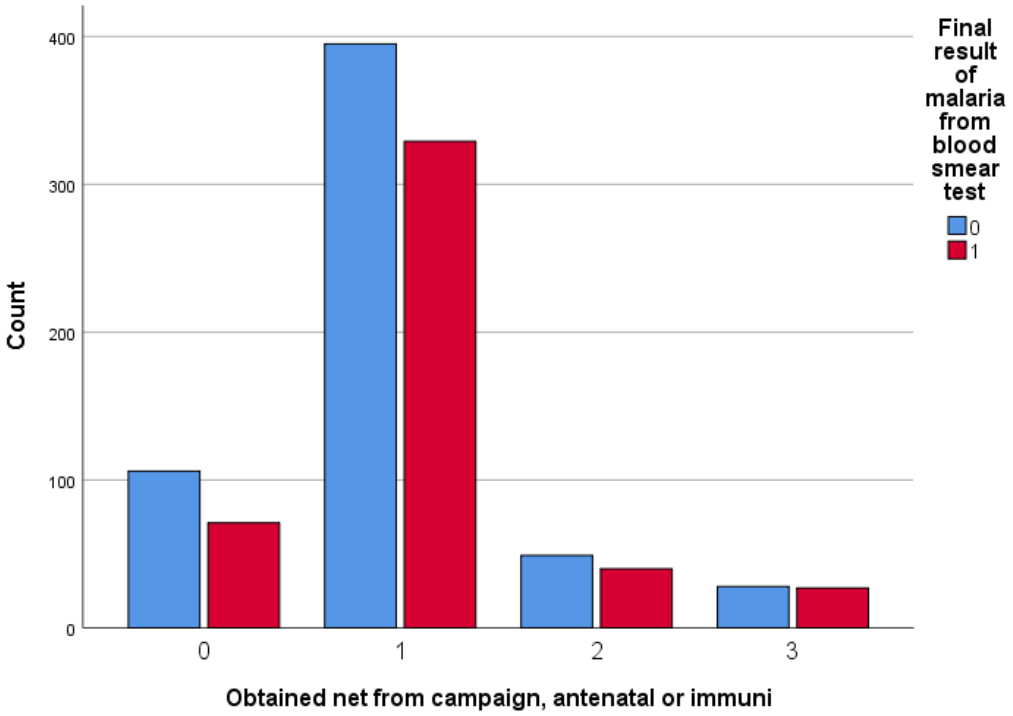

Boxplot grouped by Final result of malaria from blood smear test
Hemoglobin level (g/dl - 1 decimal)
160
140
120
100
80
60
40
20
0.0
1.0
Final result of malaria from blood smear test

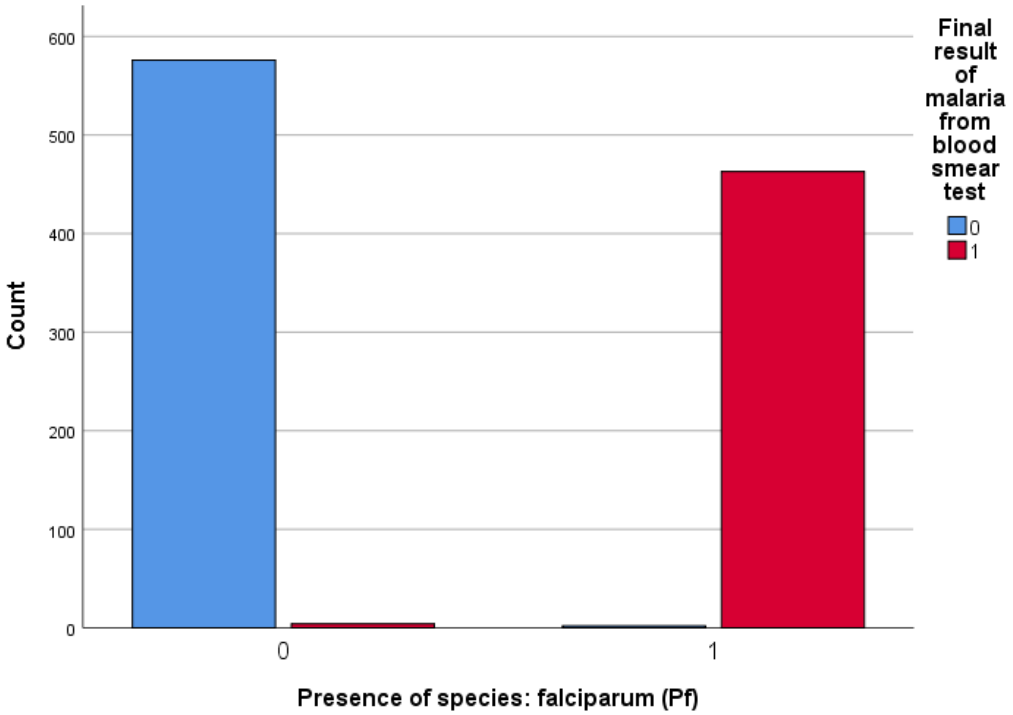

Final result of malaria from blood smear test
0
1
Count
400
300
200
100
0
1
2
3
4
Anemia level

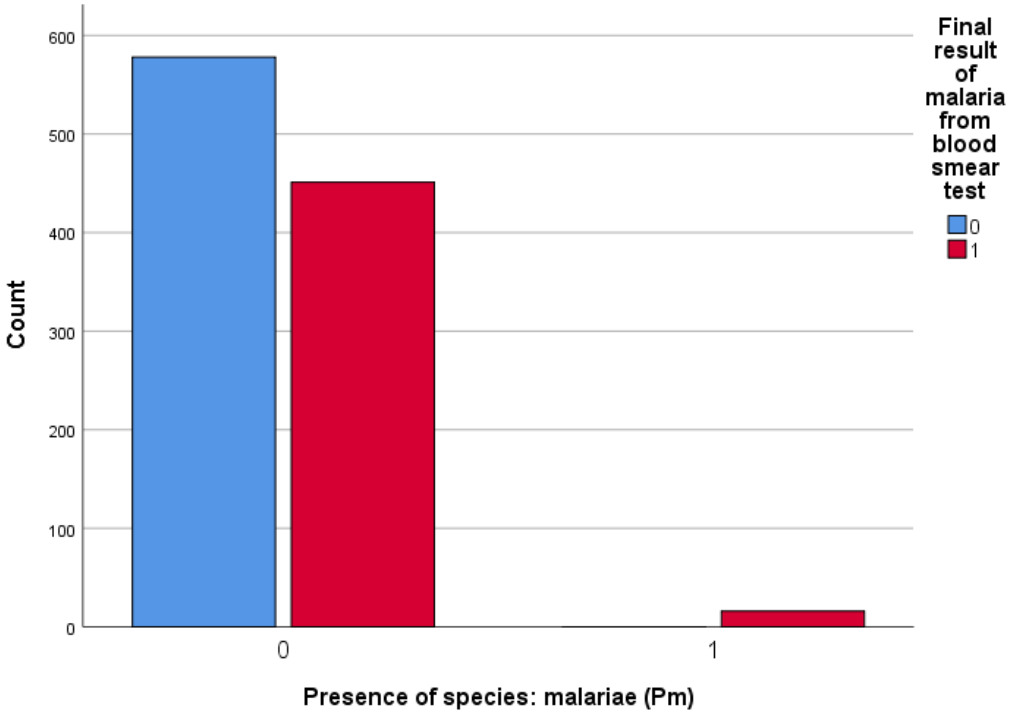

Final result of malaria from blood smear test
0
1
Count
400
300
200
100
0
0
1
2
3
Mother's highest educational level

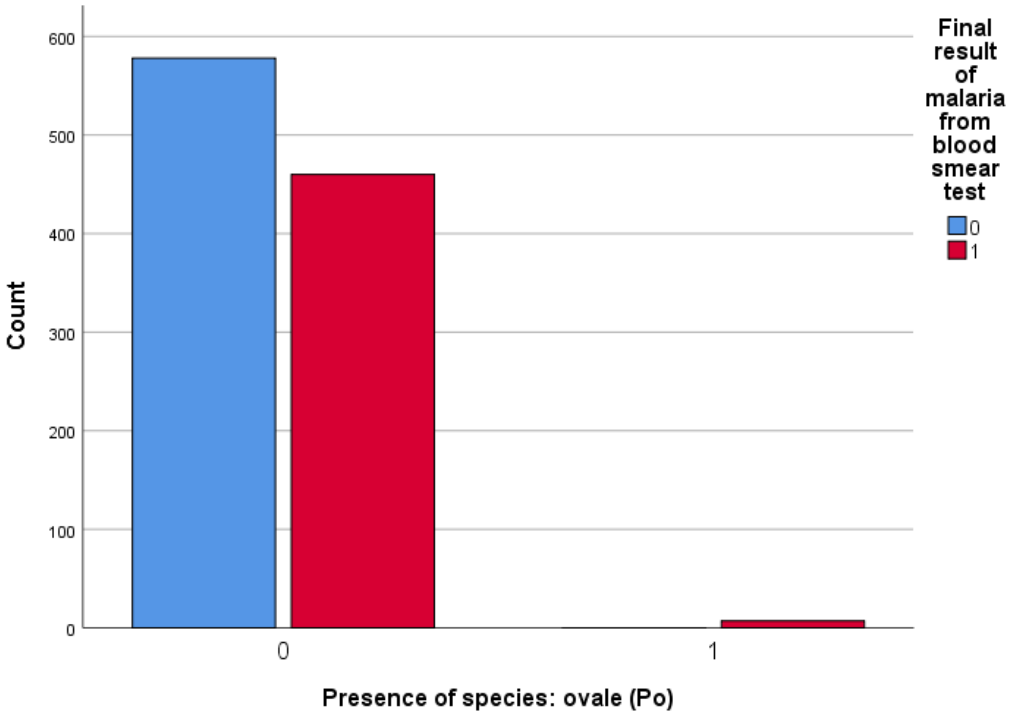

Count
500
400
300
200
100
0
0
1
Net observed by interviewer
Final result of malaria from blood smear test
0
1

