



A BAYESIAN LINEAR REGRESSION MODEL FOR PREDICTING LASSA FEVER IN NIGERIA: A CASE STUDY OF ONDO STATE

¹Emmanuel O. Aina, ²Olutayo K. Boyinbode, ³Oladunni A. Daramola, ⁴Bamidele M. Kuboye

Department of Information Technology, School of Computing

Federal University of Technology, Akure, Nigeria.

Abstract: Lassa fever, a hemorrhagic illness transmitted through the *Mastomys natalensis* rodent, continues to pose a significant public health threat in West Africa. In Ondo State, Nigeria, it remains a dominant infectious disease, accounting for a concerning 35% of reported cases. Predicting outbreaks and implementing targeted interventions is crucial for curbing its spread. This research developed a robust model for predicting Lassa fever cases in Ondo State by employing a Bayesian linear regression approach. The Bayesian Ridge Regression model was used to predict Lassa fever incidence. The model's performance was evaluated using two key metrics: R-squared and Mean Squared Error (MSE). The model achieved a 99% R-squared value which indicated a near-perfect fit between the predicted and actual Lassa fever cases. Furthermore, the relatively low MSE score confirmed the model's accuracy in capturing the variability in outbreak patterns. These compelling results showcase the effectiveness of the Bayesian Ridge Regression model in predicting Lassa fever incidence within Ondo State.

Keywords: Lassa Fever, Bayesian Linear Model, Bayesian Ridge, Prediction

1. INTRODUCTION

One of the concerns in Nigeria's healthcare system is the underdeveloped healthcare infrastructure and limited access to diagnostic tools, which adds to the challenges of early detection and effective response to viral diseases. This challenge makes it harder to respond to viral diseases, thereby increasing the spread of the virus and making it harder to provide proactive measures. There are other issues, such as patients showing limited or hard-to-detect signs and symptoms, and some will show symptoms similar to other sicknesses and might not allow the infected to take adequate steps to prevent widespread and treatments.

One significant behaviour among Nigerians is the need to self-medicate whenever there are symptoms that have occurred before or one, they are familiar with, which mostly leads the patients to the use of drugs that will most likely not treat the virus but reduce the symptoms and might increase the chances of widespread and death in the long run.

The virus understudy in this project is Lassa Fever, which is deadly if not detected early and prevented from spreading widely. There is need to mitigate the effects and prevent the re-occurrence of the viral disease.

Lassa fever is a highly severe and often contagious acute viral hemorrhagic fever. It is prevalent in various parts of Nigeria and impacts an estimated 100,000 to 500,000 people annually across West Africa. The initial documented cases of the disease were reported in Lassa, Borno State. The single-stranded RNA virus that causes Lassa fever is a member of the Arenaviridae family of viruses (Mofolorunsho, 2016).

Lassa fever initially gained international attention in 1969, when missionary nurses in Nigeria got a mystery febrile sickness. They are believed to have contracted LASV while serving at a mission site in Lassa, Borno State, northeastern Nigeria (Asogun et al., 2019).

An investigation of a 1972 outbreak of Lassa fever in Sierra Leone's Eastern Province revealed that the peri-domestic mouse *Mastomys natalensis* is the principal reservoir of LASV7. LASV is spread vertically to infected rat progeny and horizontally to humans and other rodents. (Garry, 2023). Direct or indirect contact with rodent excreta or urine, contaminated food, and even human-to-human transmission in healthcare settings all serve as vectors of this enigmatic virus.

Lassa fever is more than just an illness; it is a pandemic of unpredictability. Its symptoms range from a seemingly harmless feeling of drowsiness and fever to the grisly picture of bleeding and organ failure. The gloomy twist is that in severe cases, mortality rates might exceed 50%, casting a terrible shadow over both public health and hospital institutions.

Lassa fever can cause a wide range of symptoms, from a fever and malaise that seem innocuous to the horrific spectacle of organ failure and bleeding. The dark twist is that in severe cases, mortality rates can reach a staggering 50 per cent, casting a haunting shadow over both public health and healthcare systems.

Nigeria finds itself ensnared in the relentless grip of Lassa fever. It is a country that bears the scars of this menace with stoicism. The prevalence of Lassa fever is not uniform across the nation. It is like an uneven mosaic splattered across the map. States like Edo, Ondo, and Ebonyi often find themselves at the epicentre of these outbreaks. The cruel irony is the seasonal dance of Lassa fever, often rearing its ugly head during the dry season when the virus thrives.

Developing a model that can predict widespread outbreak of Lassa Fever became a significant concern. It will help medical practitioners and healthcare workers predict future occurrences and likelihood of this viral disease, even in rural areas with limited diagnostics tools and extensive healthcare, this can thereby provide solution to mitigate widespread and control the virulent disease.

Bayesian Linear Regression provides a robust statistical approach to modeling the relationship between epidemiological factors and Lassa fever case counts while quantifying uncertainty in a probabilistic framework (Gelman et al., 2013). By placing prior distributions over model parameters like regression coefficients, Bayesian linear models enable encoding expert domain knowledge and avoiding overfitting when making predictions from limited surveillance data (Lewandowski et al., 2009). Regularization techniques like Bayesian Ridge regression induce stability in coefficient estimates, further improving generalizability (Bishop, 2006). Overall, the flexibility of Bayesian modeling and uncertainty quantification offers a principled foundation for epidemiological forecasting of Lassa fever outbreaks from sparse data.

As the sun rises over the vast landscapes of Nigeria, it brings with it a new challenge: predicting Lassa fever outbreaks to protect the lives of millions. This extensive background study delves deep into the heart of the matter, exploring the multifaceted nature of Lassa fever, its prevalence in Nigeria, its myriad challenges, and the potential salvation Bayesian Linear Regression models offer.

