Spatio-Temporal Modelling and Malaria Outbreak Prediction in Africa.

Abigail Annkah, Mustafa Alghali, Rua Mohammed,
Kobby Panford-Quainoo, Abemgnigni Njifon Marianne.
African Master of Machine Intelligence (AMMI) - AIMS Rwanda

AIMS African Institute for Mathematical Sciences RWANDA

Abstract

Preceding the design of national malaria control and elimination programs, there was a need for understanding the distribution of anopheline mosquitoes. This was the motivation for the creation of a single geo-coded inventory of anophelines using all published and unpublished data records from 1960. This geo-coded and referenced inventory of anophelines in the Afrotropical Region south of the Sahara was the database to be used for the (spatio-temporal) modelling to predict malaria outbreaks in Africa. After pre-processing the data, the task consisted of predicting species types given the location and the year of survey. Therefore we used several methods for training and testing procedures including binary classification for individual species and multilabel classification. Then we compared the performance of those models to select the most accurate one for our work.

Introduction

Malaria is a widely known disease that threatens the lives of many around the globe. The causative parasite, Plasmodium, is transmitted from one person to another when bitten by female mosquitoes of certain species of the genus Anopheles. About 1.143% of the total 3500 mosquito species of Anopheles can transmit malaria. In Africa, particularly the sub Saharan Africa, malarial cases and mortality are increasing fast. The World Health Organization (WHO) malaria report released in November 2018 showed that cases of malaria and mortality were highest in the Sub Saharan Africa. Scientifically, a host of mathematical and scientific tools and approaches may be utilized in making informed theoretical and more pragmatic decisions towards solving the lethal effect of this disease.