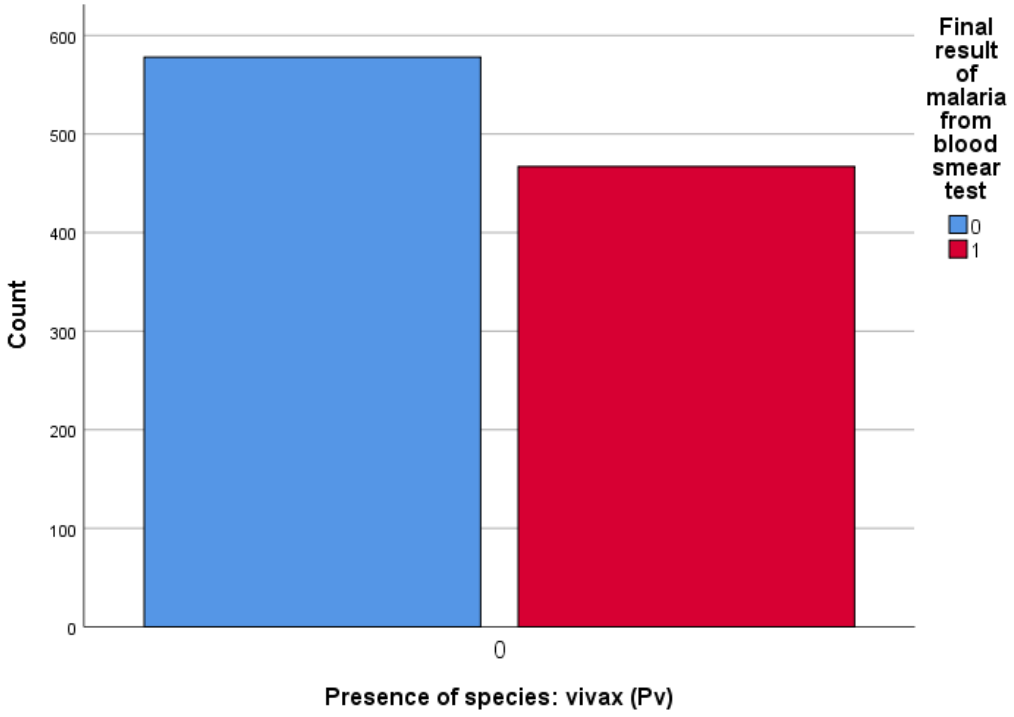

Count
600
500
400
300
200
100
0
0
1
Insecticide-Treated Net (ITN)
Final result of malaria from blood smear test
0
1

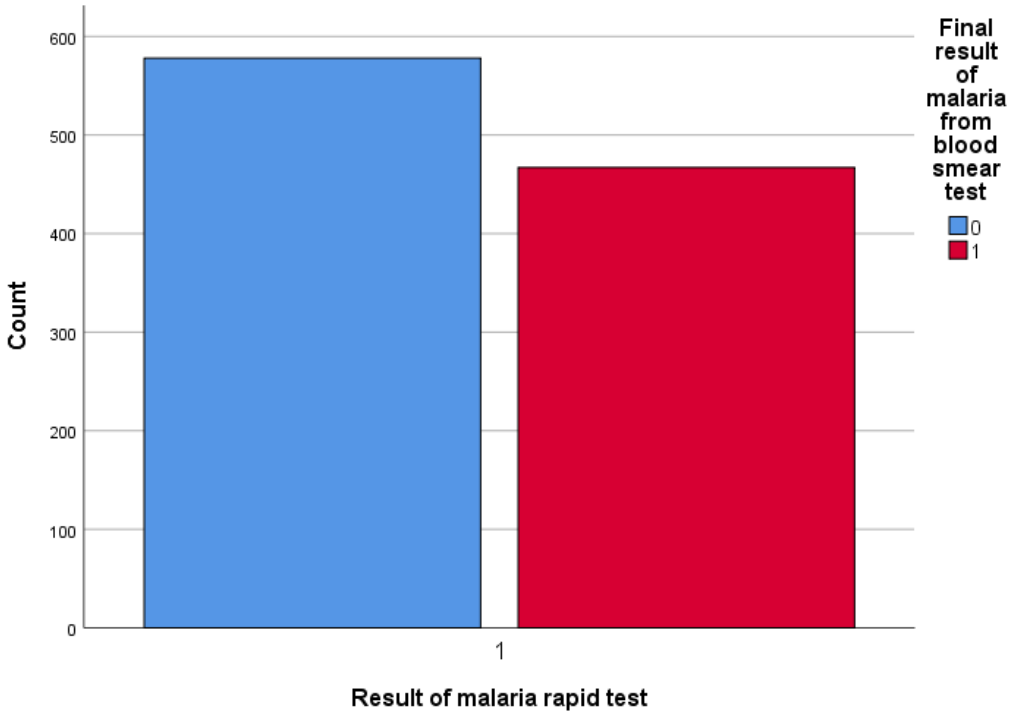

Count
400
300
200
100
0
0
1
2
3
Obtained net from campaign, antenatal or immuni
Final result of malaria from blood smear test
0
1

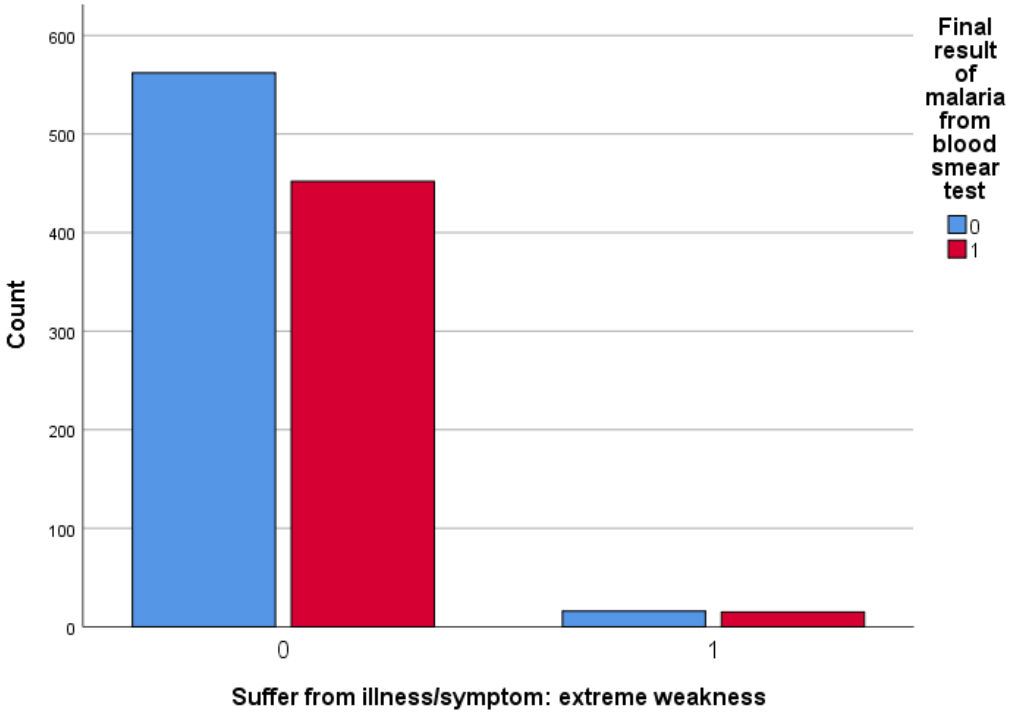

Count
600
500
400
300
200
100
0
0
1
Presence of species: falciparum (Pf)
Final result of malaria from blood smear test
0
1

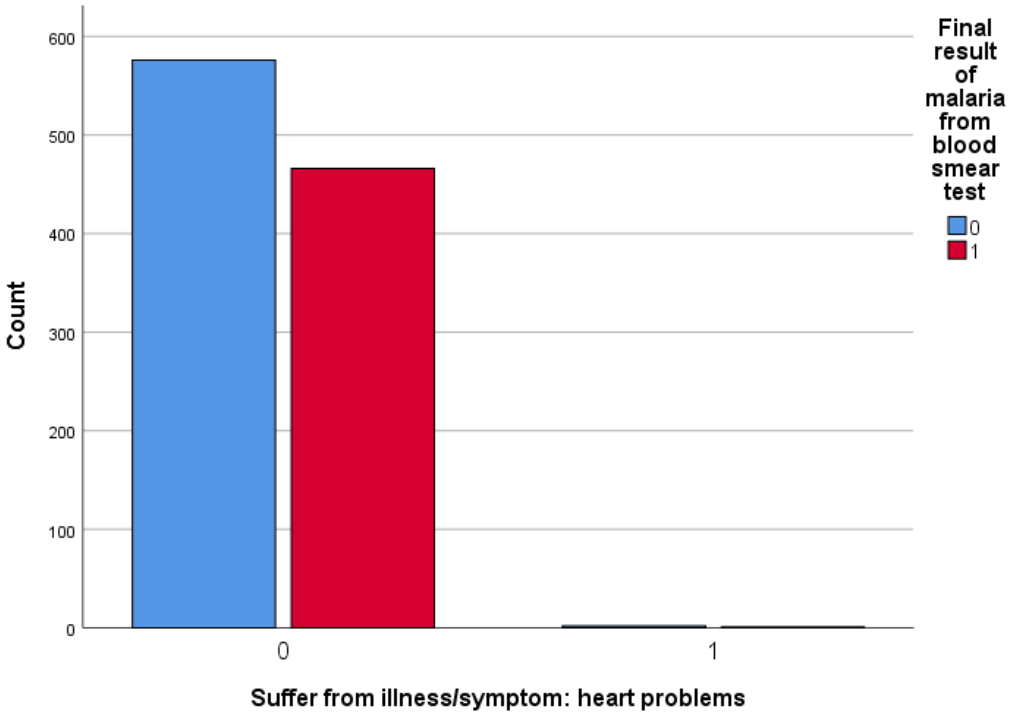

Count
600
500
400
300
200
100
0
0
1
Presence of species: malariae (Pm)
Final result of malaria from blood smear test
0
1

