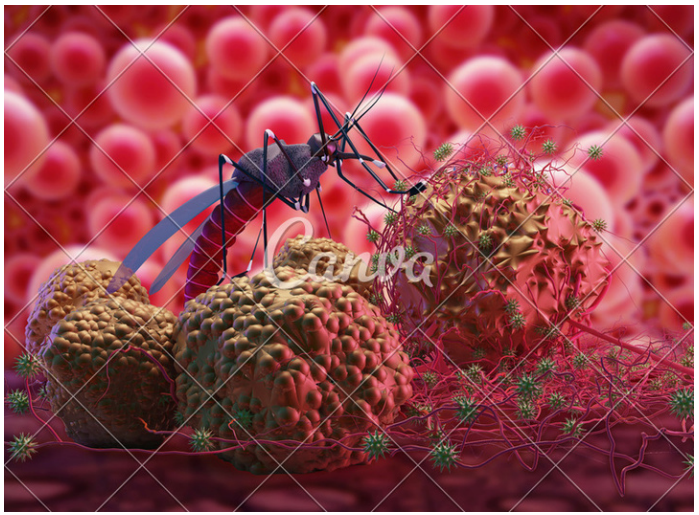
UCL
SCHOOL OF
MANAGEMENT

AI in healthcare :

Predicting malaria in Kenyan patients using ML algorithms

UCL



Malaria

OVERVIEW

- Species: female plasmodium mosquitoes

- Infectious disease: Spreads when an infected mosquito bites people

- Only 87 countries worldwide have yet to eradicate it

Context

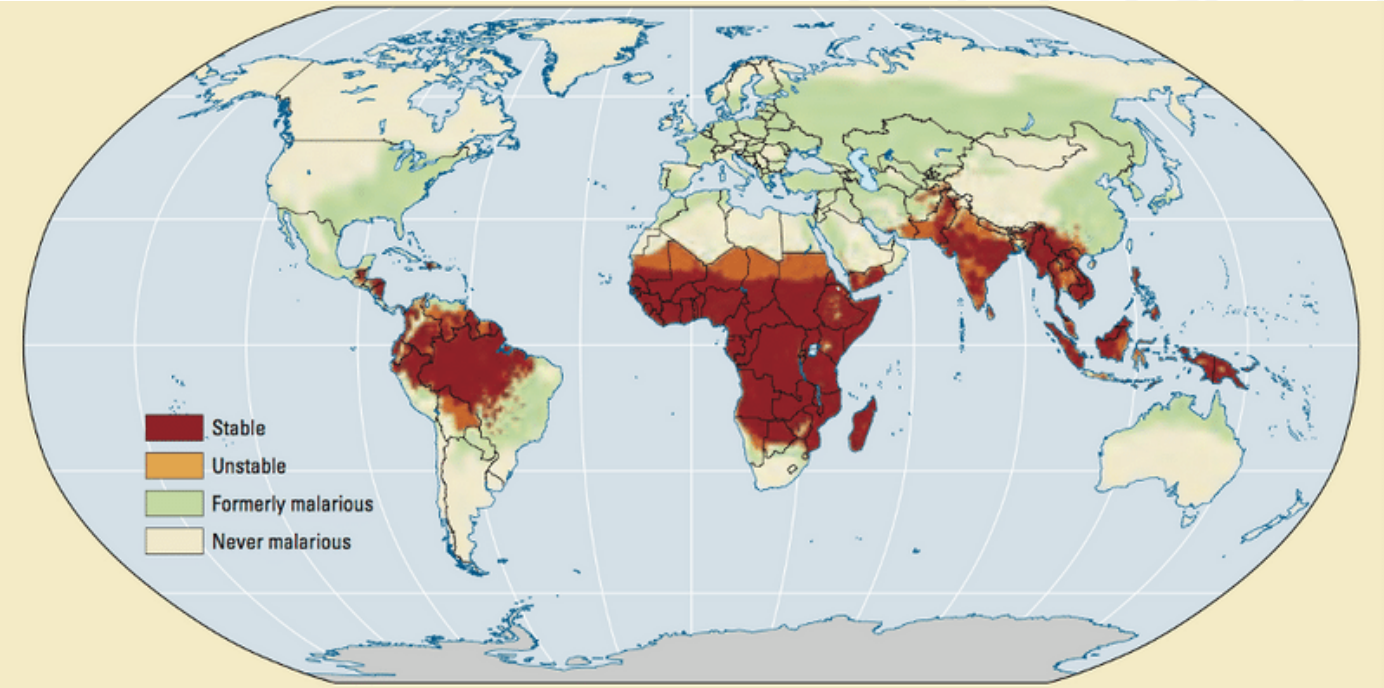
MALARIA IN KENYA

.