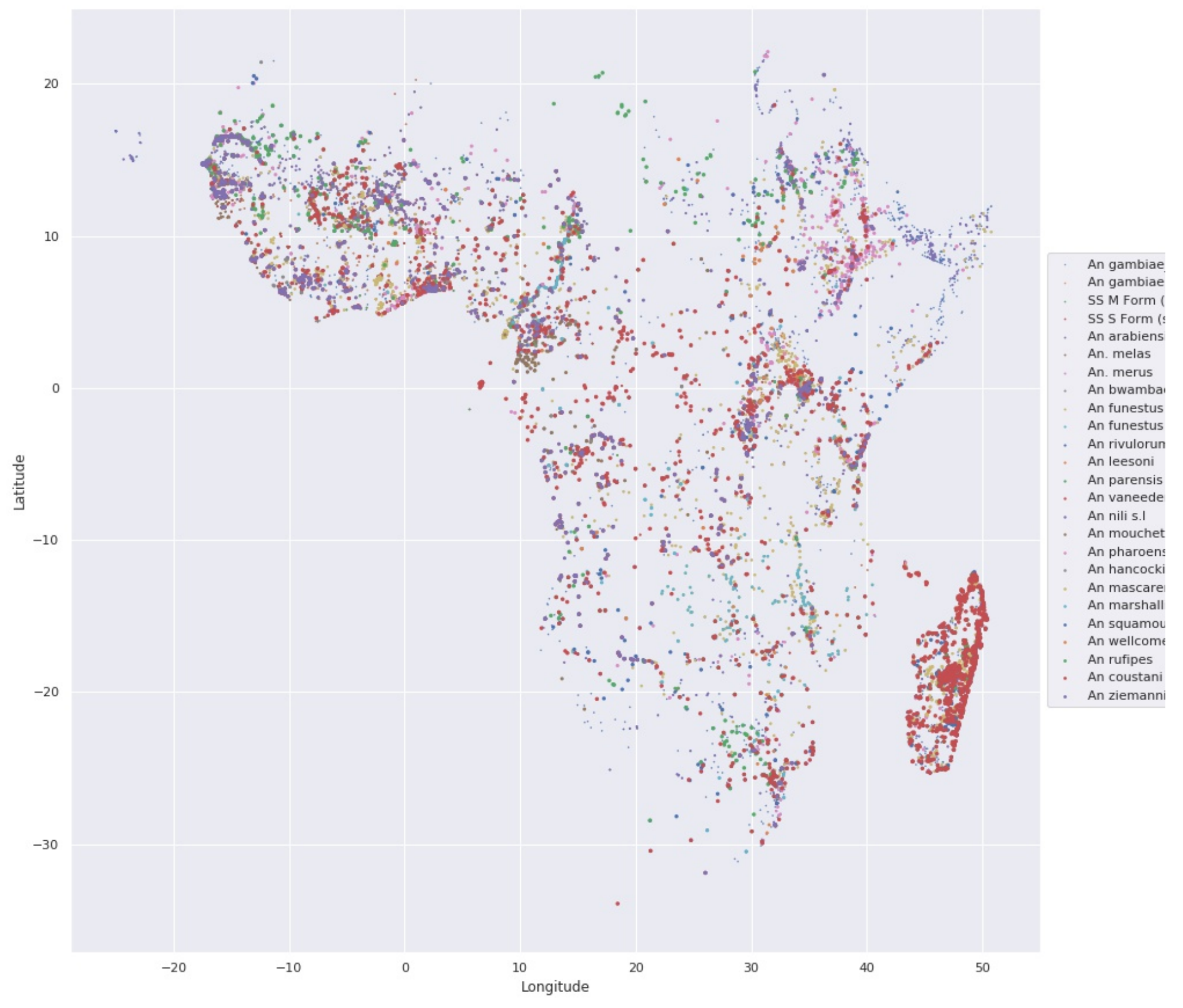

Figure 1: Geo-coded anophelines

Main Objectives

1. Binary classification of individual anopheline species given *Country*, *Region*, *Latitude*, *Longitude* and *Year*.
2. Multilabel classification of anopheline species given *Country*, *Region*, *Latitude*, *Longitude* and *Year*.

Data pre-processing and data exploration

- Record of 13,464 observations from 48 different countries in Africa with 41 features, since 1960.
- No description of the data, but Features include informations about location, time, 26 unique species and sampling methods.
- The time provided in the dataset had no relation with the time of invasion but, corresponded just to the sampling dates of mosquitoes.
- data cleaning and encoding before exploration