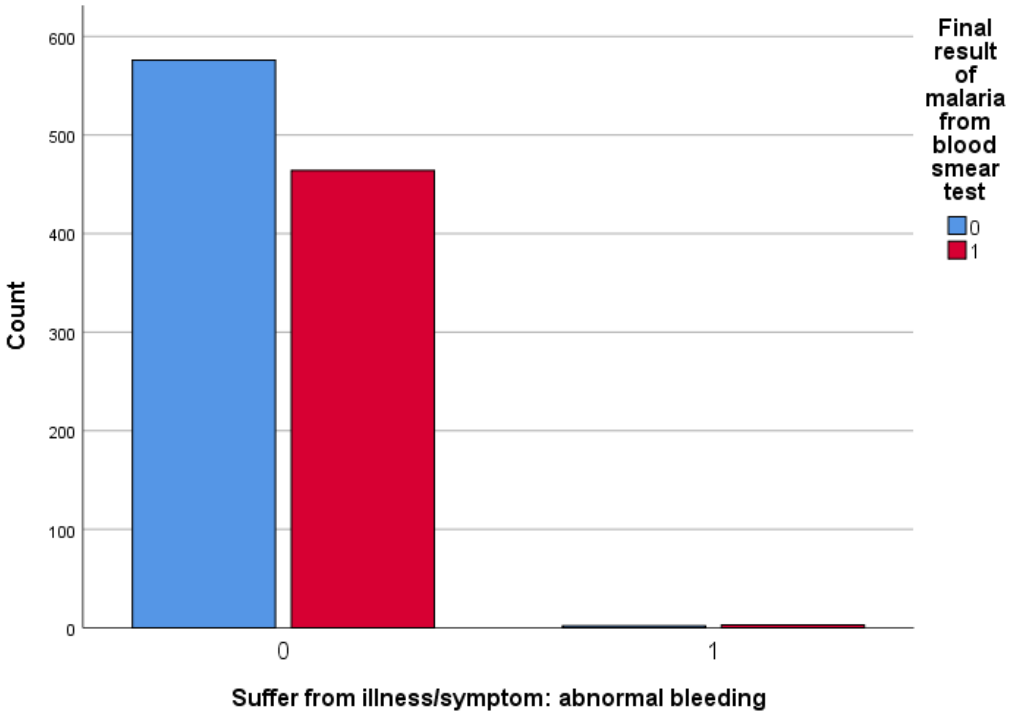

Count
600
500
400
300
200
100
0
0
1
Presence of species: ovale (Po)
Final result of malaria from blood smear test
0
1

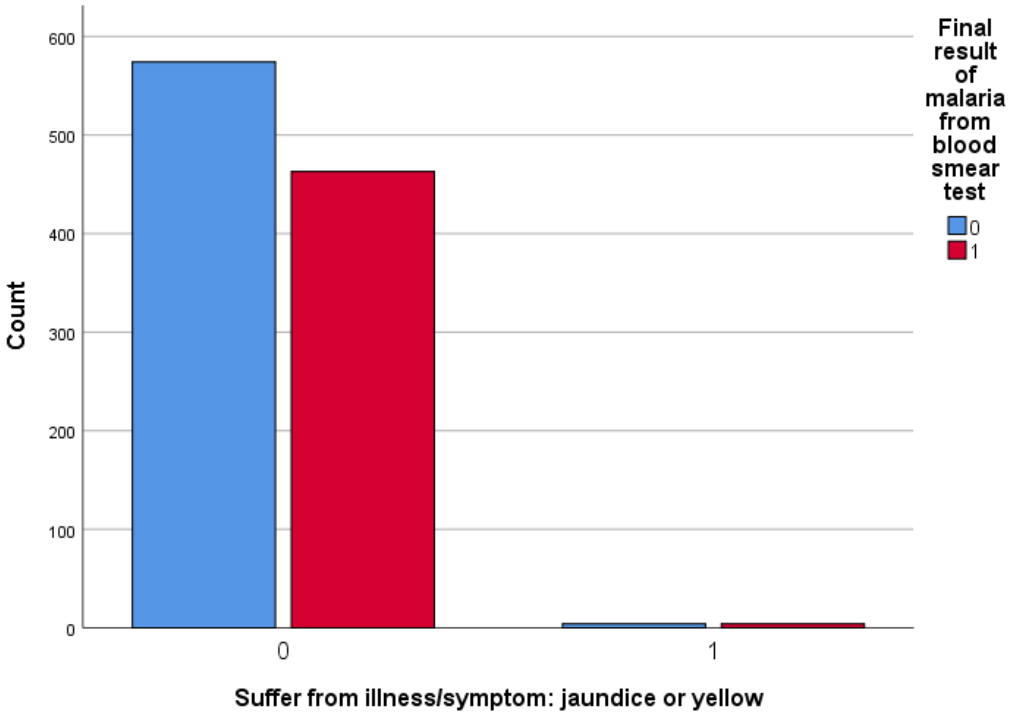

Count
600
500
400
300
200
100
0
0
Presence of species: vivax (Pv)
Final result of malaria from blood smear test
0
1

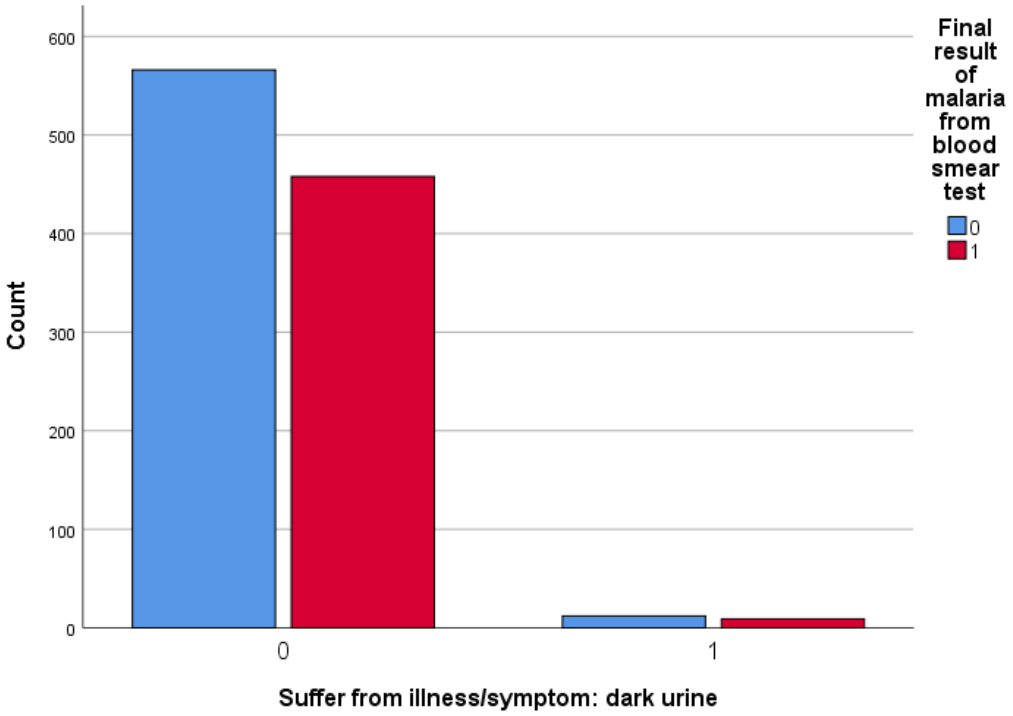

Count
600
500
400
300
200
100
0
1
Result of malaria rapid test
Final result of malaria from blood smear test
0
1

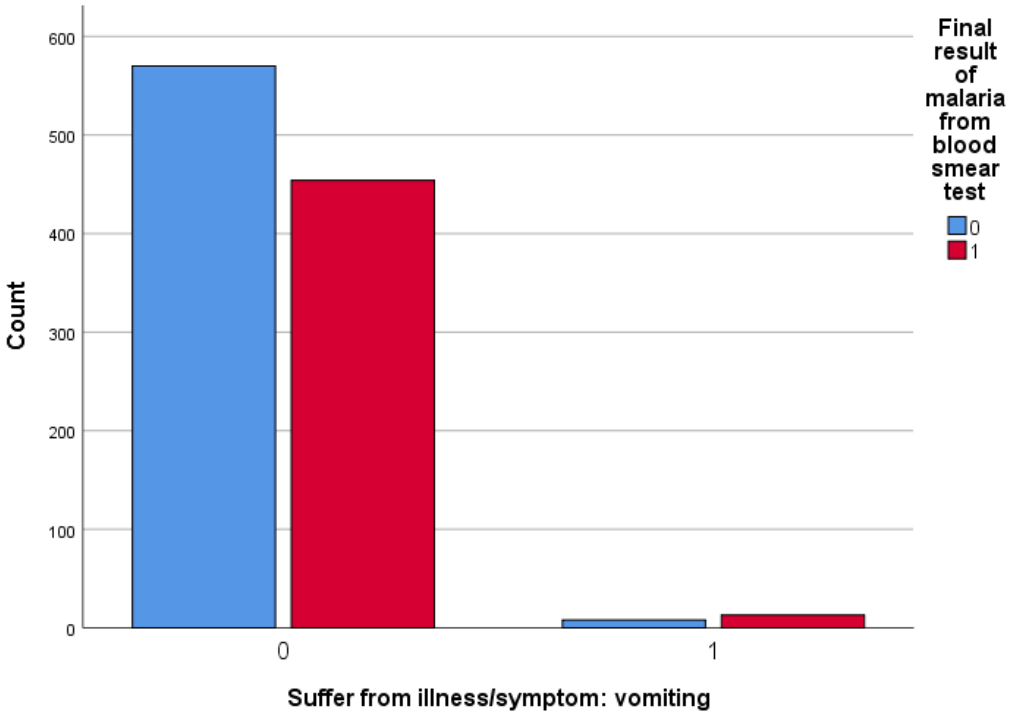

Count
600
500
400
300
200
100
0
0
1
Suffer from illness/symptom: extreme weakness
Final result of malaria from blood smear test
0
1