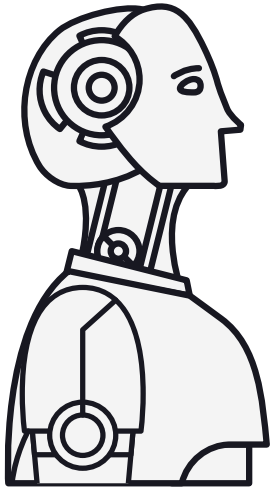
- 6 million cases & 5.1% moratility rate

- 15% to 20% of medical consultations are related to malaria in Kenya

- 90% of malaria deaths occur in Africa



Malaria incidence rate vary based on geographical location



Machine Learning

explanation

Purpose is to solve a problem through a set of mathematical instructions, following certain rules

---

Algorithms will learn from past data to predict classes of data + solve the problem again when new data points are added.

malaria is often diagnosed through RDTs, microscopy, or even images of malaria cells (deep learning). But rarely by supervised learners

Research objective

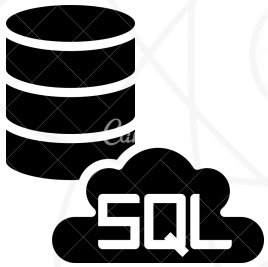
Question:

To what extent can supervised learning algorithms accurately diagnose the presence of the Malaria parasite in Kenyan patients?
Could this replace traditional medical diagnosis methods?

Aim:

Use ML algorithms to detect the presence of the parasite & gain insights on the disease for better prevention and treatment

Framework



- **The data**

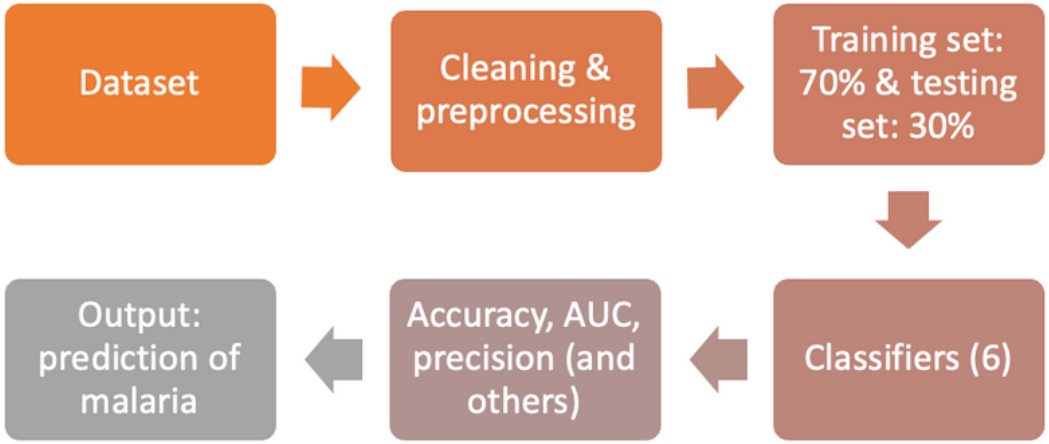
- from Carepay

⟶ a smart payment solution for healthcare

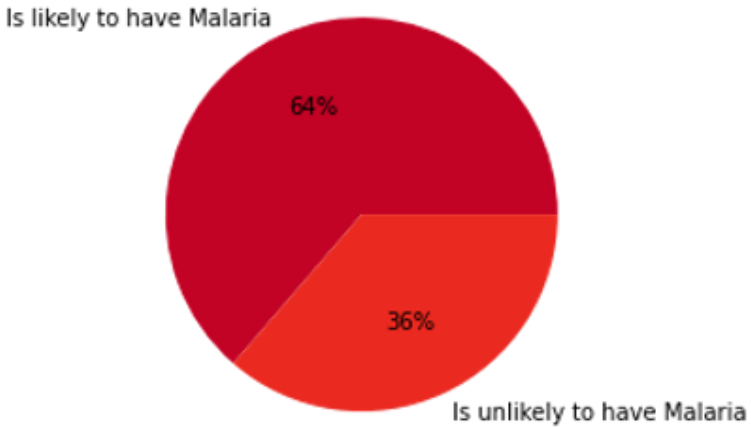
- Scope: Kenya, 2015-2022

- 19 columns: age, gender, symptoms, diagnosis, treatment, etc.