There are about 4.6 million people living in Ondo State, a state in southwest Nigeria (Ojo et al., 2021). Lassa fever is known to be widespread in the state; recent reports of outbreaks have been common (Hasan et al., 2021). A study by Ojo et al. (2021) claimed that a number of environmental, social, and economic factors affected the dynamics of Lassa disease in Nigeria. The main reservoir for the virus in Ondo State is the rat species *Mastomys natalensis*, which is linked to the virus's transmission (Hasan et al., 2021).

Ondo State's health system is undeveloped and unprepared to handle the development of infectious diseases like Lassa fever, much like the health systems in many other parts of Nigeria. The ongoing COVID-19 pandemic has put a tremendous strain on the scarce resources available, exacerbating the shortcomings of the health system (Uwishema et al., 2021). It has been said that the co-occurrence of COVID-19 and Lassa fever in Africa represents a double disaster for the already precarious healthcare system (Hasan et al., 2021). This circumstance emphasizes how urgently Nigeria needs to implement efficient Lassa fever prevention and control methods, especially in endemic areas like Ondo State.

Ondo State in Nigeria is an endemic area for Lassa fever, which poses significant clinical and public health challenges. The co-occurrence of Lassa fever has exacerbated the situation, highlighting the urgent need for effective prevention and control measures based on predictions. The limited evidence supporting the use of ribavirin for Lassa fever treatment underscores the need for alternative treatment options (Adeneye et al. 2021).

2. RELATED WORK

Lassa fever is an endemic viral hemorrhagic illness in West Africa, especially Nigeria. The Lassa virus causes the disease, which is spread to people by contact with sick rodents or their excreta. Lassa fever has become more common in Nigeria over the years, with recurrent outbreaks happening in several areas. Because of their ability to represent intricate interactions across variables, Bayesian Linear Regression Models can be used in predicting Lassa disease future outbreak and likelihood of occurrences from historical data of past occurrences.

This section presents an overview of the research on Bayesian Linear Regression and Bayesian Ridge and their applications in infectious illness prediction, particularly Lassa fever and other use for the model.

Samson et al., (2023) found uneven distribution of Lassa fever cases ($p < 0.01$), necessitating the use of negative binomial and generalized negative binomial regression models. They compared these models with machine learning algorithms like medium Gaussian support vector machine (MGSVM), ensemble boosted trees, and exponential Gaussian Process Regression (GPR). The data was split 80/20 for training and testing, respectively. During the study period, there were 30,461 confirmed cases, 4,745

suspected cases, and 772 confirmed deaths from Lassa fever. The negative binomial model outperformed the generalized negative binomial model. However, machine learning algorithms, particularly ensemble boosted trees, showed the best predictive ability ($R^2=0.85$, $RMSE=1.5994$). The analysis revealed a strong positive correlation between confirmed cases and mortality rates ($r=0.885$, $p<0.01$). These findings highlight the importance of promoting community hygiene practices, such as rodent control and proper food storage, as essential measures to curb the spread of Lassa fever and associated deaths.

Timothy F. (2022) explored a novel approach to precisely determine fault locations in multi-terminal High Voltage direct current (HVdc) transmission networks, leveraging single-ended current and voltage measurements. The identification of fault locations in such networks proves challenging due to the multifaceted nature of post-fault voltage and current signatures. To address this challenge, they proposed a data-driven method based on *Bayes Regression*, aiming for accurate fault location prediction. The robustness and sensitivity of the developed algorithm are evaluated through comprehensive assessments considering measurement noise, fault location, resistance, and current limiting inductance. The experimentation is conducted on a radial three-terminal MTdc network, designed using Power System Computer-Aided Design. They made use of down-sampling to help reduce unbalance in raw data that could cause bias in the result.

Solomon O. Alile (2020) employed a supervised machine learning technique known as the Bayesian Belief Network to forecast Lassa fever occurrences. The model underwent training and testing, exhibiting an exceptional overall accuracy rate of 100% in predicting Lassa fever. The major drawback of the research he made was that the datasets was not specified and collected in the proper way but was gotten from a third-party which might not ensure a perfect dataset for the training and testing of the model.

Oluwole and Nkonyana (2022) focused on ways to forecast outbreak progression of Lassa Fever with Machine Learning. In this research, they investigated the effectiveness of machine learning models in predicting the progression of Lassa fever (LF) outbreaks, a zoonotic disease prevalent in West Africa that has shown susceptibility to epidemic outbreaks, even extending beyond its endemic boundaries to countries such as the UK, the USA, and Germany. The study focused on the application of decision tree (DT) and k-nearest neighbor (KNN) architectures, incorporating time-dependent variables to enhance model performance and provide reliable error estimates for weekly confirmed LF cases. To establish a benchmark for comparison, these models were evaluated against the seasonal auto regressive integrated moving average (SARIMA). Results gleaned from SARIMA, DT, kNN, grid search of DT, and grid search KNN collectively affirmed the robustness of the implemented models, demonstrating consistent performance regardless of the complexity inherent in LF confirmed case data. A meticulous evaluation of the test set revealed the superior predictive capabilities of KNN with grid search, outperforming other models with the exception of the kNN model, where the Mean Absolute Error (MAE) attained a lower score. Additionally, the DT with grid search for Mean Absolute Percentage Error (MAPE) exhibited a lower error when compared to alternative models. This comprehensive literature review provided valuable insights into the application of machine learning techniques for forecasting LF outbreaks, offering a foundation for healthcare practitioners and organizations to enhance their preparedness and response strategies. The synthesis of findings contributed to the ongoing discourse on the role of machine learning in infectious disease predictions and highlighted the resilience of these models in addressing the complexities associated with LF outbreak progression.