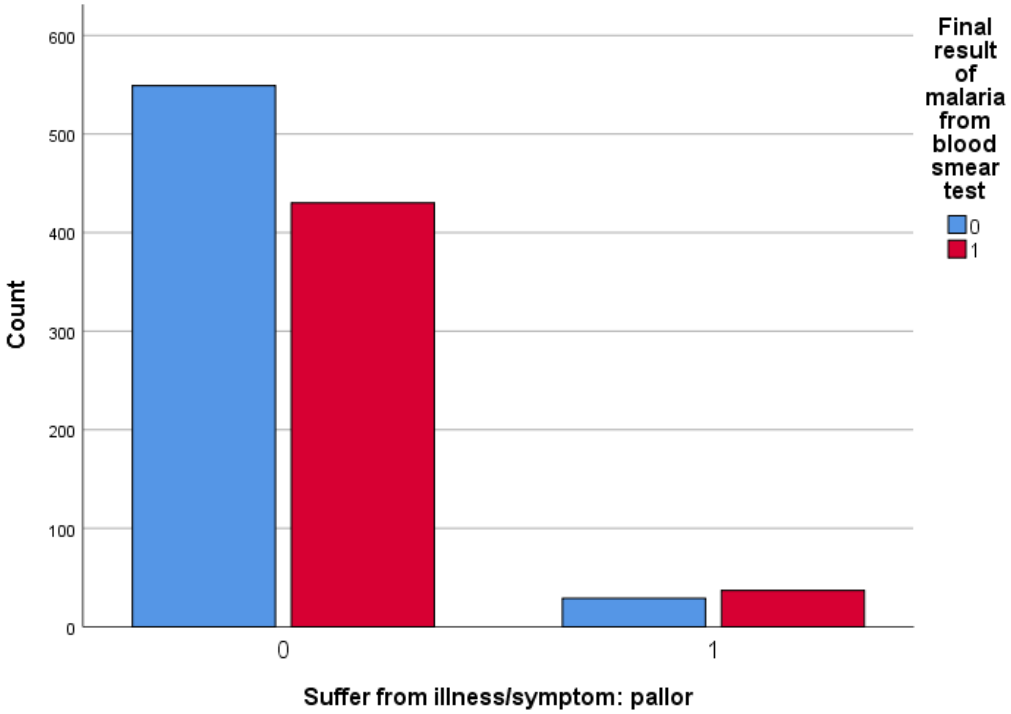

Count
600
500
400
300
200
100
0
0
1
Suffer from illness/symptom: heart problems
Final result of malaria from blood smear test
0
1

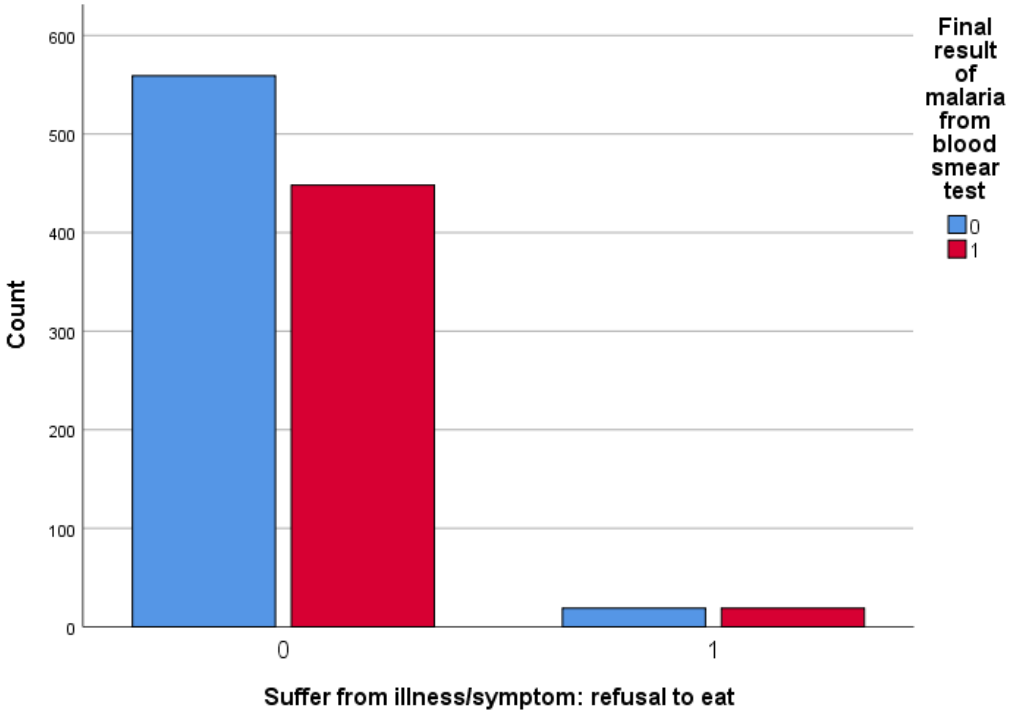

Count
600
500
400
300
200
100
0
0
1
Suffer from illness/symptom: abnormal bleeding
Final result of malaria from blood smear test
0
1

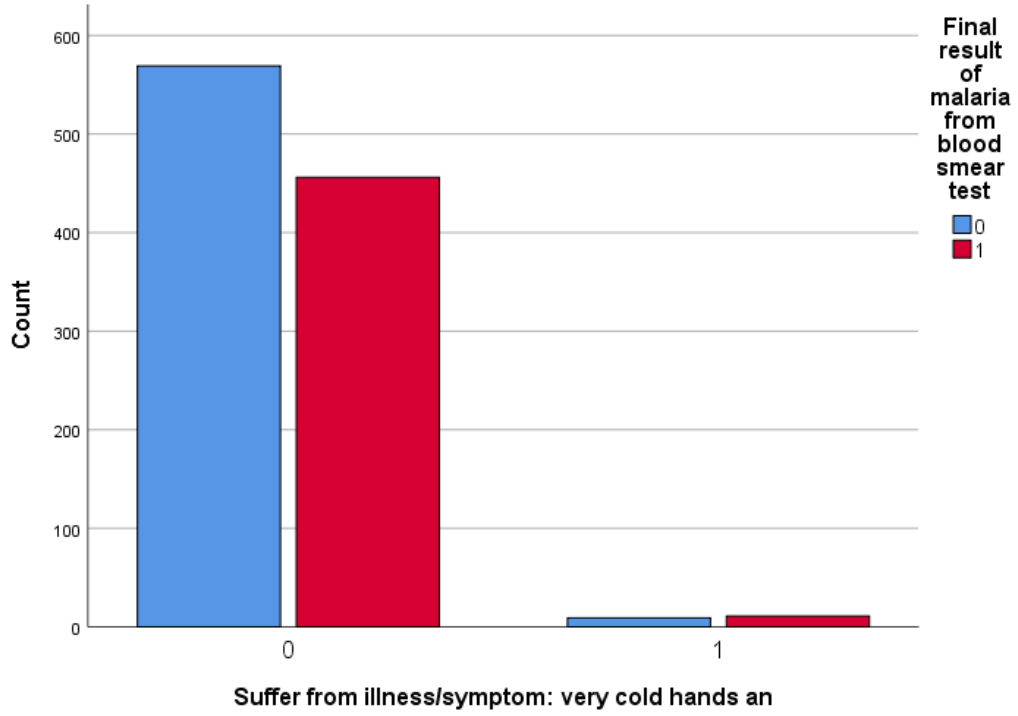

Count
600
500
400
300
200
100
0
0
1
Suffer from illness/symptom: jaundice or yellow
Final result of malaria from blood smear test
0
1


Count
600
500
400
300
200
100
0
0
1
Suffer from illness/symptom: dark urine
Final result of malaria from blood smear test
0
1


Count
600
500
400
300
200
100
0
0
1
Suffer from illness/symptom: vomiting
Final result of malaria from blood smear test
0
1


Count
600
500
400
300
200
100
0
0
1
Suffer from illness/symptom: pallor
Final result of malaria from blood smear test
0
1


Count
600
500
400
300
200
100
0
0
1
Suffer from illness/symptom: refusal to eat
Final result of malaria from blood smear test
0
1



4.4 Correlation Matrix

From section 4.2 distribution of features we found that features 'Has mosquito bed net for sleeping', "Presence of species: vivax (Pv)", "Result of malaria rapid test" have only single values which will not play any role in our further analysis. At this stage we will eliminate there three features.