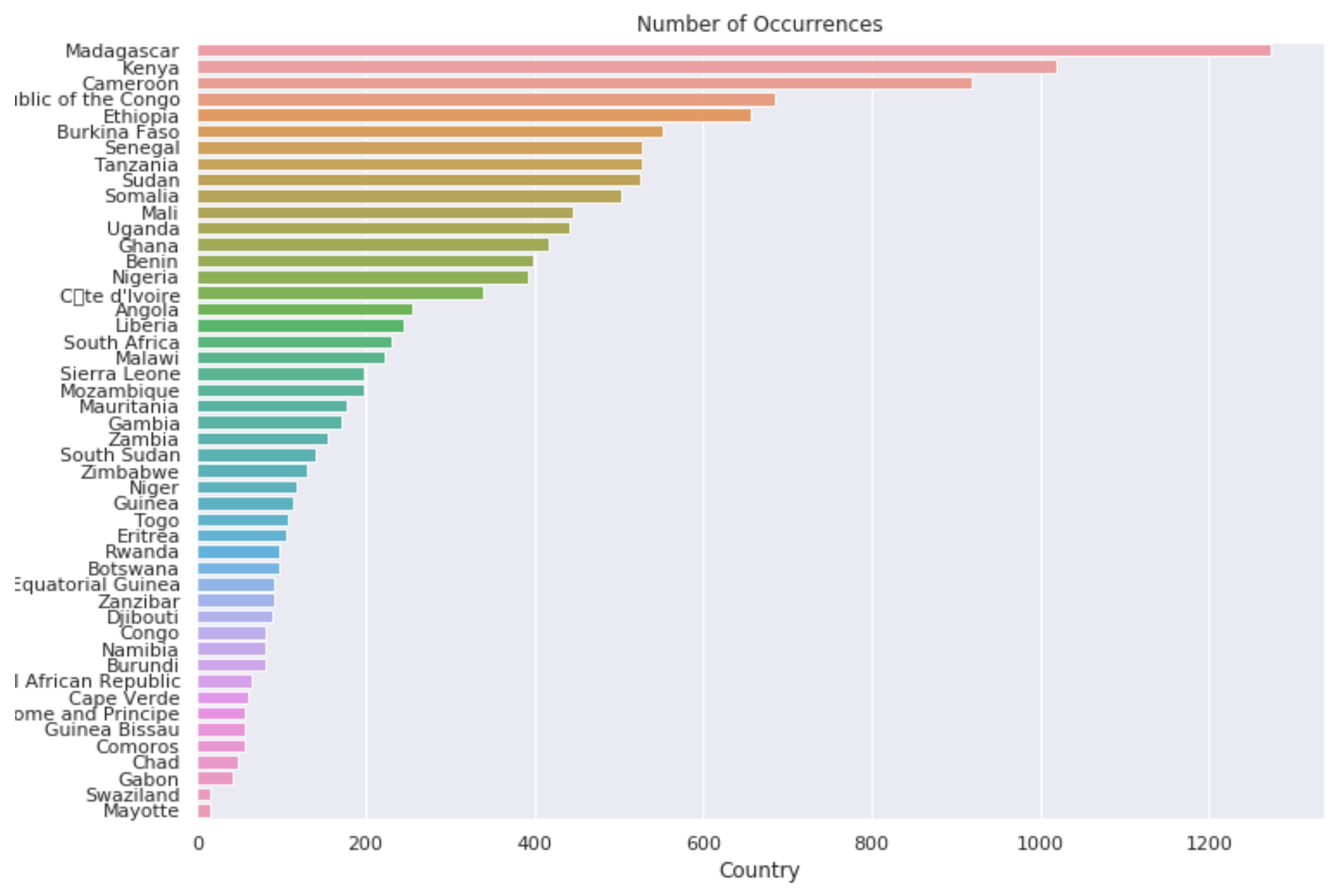

Figure 2: Number of occurences

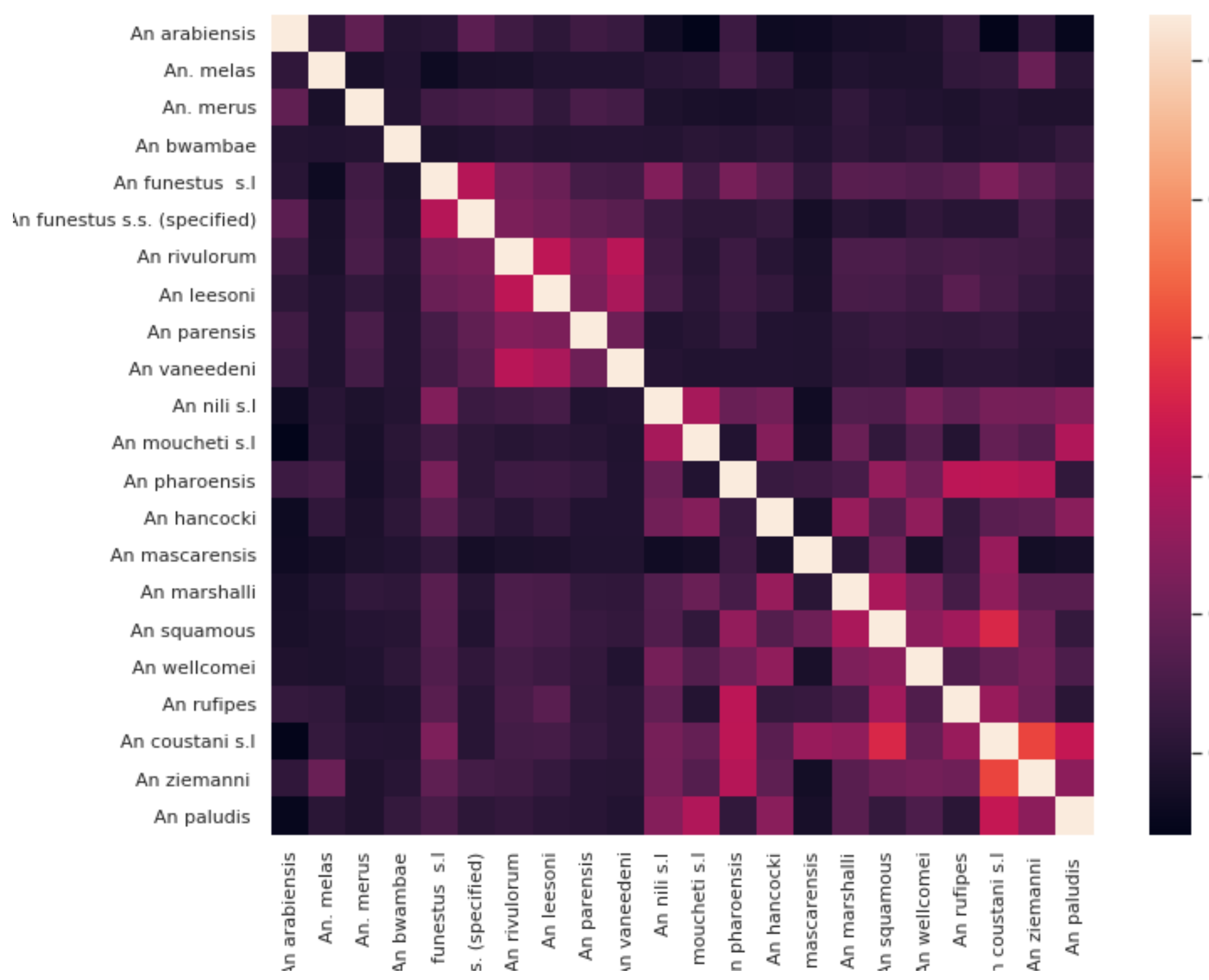

Figure 3: Correlation matrix of the species

Methods and results

Binary classification

Classifier	Avg Acc Score
L D A	91.181233
Logistic Regression	90.967333
Gaussian Naive Bayes	90.664879
Q D A	86.469166
Random Forest	74.255393

Table 1: mean of the accuracy score from individual species for various classifiers accuracy predictions

Classifier	Train time
L D A	6.66s
Logistic Regression	17.74s
Gaussian Naive Bayes	3.73s
Q D A	4.55s
Random Forest	18.60s

Table 2: Training time for various classifiers

Multilabel classification

Classifier	Avg Acc Score
Random Forest	94.221407
Decision Tree	93.443417
K Neighbors	94.221407
Logistic Regression	91.723072

Table 3: averages of the scores for each model used

Anopheline species	Precision	Recall	f1 score
An gaambiae complex	0.92	0.96	0.94
An gaambiae ss	0.83	0.80	0.82
An parensis	1.00	0.25	0.40
An bwambae	0.00	0.00	0.00

Table 4: Some classification reports

Model selection

- After cross-validation (20 iterations), Selection of the Random Forest Classifier because of the bias variance trade-off.
- The metric used for the training was Hamming loss $= \frac{1}{NL}\sum_{i=1}^{N}\sum_{j=1}^{L}\mathbb{1}_{\{\bar{y}_j^i \neq y_j^i\}}$. This function computes the ratio of all misclassified target labels of each species but can be very misleading. But evaluation done with F1 score.
- Hyper-parameter tuning: max depth 80, num of estimators 77, min samples split 3, min samples leaf 1, mx features, square root function.