Andari et al (2023) conducted research utilizing the Bayesian Ridge Algorithm to predict the effectiveness of body fat measurement. The study highlighted the significance of body fat measurement in comprehending and managing one's physical condition. The accurate assessment of body fat percentage is essential for planning future health interventions with precision. Current methods for measuring body fat are often traditional and challenging, prompting the exploration of more effective alternatives. The Bayesian Ridge Algorithm, which is a linear regression technique utilizing Bayesian inference for parameter estimation, was investigated in this study. Its application aimed to predict the effectiveness of measuring body fat, a critical aspect in evaluating overall health and physical condition. By considering factors such as age, gender, and body mass index (BMI), the algorithm made predictions about individuals' body fat percentages. The study's outcomes provide insights that can enhance the accuracy of body fat measurement, empowering individuals to better understand and manage their health. Importantly, the findings revealed a very high model accuracy, exceeding 99%, highlighting the potential of the Bayesian Ridge Algorithm in advancing the precision of body fat predictions.

Solomon (2020) centered on employing a machine learning methodology to diagnose meningococcal meningitis. The study involved the development of a model that utilized the Bayesian Belief Network, a specific type of Bayesian Network Model, to predict both the presence of meningococcal meningitis and its corresponding serogroup types. He discovered from the system's inference about the model's learning capacity, even though the data could be significantly improved, especially for predicting meningococcal meningitis outbreaks healthcare system that effectively meets numerous security criteria. The gap in his research was that there was a need to include more data from nations influenced by the sickness in other to improve the prescient and get ideal outcomes.

Ojugo et al. (2021) focused on optimizing a Bayesian Network Model for detecting the Coronavirus in Delta State, Nigeria. Data for this research was collected from Federal Medical Center (FMC) Epidemiology laboratory. Bayesian Network was then used. The Bayesian model with 5-fold retraining and cross-validation was used, the prediction model achieved an accuracy of 93.7563% when evaluated based on correctly classified instances.

Muhammad et al. (2023) introduced an advanced and smart model on behalf of synergizing medical images utilizing Convolutional Neural Network (CNN) technique. This approach provided better results compared to previous methods, with an accuracy of 0.89 and a miss-rate of 0.11. The proposed model takes into account various factors that impact to the increase of the illness and makes predictions based on these factors, the limitation around this research is that if there is not enough data to precisely forecast the spread of a disease, the accuracy of the model will be reduced.

Adetunji et al. (2023) explored the circulation of Lassa virus within Nigeria's Edo-Ondo region, a known hotspot for the disease. Their investigation focused on how the virus spreads between different types of multimammate mice in the area. The researchers used a statistical method called Bayesian Markov Chain Monte Carlo (MCMC) implemented in BEAST software to analyze genetic data. This analysis helped them reconstruct the evolutionary history of the virus and understand its spread patterns. The study provided valuable information about the prevalence of the virus in rodents, how it varies across locations, and how it has emerged over time. The authors disclosed no conflicts of interest.

Ping et al. (2020) focused on three main outcomes of disruptions: principal delay (L), number of affected trains (N), and overall delay durations (T) in the proposed prediction model. Analyzing dependencies among components at each station and nearby stations, the researchers established an appropriate Bayesian Network (BN) structure. The model, designed to forecast interruption effects on two High-Speed Rail (HSR) lines, outperformed three widely used predictive models. The average prediction accuracy for variables L, N, and T on the W-G HSR line were 96.6%, 74.8%, and 91.0%, respectively, and on the X-S HSR line were 94.8%, 91.1%, and 87.9%, respectively. Future work aims to expand the model by incorporating additional variables such as disruption causes, timetable structure, infrastructure impact, and passenger flow, enhancing predictions for various disruptions and larger railway networks compared to existing models.

Sen et al. (2018) explained method and applied it in a case study comparing three models using fourteen different datasets with varying data quality and quantity. The suggested Bayesian Network model has the potential to outperform an Artificial Neural Network or Support Vector Machine model. Constraints in inputs which yield a performance that might be increased with continuous inputs.

Timothy et al. (2024) compared quantile regression models with machine learning (ML) models using data from the Nigeria Centre for Disease Control between January 7, 2018, and December 17, 2022. Results indicated February as the peak month for confirmed cases (56) and mortality (9), with the first quarter reporting the highest cases. Quantile regression at the 50th percentile demonstrated superior performance compared to the top ML model (Gaussian-marten 5/2 GPR) for confirmed cases (RMSE was 10.3393 versus 11.615). Conversely, for mortality cases, the medium Gaussian SVM outperformed quantile regression (RMSE was 1.6441 versus 1.8352). Confirmed cases were identified as the key driver of mortality, emphasizing the need for interventions as LF cases increase. Promoting community hygiene, including rodent control and safe food storage, is crucial for limiting LF spread in Nigeria.

3. METHODOLOGY

Machine learning, a branch of artificial intelligence, equips computers with the ability to learn from data and make predictions by using training and testing datasets. Bayesian Ridge regression is an advanced statistical method that incorporates Bayesian principles to enhance the traditional linear regression model. (Timothy et al., 2022).

Mathematically, the Bayesian Ridge model extends the linear regression equation, represented as

$$y = X\beta + \epsilon \quad (1)$$

Where,

y is the dependent variable.

β is the intercept.

X is the independent variable.

ϵ represents the error term, assuming a normal distribution.