Table.4 Correlation Matrix: Chi-Square Test for Categorical vs Categorical and Kruskal-Wallis H (K-H) Test for Numerical vs Categorical features

SL. No.	Features	Chi-square/ K-H	p-value	dof	Dtype	Decision
1	Number of household members	1620.872780	0.000000e+00	1.0	int64	Yes

SL. No.	Features	Chi-square/ K-H	p-value	dof	Dtype	Decision
2	SubRegion	62.151036	3.153244e-03	35.0	category	Yes
3	Type of place of residence	10.339342	1.302244e-03	1.0	category	Yes
4	Native language of respondent	17.321305	1.673913e-03	4.0	category	Yes
5	Source of drinking water	21.469956	2.881652e-02	11.0	category	Yes
6	Time to get to water source (minutes)	1591.066796	0.000000e+00	1.0	int64	Yes
7	Type of toilet facility	17.943525	5.592380e-02	10.0	category	No
8	Share toilet with other households	2.660290	1.028820e-01	1.0	category	No
9	Children under 5 slept under mosquito bed net	1.278780	5.276140e-01	2.0	category	No
10	Location of source for water	0.687799	7.090000e-01	2.0	category	No

SL. No.	Features	Chi-square/ K-H	p-value	dof	Dtype	Decision
11	Location of toilet facility	1.326406	5.151985e-01	2.0	category	No
12	Owns land usable for agriculture	0.015100	9.022006e-01	1.0	category	No
13	Owns livestock, herds or farm animals	5.555287	1.842495e-02	1.0	category	Yes
14	Owns pigs	445.052429	8.607102e-99	1.0	int64	Yes
15	Wealth index combined	10.275718	3.603073e-02	4.0	category	Yes
16	Region	18.924853	1.985076e-03	5.0	category	Yes
17	Child's age in days	1619.058802	0.000000e+00	1.0	int64	Yes
18	Sex of the Child	0.017852	8.937107e-01	1.0	category	No
19	Hemoglobin level (g/dl - 1 decimal)	1619.153732	0.000000e+00	1.0	float64	Yes
20	Anemia level	38.083649	2.713533e-08	3.0	category	Yes

SL. No.	Features	Chi-square/ K-H	p-value	dof	Dtype	Decision
21	Mother's highest educational level	10.695928	1.348904e-02	3.0	category	Yes
22	Net observed by interviewer	1.680885	1.948069e-01	1.0	category	No
23	Insecticide-Treated Net (ITN)	0.030337	8.617287e-01	1.0	category	No
24	Obtained net from campaign, antenatal or immune	2.099025	5.521101e-01	3.0	category	No
25	Final result of malaria from blood smear test	1040.958270	2.246937e-228	1.0	category	Yes
26	Presence of species: falciparum (Pf)	1016.876105	3.855134e-223	1.0	category	Yes
27	Presence of species: malariae (Pm)	17.902631	2.324995e-05	1.0	category	Yes
28	Presence of species: ovale (Po)	6.614922	1.011278e-02	1.0	category	Yes

SL. No.	Features	Chi-square/ K-H	p-value	dof	Dtype	Decision
29	Suffer from illness/symptom: extreme weakness	0.056198	8.126090e-01	1.0	category	No
30	Suffer from illness/symptom: heart problems	0.000000	1.000000e+00	1.0	category	No
31	Suffer from illness/symptom: abnormal bleeding	0.057332	8.107642e-01	1.0	category	No
32	Suffer from illness/symptom: jaundice or yellow	0.000000	1.000000e+00	1.0	category	No
33	Suffer from illness/symptom: dark urine	0.000000	1.000000e+00	1.0	category	No
34	Suffer from illness/symptom: vomiting	1.908041	1.671810e-01	1.0	category	No

SL. No.	Features	Chi-square/ K-H	p-value	dof	Dtype	Decision
35	Suffer from illness/symptom: pallor	3.210897	7.314935e-02	1.0	category	No
36	Suffer from illness/symptom: refusal to eat	0.254647	6.138218e-01	1.0	category	No
37	Suffer from illness/symptom: very cold hands an	0.503297	4.780551e-01	1.0	category	No

After eliminating the features with p-value < 0.05, features on Table.5 will be used to implement ML algorithms

Table.5 Features having p-value < 0.05

SL. No.	Features	Chi-Square/ K-H	p-value	Dof	Dtype	Decision
1	Number of household members	1620.872780	0.000000e+00	1.0	int64	True
2	SubRegion	62.151036	3.153244e-03	35.0	category	True
3	Region	18.924853	1.985076e-03	5.0	category	True

SL. No.	Features	Chi-Square/ K-H	p-value	Dof	Dtype	Decision
4	Type of place of residence	10.339342	1.302244e-03	1.0	category	True
5	Native language of respondent	17.321305	1.673913e-03	4.0	category	True
6	Source of drinking water	21.469956	2.881652e-02	11.0	category	True
7	Time to get to water source (minutes)	1591.066796	0.000000e+00	1.0	int64	True
8	Owns livestock, herds or farm animals	5.555287	1.842495e-02	1.0	category	True
9	Owns pigs	445.052429	8.607102e-99	1.0	int64	True
10	Wealth index combined	10.275718	3.603073e-02	4.0	category	True
11	Child's age in days	1619.058802	0.000000e+00	1.0	int64	True
12	Hemoglobin level (g/dl - 1 decimal)	1619.153732	0.000000e+00	1.0	float64	True

SL. No.	Features	Chi-Square/ K-H	p-value	Dof	Dtype	Decision
13	Anemia level	38.083649	2.713533e-08	3.0	category	True
14	Mother's highest educational level	10.695928	1.348904e-02	3.0	category	True
15	Final result of malaria from blood smear test	1040.958270	2.246937e-228	1.0	category	True
16	Presence of species: falciparum (Pf)	1016.876105	3.855134e-223	1.0	category	True
17	Presence of species: malariae (Pm)	17.902631	2.324995e-05	1.0	category	True
18	Presence of species: ovale (Po)	6.614922	1.011278e-02	1.0	category	True

5 Analytical Results

5.1.1 Model Implementation

The dataset used in this study has dimensionality of 70428 observations and 225 features. After cleaning, filtering, preprocessing and statistical analysis (discussed in Section 4) we got high quality clean dataset (DS1: 1045 observations and 17 predictor features and one target feature). To enlarge the dataset, we have implemented Random Oversampling (ROS) model which enlarged the dataset.

In original dataset (DS1) the percentage of malaria positive/negative ratio is 55% negative & 45% negative. For ROS we also followed this positive/negative ratio of DS1. The newly generated ROS dataset (DS2) has 5500 negative observations and 4500 positive observations.

In this study two traditional machine learning model Logistic Regression (LR) & Decision Tree (DT) and one AutoML framework H2O AutoML were implemented on both datasets DS1 and DS2. For training, testing and validation datasets were split as 80%, 10% and 10% respectively.

5.1.2 Model Diagnostics

Random Oversampling

Random Oversampling (ROS) [28] was done with the help of *imblearn* a python package. Here *RandomOverSampler()* method was called with sampling strategy (0 = 5500 and 1 = 4500). ROS uses Random Forest algorithm in its backed to enlarge the existing dataset.

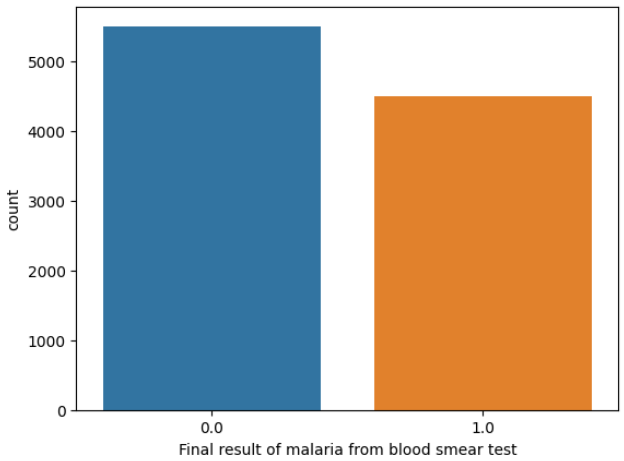


Fig.5.1 Random Oversampled Dataset (DS2) target column value count

Logistic Regression

Firstly Logistic Regression (LR) was implemented as ML model. In this study hyperparameter tuning was done. Here regularization (penalty) *"l2"* was set with solver *"newton-cholesky"* and max iteration *"1000"*. 10-fold cross-validation was done.

Decision Tree

The second model was Decision Tree (DT). Initially the DT model was trained with default parameters to find the optimum pruning with cost complexity pruning method. By this we found the below graph.

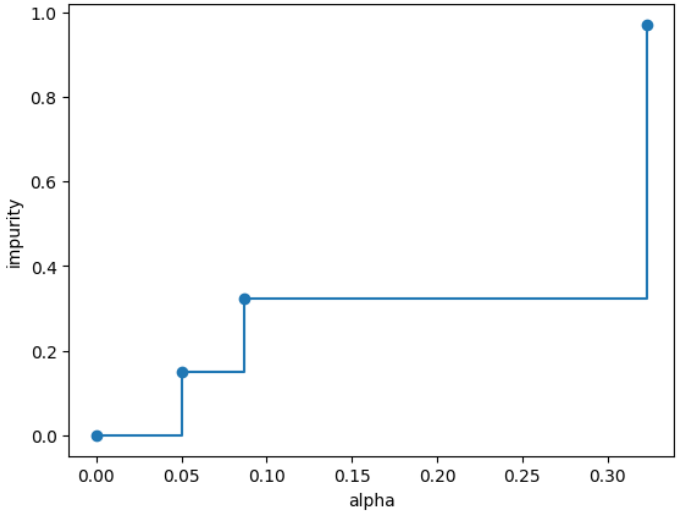


Fig.5.2 impurity vs alpha (default hyperparameter)

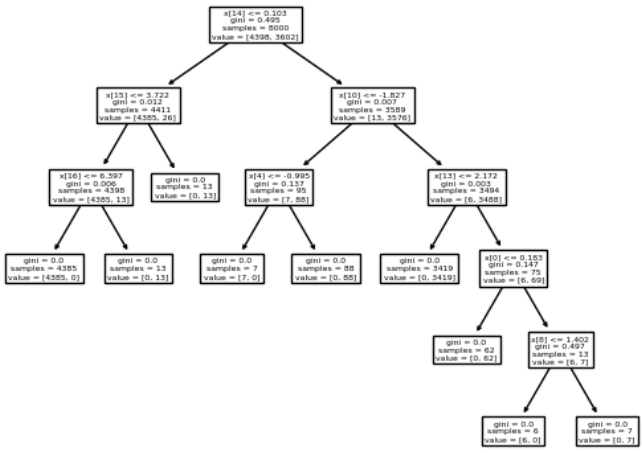


Fig.5.3 Tree expansion with default hyperparameters (overfitted tree)

Then DT was implemented with hyperparameter max depth *"4"*, method used for maximum feature used for each tree was *"sqrt"* (square root of total features), splitter was set to *"random"* and pruning parameter alpha *"0.05"* from the figure "Fig: impurity vs alpha (pruning parameter)". And also, cross-validation was done with 10-fold