- Training error: 0.452997% and test error: 5.590339%

Discussion

- Multilabel classification algorithms are computationally efficient in comparison to several binary classifications.
- High accuracy of the "multiple binary classification" of species due to the lack of correlation between species shown by the correlation matrix, although the biological correlation.
- Understanding the data set was challenging (misleading names, no meta data files, Lot of assumptions made during the modelling) More refining need to be done in the data set.

Conclusion

In order to accomplish the tasks defined in our objectives, we had to make a crucial assumption: each record in the dataset was a collapse of multiple records from different surveys into one record and work with those data as year of invasion. Within that assumption, we were able to:

- Learn to analyze a real world problem as well as implementing and conceptual solutions into codes.
- Lear n to interpret results and their feasibility and applicability

Forthcoming Research

- If the time of invasion is available, try our model with the selected parameters and see the results.
- Use our model to predict next invasion in Africa
- Extend our model to other countries in the world
- Retry training a sustainable neural network and compare the results

References

1] David Kyalo, Punam Amratia, Clara W. Mundia, Charles M. Mbogo, Maureen Coetzee, Robert W. Snow Version 1. Wellcome Open Res. 2017; 2: 57. (2017).
[2] Marc Deisenroth, Logistic Regression, Afriacn Masters of Machine In- telligence (2018).