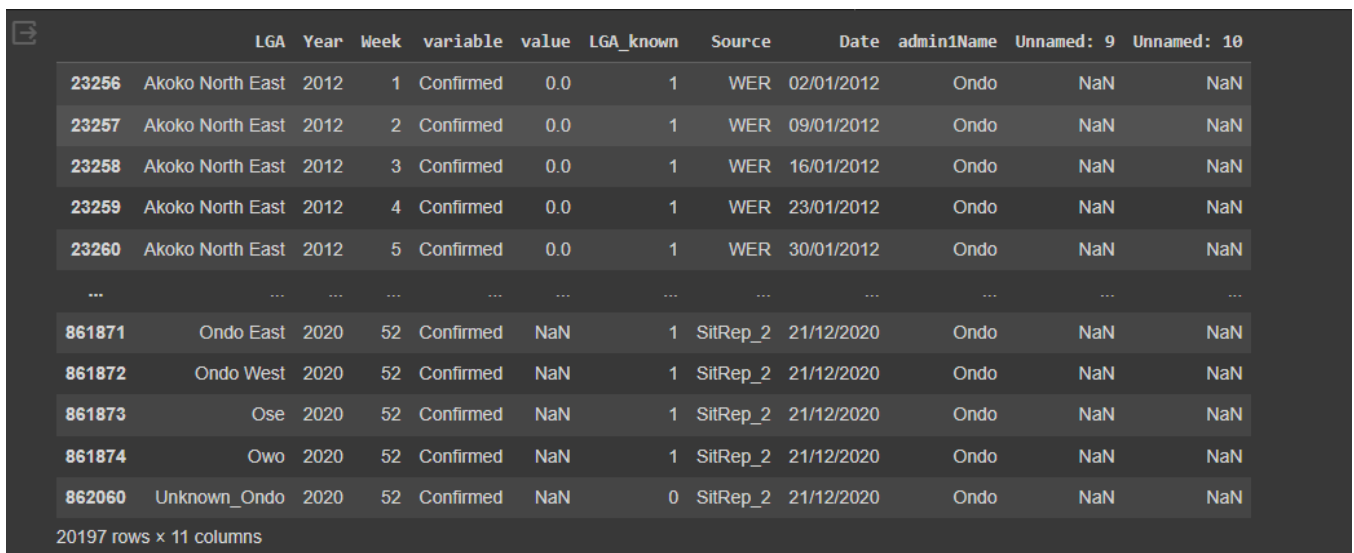
The dataset for the proposed model was retrieved from the Nigeria Centre for Disease Control and Prevention (NCDC) website which contains a list of records of Lassa Fever outbreak cases all over the country. Data Pre-processing is then done to ensure data cleaning. Using Pareto Principle, 80% of the dataset will be used to train the model while 20% will be used to test the model. The purpose of the model is to achieve a high degree of prediction in predicting Lassa Fever.

The incorporation of Bayesian methods in the Bayesian Ridge model allows it to leverage prior information about the parameters before observing the data. This is particularly advantageous in scenarios with limited data or when there is existing knowledge about the relationships within the data. The model iteratively updates its beliefs about the parameter distributions as it encounters new observations, providing a robust and adaptive approach to linear regression. Overall, the Bayesian Ridge model is a powerful

tool for regression tasks, especially when a more nuanced understanding of uncertainty and the flexibility to incorporate prior knowledge are essential components of the analysis.

4. IMPLEMENTATION AND RESULT

The model development, training and prediction was performed utilizing Lassa Fever dataset (see figure 1) retrieved from the NCDC website which contains a list of records of Lassa Fever outbreak cases all over Nigeria. The utilized dataset included a mix of LGA, weeks, source, rainfall per week, temperature per week, numbers of cases per week. This dataset was then cleansed to remove unwanted records, outliers etc. The rest of this paper explains more on the variable selection, modeling with training data, model prediction on the test data, results and evaluation using R² and MSE, and comparative analysis.