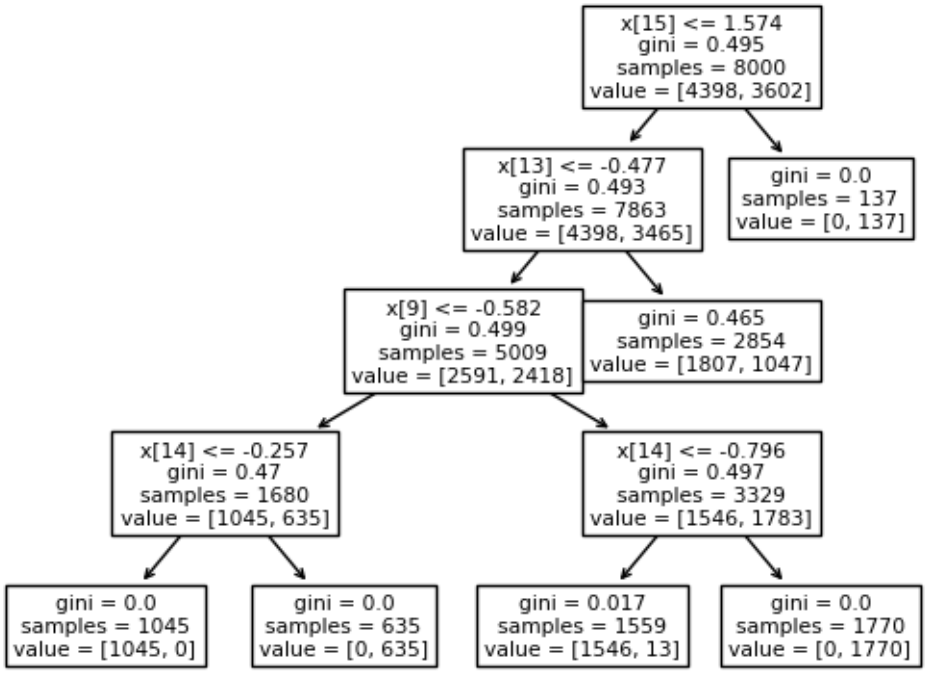


Fig.5.4 Tree expansion after fine tuning hyperparameters

H2O AutoML

In H2O AutoML there are hundreds of traditional, modified and new algorithms including *deep learning* models. Due to computational and time limitation we restricted the model with *max_model = 5* which means the model will give best 5 model from its framework after training and set cross-validation as 10-fold. The best part of AutoML is that, it doesn't need any hyperparameter tuning or cleaning the dataset. By this advantages AutoML dominates the traditional ML algorithms. It also doesn't need any ML expertise to operate. It opens a new door

for those who wants to take advantages of predictive analysis (AI and ML) but don't have expertise on ML or AI like, Health experts.

After training the H2O AutoML had given the best model. The name of the model for DS1 was Generalized Linear Model (GLM) and for DS2 was *Stacked Ensemble*. *Stacked Ensemble* model follows a stacking strategy which is given below

```
GLM Model: summary
    family    link    regularization
--  --------  ------  --------------------------
    binomial  logit   Ridge ( lambda = 5.341E-4 )

  lambda_search
  ------------------------------------------------------------------------------
  nlambda = 30, lambda.max = 49.331, lambda.min = 5.341E-4, lambda.1se = 0.00223

number_of_predictors_total    number_of_active_predictors
----------------------------  ----------------------------
17                            17

number_of_iterations    training_frame
----------------------  ------------------------------------------------
62                      AutoML_2_20230405_144824_training_py_21_sid_bf47
```

Fig.5.5 the optimum hyperparameters tuned by H2O AutoML in GLM for DS1

```
Model Key: StackedEnsemble_BestOfFamily_1_AutoML_4_20230405_153849


Model Summary for Stacked Ensemble:
key                                  value
-----------------------------------  ----------------
Stacking strategy                    cross_validation
Number of base models (used / total)  2/4
# GBM base models (used / total)      1/1
# XGBoost base models (used / total)  0/1
# GLM base models (used / total)      0/1
# DRF base models (used / total)      1/1
Metalearner algorithm                GLM
Metalearner fold assignment scheme    Random
Metalearner nfolds                   10
Metalearner fold_column
Custom metalearner hyperparameters    None
```

Fig.5.6 Stacking strategy (by H2O AutoML) of the best model "*Stacked Ensemble"* for DS2

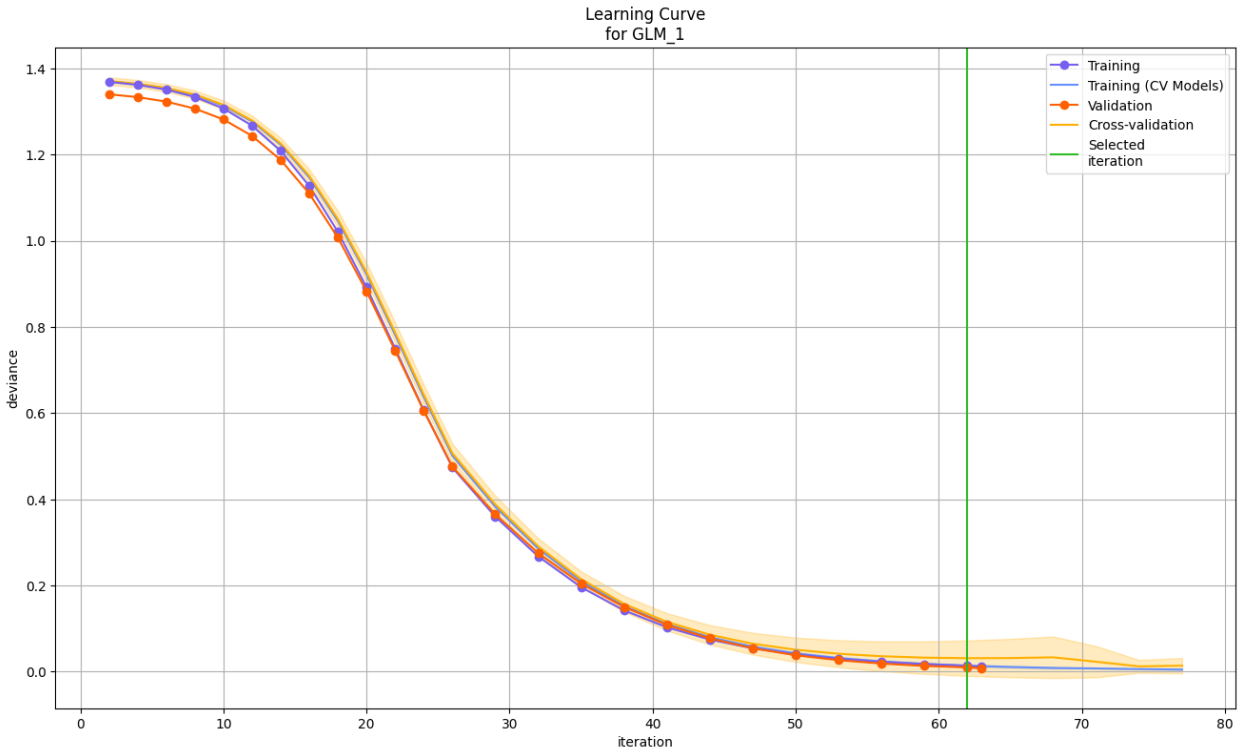


Fig.5.7 Learning Curve of GLM for DS1

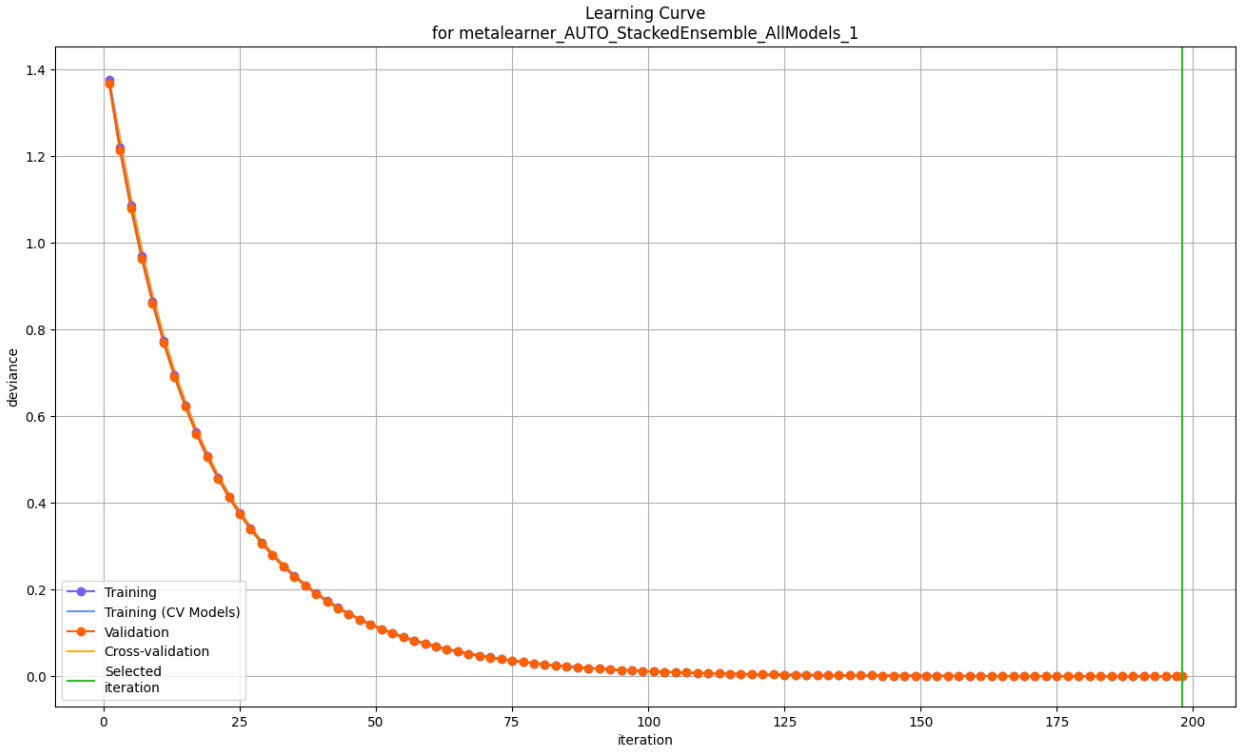


Fig.5.8 Learning Curve of Stacked Ensemble for DS2

5.2 Model Performance

Performance measures for all the ML models are given in below tables

Table.5 Performance measures of Logistic Regression (LR) for DS1

LR	MAE	RMSE	R2	Precision	Recall	f1_score	Accuracy
Train	0.007177	0.084717	0.971112	0.994778	0.989610	0.992188	0.992823
Test	0.000000	0.000000	1.000000	1.000000	1.000000	1.000000	1.000000
Val	0.000000	0.000000	1.000000	1.000000	1.000000	1.000000	1.000000
Average	0.002392	0.028239	0.990371	0.998259	0.996537	0.997396	0.997608