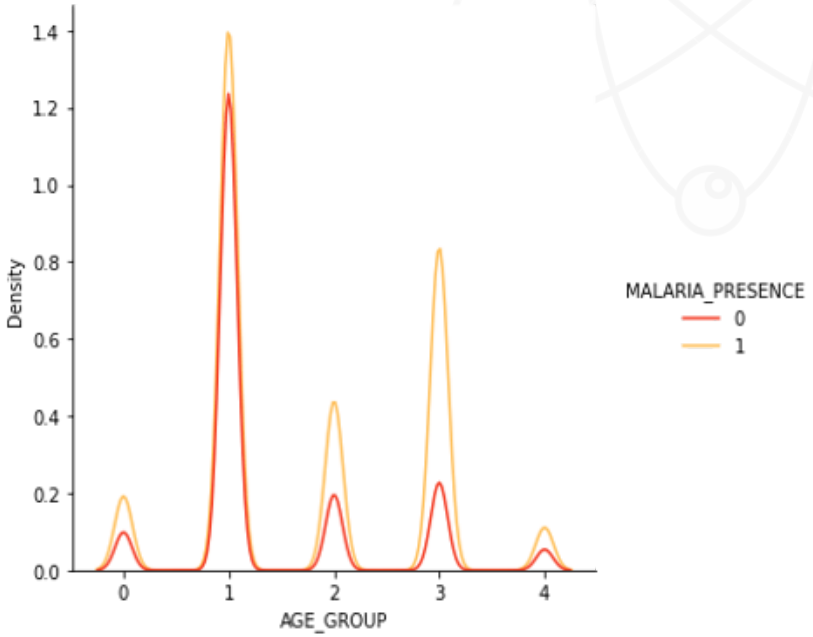
- 467,908 observations



EDA findings

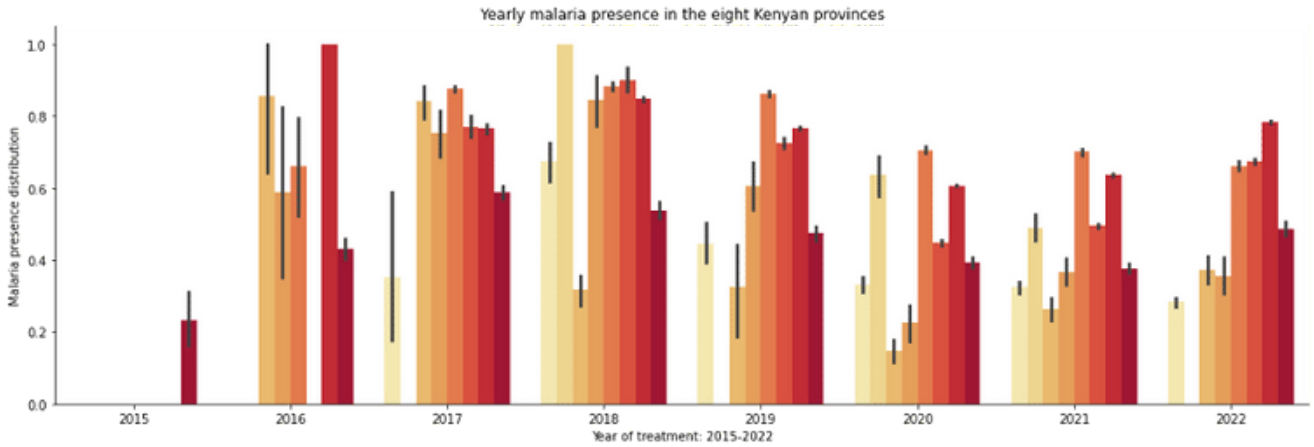


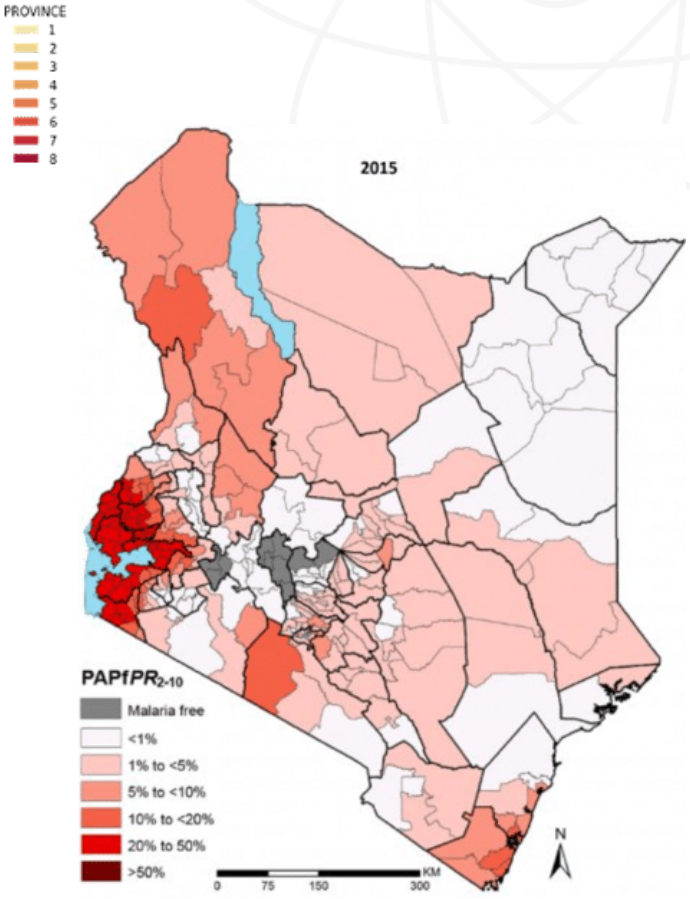
Who gets more sick?



key relationship 1:
age - malaria presence