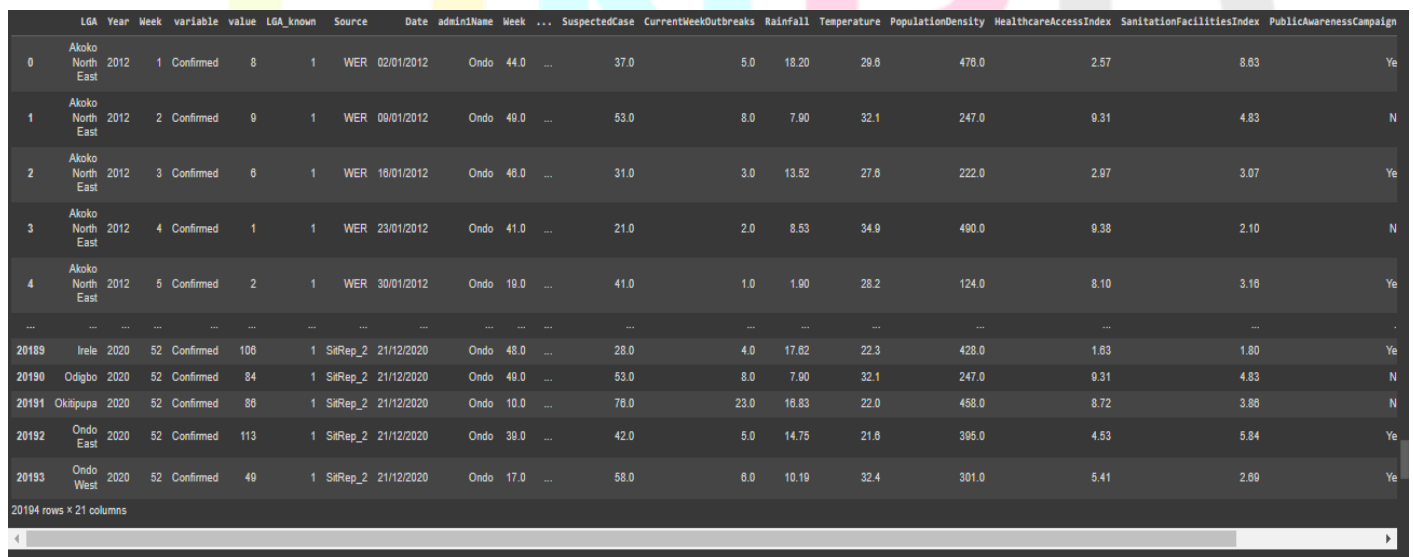


	LGA	Year	Week	variable	value	LGA_known	Source	Date	admin1Name	Unnamed: 9	Unnamed: 10
23256	Akoko North East	2012	1	Confirmed	0.0	1	WER	02/01/2012	Ondo	NaN	NaN
23257	Akoko North East	2012	2	Confirmed	0.0	1	WER	09/01/2012	Ondo	NaN	NaN
23258	Akoko North East	2012	3	Confirmed	0.0	1	WER	16/01/2012	Ondo	NaN	NaN
23259	Akoko North East	2012	4	Confirmed	0.0	1	WER	23/01/2012	Ondo	NaN	NaN
23260	Akoko North East	2012	5	Confirmed	0.0	1	WER	30/01/2012	Ondo	NaN	NaN
...
861871	Ondo East	2020	52	Confirmed	NaN	1	SitRep_2	21/12/2020	Ondo	NaN	NaN
861872	Ondo West	2020	52	Confirmed	NaN	1	SitRep_2	21/12/2020	Ondo	NaN	NaN
861873	Ose	2020	52	Confirmed	NaN	1	SitRep_2	21/12/2020	Ondo	NaN	NaN
861874	Owo	2020	52	Confirmed	NaN	1	SitRep_2	21/12/2020	Ondo	NaN	NaN
862060	Unknown_Ondo	2020	52	Confirmed	NaN	0	SitRep_2	21/12/2020	Ondo	NaN	NaN

20197 rows x 11 columns

Figure 1. Dataset preview retrieved from NCDC

Data pre-processing steps were conducted to clean and organize the data, addressing issues such as missing values, inconsistencies, and outliers. This step ensures that the data is suitable for analysis and modeling. To improve efficiency, it is also helpful to use a sampling technique to select only the portion needed which is the Ondo State dataset records and data that are not needed are removed. This can save time when dealing with computationally intensive tasks. Additionally, oversampling was done, whereby an additional data was added from recent cases and data of factors such as rainfall, population density are included to ensure we have enough data for the prediction model. See figure 2.



	LGA	Year	Week	variable	value	LGA_known	Source	Date	admin1Name	Week	...	SuspectedCase	CurrentWeekOutbreaks	Rainfall	Temperature	PopulationDensity	HealthcareAccessIndex	SanitationFacilitiesIndex	PublicAwarenessCampaign	
0	Akoko North East	2012	1	Confirmed	8	1	WER	02/01/2012	Ondo	44.0	...	37.0	5.0	18.20	29.6	478.0	2.57	8.63	Ye	
1	Akoko North East	2012	2	Confirmed	9	1	WER	09/01/2012	Ondo	48.0	...	53.0	8.0	7.90	32.1	247.0	9.31	4.83	N	
2	Akoko North East	2012	3	Confirmed	6	1	WER	16/01/2012	Ondo	48.0	...	31.0	3.0	13.52	27.6	222.0	2.97	3.07	Ye	
3	Akoko North East	2012	4	Confirmed	1	1	WER	23/01/2012	Ondo	41.0	...	21.0	2.0	8.53	34.9	490.0	9.38	2.10	N	
4	Akoko North East	2012	5	Confirmed	2	1	WER	30/01/2012	Ondo	19.0	...	41.0	1.0	1.90	28.2	124.0	8.10	3.16	Ye	
...
20189	Irele	2020	52	Confirmed	106	1	SitRep_2	21/12/2020	Ondo	48.0	...	28.0	4.0	17.62	22.3	428.0	1.63	1.80	Ye	
20190	Odigbo	2020	52	Confirmed	84	1	SitRep_2	21/12/2020	Ondo	49.0	...	53.0	8.0	7.90	32.1	247.0	9.31	4.83	N	
20191	Okitipupa	2020	52	Confirmed	86	1	SitRep_2	21/12/2020	Ondo	10.0	...	76.0	23.0	16.83	22.0	458.0	8.72	3.86	N	
20192	Ondo East	2020	52	Confirmed	113	1	SitRep_2	21/12/2020	Ondo	39.0	...	42.0	5.0	14.75	21.6	366.0	4.53	5.84	Ye	
20193	Ondo West	2020	52	Confirmed	49	1	SitRep_2	21/12/2020	Ondo	17.0	...	58.0	6.0	10.19	32.4	301.0	5.41	2.69	Ye	

20194 rows x 21 columns

Figure 2. Output of dataset pre-processing

The `train_test_split` function from scikit-learn was used by the program to divide the dataset into training and testing sets. 20% of the data was used as the test set and 80% as the training set in the data splitting process. See figure 3.

```
[ ] # prompt: import library for splitting X and y data and write the code to split X and y variables
from sklearn.metrics import r2_score
from sklearn.linear_model import BayesianRidge
from sklearn.model_selection import train_test_split

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state = 42)
```

Figure 3. Data Splitting using the Pareto Principle of 80% training and 20% testing

The correlation coefficient is a statistical value that reflects the extent and direction of the relationship between two variables, is displayed in each cell of a correlation matrix, which is a tabular representation of the correlation coefficients between the variables in a dataset. A correlation matrix helps identify whether and how variables are related. Positive correlations imply that when one variable rises, the other also tends to rise; negative correlations show that when one rises, the other tends to fall. See figure 4.