Table.6 Performance measures of Decision Tree (DT) for DS1

DT	MAE	RMSE	R2	Precision	Recall	f1_score	Accuracy
Train	0.138756	0.372500	0.441495	1.000000	0.698701	0.822630	0.861244
Test	0.153846	0.392232	0.365612	1.000000	0.627907	0.771429	0.846154
Val	0.270000	0.490118	-0.158333	0.750000	0.300000	0.400000	0.730000
Average	0.187534	0.418283	0.216258	0.916667	0.542203	0.664686	0.812466

Table.7 Performance measures of H2O AutoML Generalized Linear Model (H2O_GLM) for DS1

H2O_GLM	MAE	RMSE	R2	Precision	Recall	f1_score	Accuracy
Train	0.000968	0.031109	0.996105	1.000000	1.000000	0.998689	0.998791
Test	0.000111	0.010544	0.999522	1.000000	1.000000	1.000000	1.000000
Val	0.002511	0.050115	0.989892	0.997582	1.000000	0.997375	0.997582
Average	0.001197	0.030589	0.995173	0.999194	1.000000	0.998688	0.998791

Table.8 Average of Table1, Table.2 and Table.3 (DS1)

Model	MAE	RMSE	R2	Precision	Recall	f1_score	accuracy
LR	0.0023923	0.0282390	0.99037067	0.9982593	0.9965367	0.9973960	0.9976076
DT	0.1875340	0.4182833	0.21625800	0.9166666	0.5422026	0.6646863	0.8124660
H2O_GLM	0.0011968	0.0305893	0.99517269	0.9991940	1.0000000	0.9986880	0.9987910

Table.9 Performance measures of LR for DS2

LR	MAE	RMSE	R2	Precision	Recall	f1_score	accuracy
Train	0.0016250	0.040311	0.9934350	0.9964040	1.0000000	0.998199	0.998375
Test	0.0020000	0.044721	0.9919540	0.9956900	1.0000000	0.997840	0.998000
Val	0.0030000	0.030000	0.9878030	0.9932830	1.0000000	0.996603	0.997000
average	0.0022083	0.038344	0.9910640	0.9951257	1.0000000	0.997547	0.997791

Table.10 Performance measures of DT for DS2

DT	MAE	RMSE	R2	Precision	Recall	f1_score	accuracy
Train	0.132500	0.364005	0.464700	1.000000	0.705719	0.827474	0.867500
Test	0.143000	0.378153	0.424677	1.000000	0.690476	0.816901	0.857000
Val	0.119000	0.315056	0.515560	0.897222	0.728171	0.800796	0.881000
average	0.131500	0.352405	0.468312	0.965741	0.708122	0.815057	0.868500

Table.11 Performance measures of H2O AutoML Stacked Ensemble (H2O_SE) for DS2

H2O_SE	MAE	RMSE	R2	Precision	Recall	f1_score	accuracy
Train	0.000000	0.000081	1.000000	1.000000	1.000000	1.000000	1.000000
Test	0.000000	0.000098	1.000000	1.000000	1.000000	1.000000	1.000000
Val	0.000000	0.000109	1.000000	1.000000	1.000000	1.000000	1.000000
average	0.000000	0.000096	1.000000	1.000000	1.000000	1.000000	1.000000

Table.12 Average of Table.5, Table.6 and Table.7

Model	MAE	RMSE	R2	Precision	Recall	f1_score	accuracy
LR	0.002208	0.038344	0.991064	0.995126	1.000000	0.997547	0.997792
DT	0.131500	0.352405	0.468312	0.965741	0.708122	0.815057	0.868500
H2O_SE	0.000000	0.000096	1.000000	0.999939	0.999267	0.999267	0.999267

5.3 Model Result

From Table.8 we can summarize result for DS1. In Table.4 we can see that GLM of H2O AutoML outperformed over LR and DT in all performance measures MAE, RMSE,R2, Precision, Recall, f1_score and accuracy with value 0.0011968, 0.0305893, 0.99517269, 0.9991940, 1.0000000, 0.9986880 and 0.9987910 respectively.

Also, from Table.12 we can recommend that Stacked Ensemble of AutoML performed better than other two models in all performance measure MAE, RMSE,R2, Precision, Recall, f1_score and accuracy with value 0.000000, 0.000096, 1.000000, 0.999939, 0.999267, 0.999267 and 0.999267 respectively.

6 Prescriptive Insight

In this study, initially 43 features were selected from 225 features of original dataset. After cleaning, filtering and some statistical analysis we ended up with 18 features including target feature. As this is a predictive study we implemented some models like Logistic Regression, Decision Tree and H2O AutoML. Among them Generalized Linear Model (GLM) and Stacked Ensemble model of H2O AutoML outperformed over other two model on DS1 and DS2 respectively. Now GLM model has the capability to provide feature importance based on its learning. A feature importance graph is given below based on GLM model where it shows top ten important features.

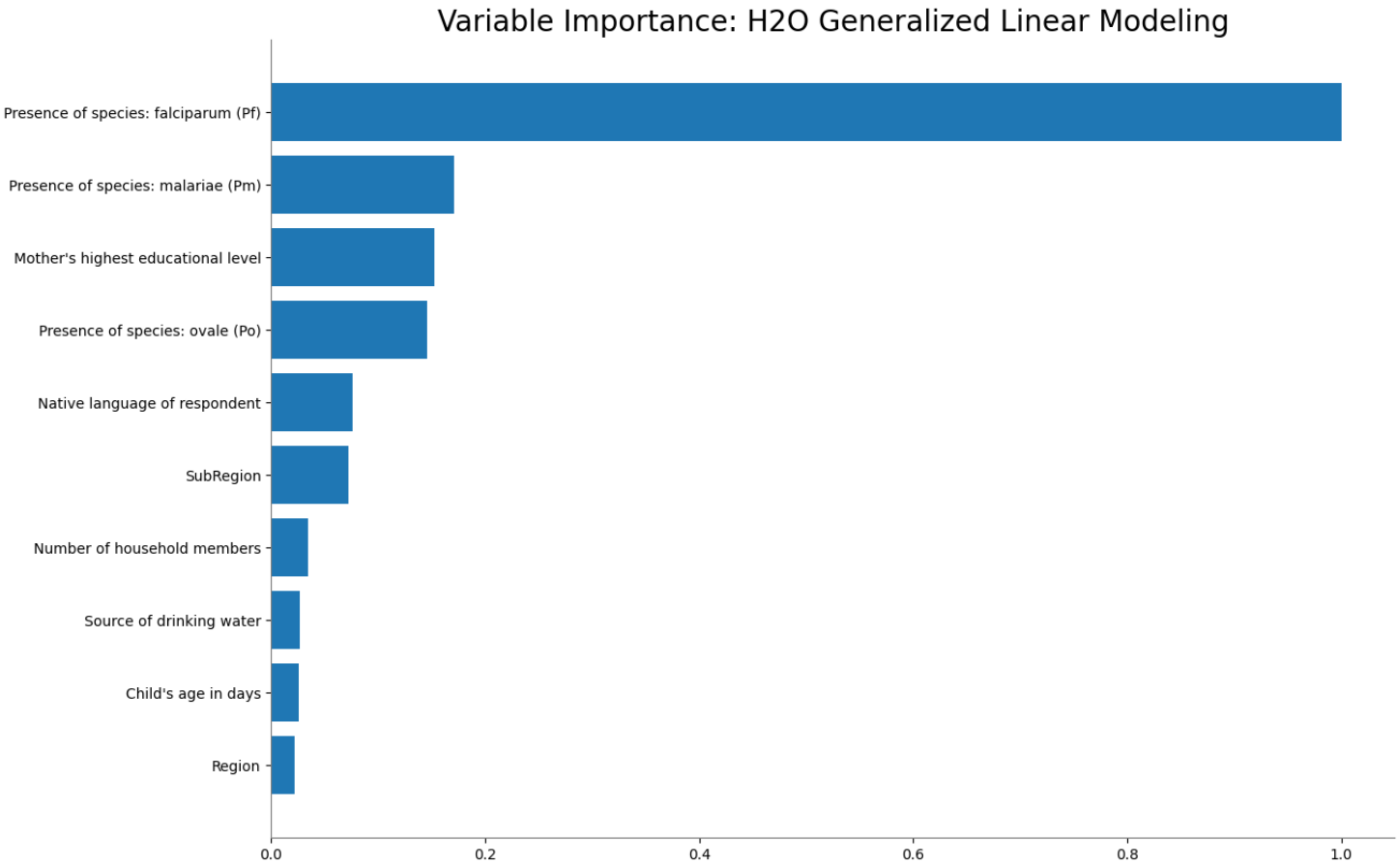


Fig.6.1 Features Importance Graph based on GLM Model

Variable Importances:

variable	relative_importance	scaled_importance	percentage
Presence of species: falciparum (Pf)	5.69101	1	0.551691
Presence of species: malariae (Pm)	0.97494	0.171312	0.0945115
Mother's highest educational level	0.870094	0.152889	0.0843476
Presence of species: ovale (Po)	0.829361	0.145732	0.0803989
Native language of respondent	0.434028	0.0762655	0.042075
SubRegion	0.413891	0.0727272	0.040123
Number of household members	0.197473	0.0346992	0.0191432
Source of drinking water	0.150973	0.0265284	0.0146355
Child's age in days	0.144652	0.0254176	0.0140227
Region	0.127021	0.0223196	0.0123135
Owns pigs	0.112229	0.0197203	0.0108795
Owns livestock, herds or farm animals	0.0977372	0.017174	0.00947473
Wealth index combined	0.0836229	0.0146939	0.00810647
Time to get to water source (minutes)	0.0540026	0.00948911	0.00523506
Type of place of residence	0.0538744	0.00946657	0.00522262
Anemia level	0.04436	0.00779475	0.0043003
Hemoglobin level (g/dl - 1 decimal)	0.0363023	0.00637889	0.00351918

Fig.6.2 Features Importance with value based on GLM Model

from Fig.6.2 we can find that Presence of species: falciparum (Pf), Presence of species: malariae (Pm), Mother's highest educational level and Presence of species: ovale (Po) are the most important features. We also come to know thar three features among top four are clinical factors and one is social factor.