EDA findings

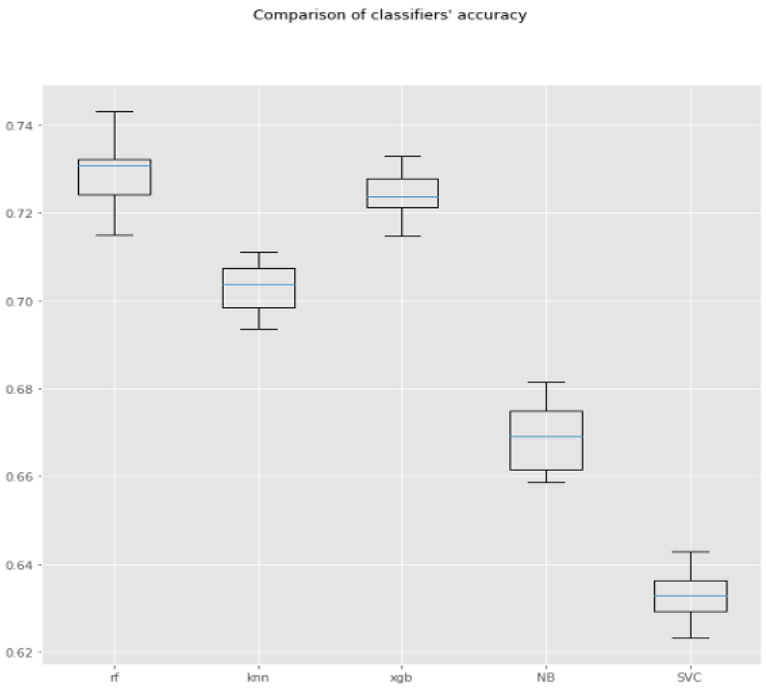




Key relationship 2:
geographical area - malaria presence

Research proves the main causes of malaria are climatic:
- rainfall
- altitude
- temperature
- humidity