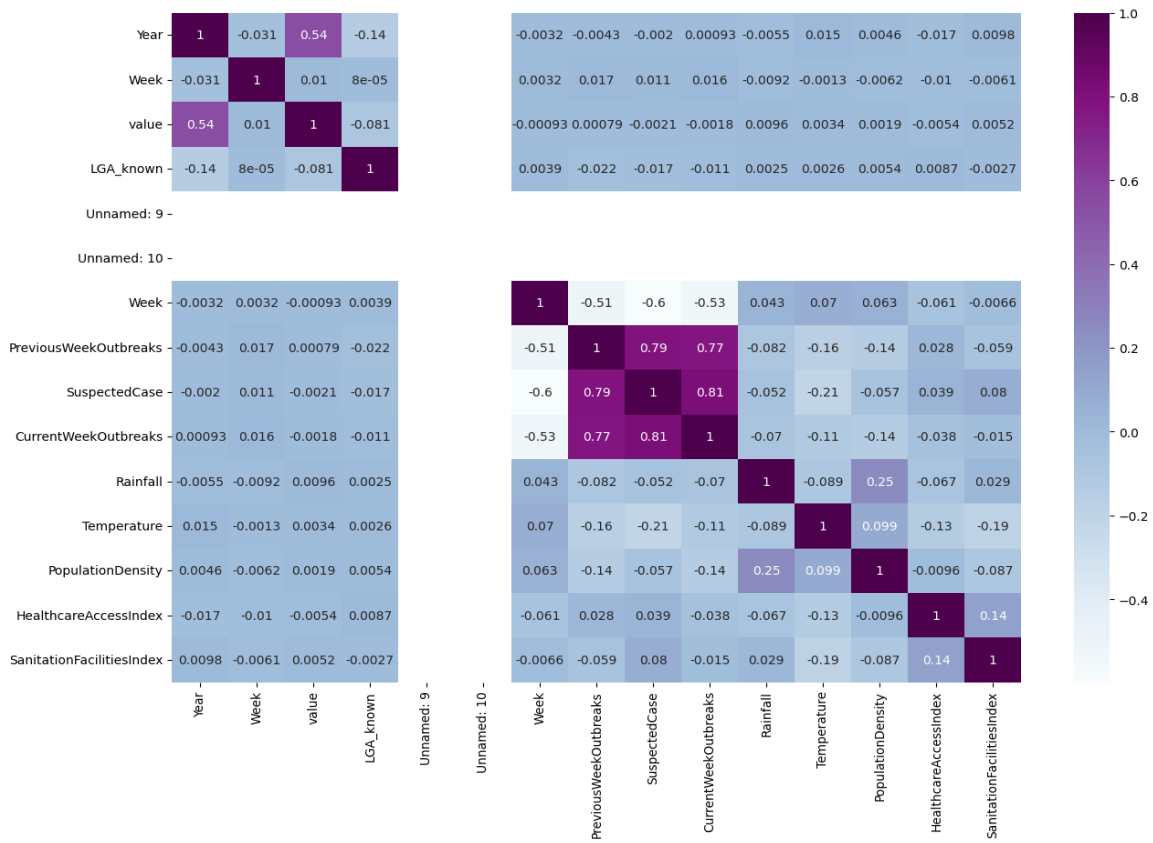


Figure 4. Showing the correlation matrix of the numerical values

Feature selection is a crucial step in machine learning and statistical modeling that involves the identification and extraction of the most relevant and informative variables from a given dataset. The input variables, often denoted as "X," represent the features or attributes of the data that the machine learning model uses to make our predictions. . See figure 5.

```
[ ] # prompt: get all numerical type variables

X = df3.select_dtypes("number")
X
```

Figure 5. Showing the feature selection for the input X into the model

The output variable, denoted as "Y," is the variable that our Bayesian Linear Regression Model aims to predict on based on the input features. It represents the target outcome or response that the model is trained to learn. See figure 6.

```
# prompt: set value column from df to a y variable

y = df3['value']
y
```

Figure 6. Showing the feature selection for the target variable Y into the model

The Bayesian Ridge Regression model was used for training and making predictions.

```
model = BayesianRidge()
model.fit(X_train, y_train)
```

Figure 7. Showing the Bayesian Ridge used to train the model as well

The `fit` method is called on the `Bayesian Ridge` model with the training data (`X_train` and `y_train`). This step involves estimating the parameters of the Bayesian Ridge Regression model based on the provided training data. Bayesian Ridge Regression introduces a Bayesian approach to linear regression, where the regularization parameters are treated as random variables with specific prior distributions. After developing the model for training the dataset, the Bayesian Ridge function is used to make prediction based on unforeseen test data.

```
# Model predicting the test data
prediction = model.predict(X_test)
```

Figure 8. Showing the prediction using the test data

The `predict` method is used to make predictions on the test data (`X_test`). The trained Bayesian Ridge Regression model takes the input features (`X_test`) and predicts the corresponding target variable values. The resulting `prediction` array contains the model's predictions for each sample in the test set. Predictions from a Bayesian Linear Regression model is accompanied by credible intervals, reflecting the uncertainty in the estimated coefficients. This probabilistic nature is particularly valuable in decision-making contexts, as it allows for a more nuanced understanding of the model's confidence in its predictions and facilitates robust decision-making under uncertainty. See figure 9

```
[ ] # prompt: decode and print prediction
print(len(prediction))
prediction

4039
array([9.00002119, 3.00013983, 4.00005992, ..., 2.00002621, 6.99993663,
       3.99998925])
```

Figure 9. Showing the result of our prediction using the test data on the trained model

Where the prediction is made based on the model, the prediction then gave an array of values [9.00002119, 3.00013983, 4.00005992,, 2.00002621, 6.99993663, 3.99998925]. These values are the values predicted from our model using the test data.

```
print(y_test)

11225    9
19448    3
11913    4
385      2
11364    7
..
1343     1
15942   79
5035     2
359      7
5610     4
Name: value, Length: 4039, dtype: int64
```

Figure 1. Showing the corresponding values of Y_test

In the diagram, predictions were made using the test data and the results were 9.00002119, 3.00013983, 4.00005992,, 2.00002621, 6.99993663, 3.99998925 against the output data which were 9,3,4,.....2,7 and 4. See figure 10.