7. Conclusion

7.1 Summary

Machine Learning (ML) approach in Health Care is booming. In this study an ML approach is taken to predict the Malaria result (positive/negative) based of Demographic Health Surveys 2021 of Nigeria dataset. Here two traditional ML algorithm Logistic Regression & Decision Tree and an AutoML framework H2O are implemented. H2O AutoML algorithms Generalized Linear Model and Stacked Ensemble performed better than other algorithms in all aspect of seven performance measures in DS1 and DS2 respectively. GLM shows top ten important features. Among Presence of species: falciparum (Pf), Presence of species: malariae (Pm), Mother's highest educational level and Presence of species: ovale (Po) are the top four features.

7.2 Limitations

During this study some limitations compromised the best result. The computational and time limitation were the main factors of compromised results. Beside these, the data uncertainty and model uncertainty with irreducible error are also the limitations. For the limitation of time and computational cost we restrict (limit) the H2O AutoML training with five model. If there were no limitations, we will be able to try more than 100 models to find the best model in H2O AutoML.

7.3 Future Research Direction

This study is limited to a historical dataset of 2021, one linear, one tress based and one AutoML model. Research with updated data, new features and new algorithms and frameworks like Deep Learning, Semi Supervised Learning, Transfer Learning models and AutoKeras, Auto-Sklearn, TPOT etc. AutoML frameworks can be implemented to get more meaningful and important insights from the datasets. Predictive analysis with AutoML frameworks should be done in a wider range as it is a user-friendly predictive tool and don't need expertise of Machine Learning, Deep learning or Artificial Intelligence.

Reference

[1] World Health Organization, "World Health Organization." https://www.who.int/news-room/fact-sheets/detail/malaria

[2] Centers for Disease Control and Prevention, "Centers for Disease Control and Prevention." https://www.cdc.gov/malaria/diagnosis_treatment/diagnosis.html

[3] MedlinePlus, "National Library of Medicine - National Institutes of Health", [Online]. Available: https://medlineplus.gov/lab-tests/malaria-tests/

[4] DHS, "The DHS Program Available Datasets", [Online]. Available: https://dhsprogram.com/data/available-datasets.cfm

[5] AutoML.org, "AutoML.org", [Online]. Available: https://www.automl.org/automl/

[6] G. Kalipe, V. Gautham, and R. K. Behera, "Predicting Malarial Outbreak using Machine Learning and Deep Learning Approach: A Review and Analysis," in *Proceedings - 2018*

International Conference on Information Technology, ICIT 2018, Dec. 2018, pp. 33–38. doi: 10.1109/ICIT.2018.00019.

[7] Y. W. Lee, J. W. Choi, and E.-H. Shin, "Machine learning model for predicting malaria using clinical information," *Comput. Biol. Med.*, vol. 129, p. 104151, Feb. 2021, doi: 10.1016/j.compbiomed.2020.104151.

[8] O. Nkiruka, R. Prasad, and O. Clement, "Prediction of malaria incidence using climate variability and machine learning," *Informatics Med. Unlocked*, vol. 22, Jan. 2021, doi: 10.1016/j.imu.2020.100508.

[9] S. S. Yadav, V. J. Kadam, S. M. Jadhav, S. Jagtap, and P. R. Pathak, "Machine learning based malaria prediction using clinical findings," in *2021 International Conference on Emerging Smart Computing and Informatics, ESCI 2021*, Mar. 2021, pp. 216–222. doi: 10.1109/ESCI50559.2021.9396850.

[10] M. O. Arowolo, "ICA LEARNING APPROACH FOR PREDICTING OF RNA-SEQ MALARIA VECTOR DATA CLASSIFICATION USING SVM KERNEL ALGORITHMS," 2022.

[11] G. Shekar, S. Revathy, and E. K. Goud, "Malaria Detection using Deep Learning," in *Proceedings of the 4th International Conference on Trends in Electronics and Informatics, ICOEI 2020*, Jun. 2020, pp. 746–750. doi: 10.1109/ICOEI48184.2020.9143023.

[12] M. Masud *et al.*, "Leveraging Deep Learning Techniques for Malaria Parasite Detection Using Mobile Application," *Wirel. Commun. Mob. Comput.*, vol. 2020, 2020, doi: 10.1155/2020/8895429.

[13] D. Harvey, W. Valkenburg, and A. Amara, "Predicting malaria epidemics in Burkina Faso with machine learning," *PLoS One*, vol. 16, no. 6 June, Jun. 2021, doi: 10.1371/journal.pone.0253302.

[14] M. K. Gourisaria, S. Das, R. Sharma, S. S. Rautaray, and M. Pandey, "A deep learning model for malaria disease detection and analysis using deep convolutional neural networks," *Int. J. Emerg. Technol.*, vol. 11, no. 2, pp. 699–704, 2020, [Online]. Available: www.researchtrend.net

[15] P. Mohapatra, N. K. Tripathi, I. Pal, and S. Shrestha, "Determining suitable machine

learning classifier technique for prediction of malaria incidents attributed to climate of Odisha," *Int. J. Environ. Health Res.*, vol. 32, no. 8, pp. 1716–1732, 2022, doi: 10.1080/09603123.2021.1905782.

[16] B. Muhammad and A. Varol, "A Symptom-Based Machine Learning Model for Malaria Diagnosis in Nigeria," in *2021 9th International Symposium on Digital Forensics and Security (ISDFS)*, Jun. 2021, pp. 1–6. doi: 10.1109/ISDFS52919.2021.9486315.

[17] H. I. Okagbue, P. E. Oguntunde, E. C. M. Obasi, P. I. Adamu, and A. A. Opanuga, "Diagnosing malaria from some symptoms: a machine learning approach and public health implications," *Health Technol. (Berl).*, vol. 11, no. 1, pp. 23–37, Jan. 2021, doi: 10.1007/s12553-020-00488-5.

[18] N. O. Adeboye, O. V. Abimbola, and S. O. Folorunso, "Malaria patients in Nigeria: Data exploration approach," *Data Br.*, vol. 28, Feb. 2020, doi: 10.1016/j.dib.2019.104997.

[19] M. O. Arowolo, M. O. Adebiyi, A. A. Adebiyi, and C. Aremu, "An ICA-ensemble learning approaches for prediction of RNAseq malaria vector gene expression data classification," *Int. J. Electr. Comput. Eng.*, vol. 11, no. 2, pp. 1561–1569, Apr. 2021, doi: 10.11591/ijece.v11i2.pp1561-1569.

[20] O. A. Abisoye and R. G. Jimoh, "Comparative Study on the Prediction of Symptomatic and Climatic based Malaria Parasite Counts Using Machine Learning Models," *Int. J. Mod. Educ. Comput. Sci.*, vol. 10, no. 4, pp. 18–25, Apr. 2018, doi: 10.5815/ijmecs.2018.04.03.

[21] N. Maurice *et al.*, "Malaria Epidemic Prediction Model by Using Twitter Data and Precipitation Volume in Nigeria," *Multimed. Soc.*, vol. 22, no. 5, pp. 588–600, 2019, doi: https://doi.org/10.9717/kmms.2019.22.5.588.

[22] M. Masinde, "Africa's Malaria Epidemic Predictor: Application of Machine Learning on Malaria Incidence and Climate Data," in *ACM International Conference Proceeding Series*, Mar. 2020, pp. 29–37. doi: 10.1145/3388142.3388158.

[23] B. J. Brown *et al.*, "Data-driven malaria prevalence prediction in large densely populated urban holoendemic sub-Saharan West Africa," *Sci. Rep.*, vol. 10, no. 1, Dec. 2020, doi: 10.1038/s41598-020-72575-6.

[24] O. N. Bridget, "Machine Learning Techniques for Malaria Incidence and Tuberculosis

Prediction," African University of Science and Technology (AUST), 2021. Accessed: Jan. 26, 2023. [Online]. Available: http://repository.aust.edu.ng/xmlui/handle/123456789/5096

[25] A. P. Idowu, N. Okoronkwo, and R. E. Adagunodo, "Spatial Predictive Model for Malaria in Nigeria," 2009. [Online]. Available: www.jhidc.org

[26] DHS, "The DHS Program Survey Methodology - Malaria Indicators Survey (MIS)", [Online]. Available: https://www.dhsprogram.com/Methodology/Survey-Types/MIS.cfm

[27] DHS, "Demographic and Health Surveys Methodology (Variable Descriptions)," 2013, [Online]. Available: https://dhsprogram.com/publications/publication-dhsg4-dhs-questionnaires-and-manuals.cfm

[28] S. Sawangarreerak and P. Thanathamathee, "Random Forest with Sampling Techniques for Handling Imbalanced Prediction of University Student Depression," *Information*, vol. 11, no. 11, p. 519, Nov. 2020, doi: 10.3390/info11110519.