ML models: performance



Threshold = 70% performance

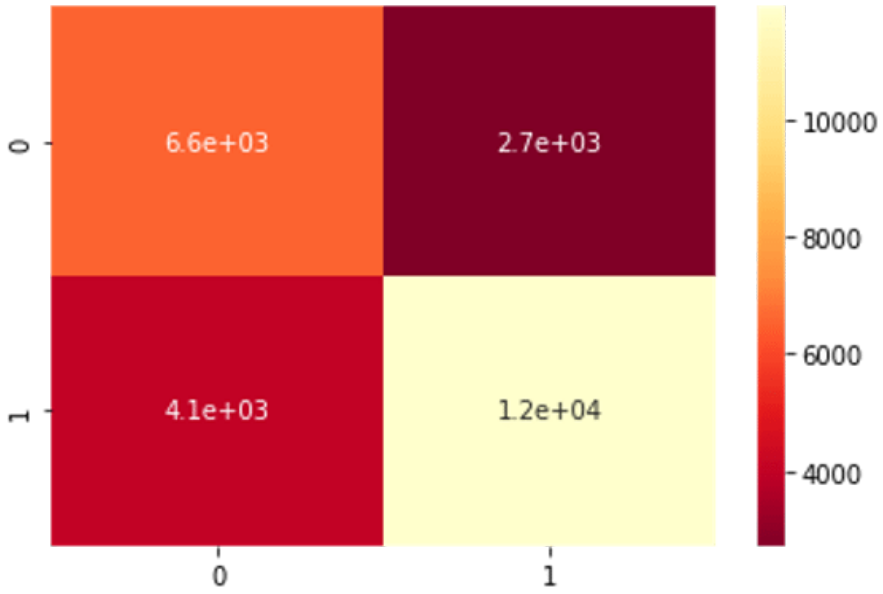
- Logistic regression is ruled out
- SVM is ruled out

Best performers:

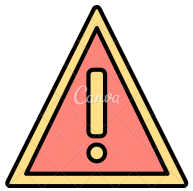
- Random forest
- XGboost

Model	Accuracy	Precision	AUC under ROC
KNN	70%	73%	77%
XGB	73%	64.9%	82%
RF	73%	76%	82%
CategoricalNB	70.7%	72.7%	77%

Problem: sensitivity and specificity



Despite Random Forest's high performance,
it struggles at detecting true positive cases
& true negative cases
LR had higher specificity, but lower accuracy



High specificity and sensitivity
are crucial in medical research

To sum up: there's a tradeoff

Classifiers

- Higher accuracy higher AUC

- Less FP and less FN but also less good at finding TP and TN

- Lower specificity & sensitivity

RDTs

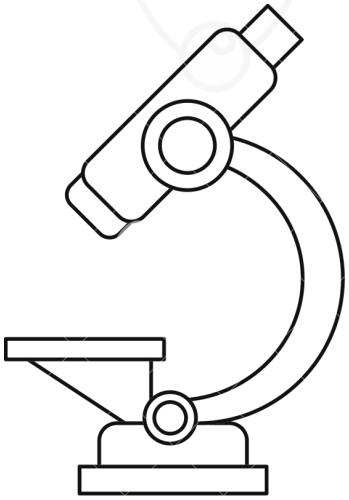
- Lower accuracy
- Better at finding TP and TN

- Higher specificity and sensitivity

Microscopy

- 100% accuracy

- Less accessible



Recommandations

- EDA showed malaria spreads differently under different parameters
  - *widen efforts to all age groups, even more in endemic regions*
- Combine RDTs with ML classifiers for increased performance
  - *Increase funding toward research on machine learning & vaccination*
- Keep prioritising preventive measures & prevent drug resistance

Prevention or treatment

TRADEOFF 2: POLICY MAKING

Preventive measures

WHO recommends **prevention** as a priority for governments:
- LLIN
- IRS
- preventive behavior

→ account for 60% of global investment in malaria control

Treating

- difference of cost between artemisinin & non artemisinin based **drugs**

↳ **cost vs effectiveness**

- FP may lead to treating negative cases ==> drug resistance

Limitations

- Study is biased : Carepay's data = unrepresentative sample of the Kenyan population

- Some medical consultations diagnosis are not accurate → reduces models accuracy + trustworthiness

Future works

to improve research results

1. Add more data: climatic factors as independent variables

2. Combine multiple classifiers for increased performance