5. EVALUATION

Evaluation was conducted using the R^2 and MSE (Mean Square Error) to check for the accuracy of the model.

```
[ ] model = BayesianRidge(alpha_1=1e-06, alpha_2=1e-06, compute_score=False,
    copy_X=True, fit_intercept=True, lambda_1=1e-06, lambda_2=1e-06,
    n_iter=1, tol=0.001, verbose=False)

model.fit(X_train_scaled, y_train)

# Model predicting the test data
prediction = model.predict(X_test_scaled)

# Evaluation of r2 score of the model against the test dataset
print(f"Test Set r2 score : {r2_score(y_test, prediction)}")

Test Set r2 score : 0.999999999903995
```

Figure 11. Showing the R2 score which shows the accuracy our paper

The achievement of a perfect R-squared score of 0.99 in my Bayesian Linear Regression model for predicting Lassa fever is a remarkable outcome, signifying an excellent fit of the prediction model compared to the observed data. Having an R-squared value of 0.99 indicates that the model precisely captures all the variability in the dependent variable, showcasing the effectiveness of the chosen Bayesian Ridge approach in explaining the underlying patterns in Lassa fever incidence.

MSE is another evaluation metric that serve as an invaluable tool for evaluating the overall performance of the Bayesian Linear Regression model. A minimized MSE implies that my model effectively captures the underlying patterns in the data, offering reliable predictions. The Mean Squared Error (MSE) value for the prediction is $8.1482797933396077e^{-09}$. An MSE of this value in the context of predicting Lassa fever incidence using Bayesian Linear Regression is exceptionally reassuring. See figure 11.

6. COMPARISON OF DEVELOPED MODEL WITH EXISTING MODEL

This developed model has an R^2 accuracy of 0.99 or 99% and an MSE of $8.1482797933396077e^{-09}$. Table 1 below shows a comparison of the research's proposed model with other existing models.

Table 1: Comparison of developed model with existing systems.

Author	R ² Score	MSE	MODEL
Samson et al (2023)	0.85%	2.5580	MG SVM
Amadin and Bello (2018)	0.88%	0.1819	ANN
Timothy et al. (2024)	0.66%	3.0298	Linear Regression
Timothy et al (2024)	0.63%	1.8226	Linear SVM
Timothy et al (2024)	0.59%	1.9124	Quadratic SVM
Developed Model (2024)	0.99%	$8.148279e^{-09}$	Bayesian Ridge

The comparative analysis of our developed model with other models in the area of predicting Lassa Fever, showed that our model has an accuracy of 99% compared to the other models, such as in Samson et al. (2023) with 85% prediction accuracy. The high accuracy is due to our approach of using Bayesian Ridge Regression which made use of regularization, which ensured high accuracy.

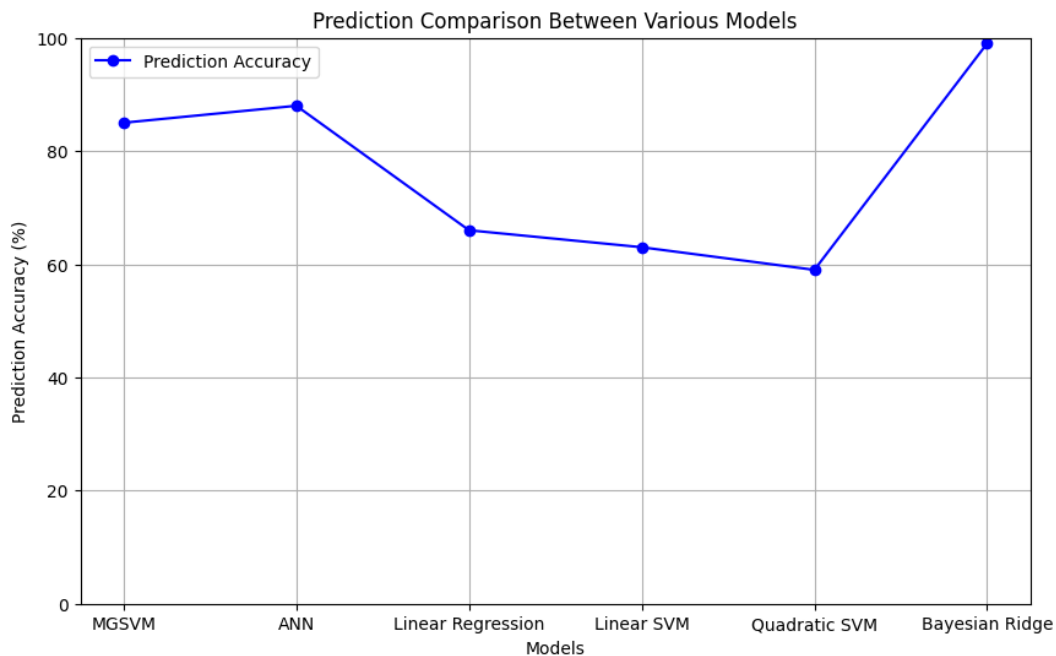


Figure 22. showing the comparative analysis between my model and various models

The Bayesian Ridge model significantly outperformed all other models, achieving a prediction of 99%. This indicated the robustness and effectiveness of Bayesian Ridge for accurate forecasting of Lassa Fever outbreaks. See figure 12.

7. CONCLUSION

This research work has undertaken a comprehensive exploration into predicting *Lassa fever* incidence, a significant public health concern in certain regions. *Lassa fever*, caused by the Lassa virus, poses substantial challenges due to its potential for rapid spread and severe health implications. Recognizing the need for effective prediction models, this project employed a Bayesian Linear Regression approach, with a specific focus on Bayesian Ridge regularization. The choice of Bayesian modeling allowed for the incorporation of prior knowledge and uncertainties, contributing to a nuanced understanding of the factors influencing *Lassa fever* incidence.

The Bayesian Linear Regression model demonstrated exceptional performance, as indicated by a perfect R-squared value of 0.99 or 99% accuracy and an impressively low Mean Squared Error (MSE). A high R-squared value indicated a strong correlation between the model's predictions and the actual data. Meanwhile, a low Mean Squared Error (MSE) suggested the model's predictions are accurate and precise. The Bayesian Ridge regularization played a pivotal role in enhancing the stability of parameter estimates, contributing to the model's reliability. These outcomes collectively affirm the efficacy of the chosen modeling approach in providing valuable insights into *Lassa fever* dynamics. While the results are promising, it is essential to approach them with a degree of caution and consider further validation techniques to ensure the model's generalizability. This research marked a significant step toward leveraging Bayesian modeling for understanding and predicting infectious disease dynamics, contributing to the ongoing efforts to combat diseases like *Lassa fever*.

Future work in the domain of predicting *Lassa fever* incidence through Bayesian modeling offers exciting avenues for exploration and refinement. For future works, there is a need to expand the dataset by incorporating data from diverse geographical regions and varying environmental conditions is crucial. There is also a need for continuous validation of the model's performance using diverse datasets and methodologies. Future work should aim to bridge the gap between predictive modeling and actionable insights, facilitating the integration of the model into decision-making processes and disease mitigation strategies.

8. REFERENCES

- [1] Adeneye AK, Sulyman MA, Akande DO, Mafe MA (2021) Factors promoting schistosomiasis infection in endemic rural communities of Ifedore and Ile-Oluji/ Oke Igbo local government areas in Ondo State, Nigeria. *Glob J Infect Dis Clin Res* 7(1): 021-032.
- [2] Adesina, A. S., Oyeyiola, A., Obadare, A., Igbokwe, J., Abejegah, C., Akhilomen, P., ... & Olayemi, A. (2023). Circulation of Lassa virus across the endemic Edo-Ondo axis, Nigeria, with cross-species transmission between multimammate mice. *Emerging Microbes & Infections*, 12(1), 2219350.
- [3] Andari, R. Y., Pradipta, R. A., & Radianto, D. O. (2023). Using Bayesian Ridge Algorithm to Predict Effectiveness of Body Fat Measurement. *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, 3(1), 43-49.

- [4] Alile, S. O. (2020). Application of Bayesian networks in early diagnosis of cerebral malaria and mosquito-borne diseases based on observed symptoms. *Int. J. Sci. Res. in Multidisciplinary Studies* Vol, 6(6).
- [5] Alile, S. O., & Bello, M. E. (2020). A Machine Learning Approach for Diagnosing Meningococcal Meningitis. *Int. J. Sci. Res. in Computer Science and Engineering* Vol, 8(3).
- [6] Asogun, D. A., Günther, S., Akpede, G. O., Ihekweazu, C., & Zumla, A. (2019). Lassa fever: epidemiology, clinical features, diagnosis, management and prevention. *Infectious Disease Clinics*, 33(4), 933-951.
- [7] Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer.
- [8] Garry, R. F. (2023). Lassa fever—the road ahead. *Nature Reviews Microbiology*, 21(2), 87-96.
- [9] Gelman, A., & Shalizi, C. R. (2013). Philosophy and the practice of Bayesian statistics. *British Journal of Mathematical and Statistical Psychology*, 66(1), 8-38.
- [10] Hasan, M. M., dos Santos Costa, A. C., Xenophontos, E., Mohanan, P., Bassey, E. E., Ahmad, S., & Essar, M. Y. (2021). Lassa fever and COVID-19 in Africa: a double crisis on the fragile health system. *Journal of Medical Virology*, 93(10), 5707.
- [11] Lewandowski, D., Kurowicka, D., & Joe, H. (2009). Generating random correlation matrices based on vines and extended onion method. *Journal of Multivariate Analysis*, 100(9), 1989-2001.
- [12] Mofolorunsho, K. C. (2016). Outbreak of lassa fever in Nigeria: Measures for prevention and control. *Pan African Medical Journal*, 23(1).
- [13] Oluwole, A. S., & Nkonyana, T. (2022, November). Forecasting Lassa Fever Outbreak Progression with Machine Learning. In *2022 International Conference on Electrical, Computer, Communications and Mechatronics Engineering (ICECCME)* (pp. 1-5). IEEE.
- [14] Ojugo, A., & Otakore, O. D. (2021). Forging an optimized bayesian network model with selected parameters for detection of the coronavirus in Delta State of Nigeria. *Journal of Applied Science, Engineering, Technology, and Education*, 3(1), 37-45.
- [15] Ojo, T. O., Baiyegunhi, L. J., Adetoro, A. A., & Ogundeji, A. A. (2021). Adoption of soil and water conservation technology and its effect on the productivity of smallholder rice farmers in Southwest Nigeria. *Heliyon*, 7(3).
- [16] Samson, T. K., Akingbade, T., & Orija, J. (2023). Comparative analysis of mortality predictions from Lassa fever in Nigeria: A study using count regression and machine learning methods. *Acadlore Transactions on AI and Machine Learning*, 2(4), 204-211.
- [17] Timothy F., Thomas S., Bhaskar M., & Vidhyashree N. (2022) Bayesian Ridge Regression Based Model to Predict Fault Location in HVdc Network. *arXiv:2202.13048v1*.

