Segmenting and Grouping Neighborhood Vulnerable to Ebola Virus Diseases

By

Michael Kumakech

Background

- According to World Health Organization (WHO, 10th February 2020) news on her website:
- Ebola virus disease (EVD) is transmitted to people from wild animals and spreads in the human population through human-to-human transmission.
- The average EVD case fatality rate is around 50%.
- Case fatality rates have varied from 25% to 90% in past outbreaks.
- Relapse-symptomatic illness in someone who has recovered from EVD due to increased replication of the virus in a specific site is a rare event.
- Studies of viral persistence indicate that in a small percentage of survivors, some body fluids may test positive.

Problem Statement

- Latitude and Longitude of location of countries hit by EVD are not known to International tourists and other visitors.
- Data that can be used to identify, subdivide and group similar or dissimilar species of Ebola Viruses outbreak in countries are Years of outbreak, Country, Ebola Virus Disease Species (EVD), Cases, Deaths, Case fatality.
- This project aimed at segmenting and clustering neighborhood vulnerable to Ebola viruses in the world map.

Significance of the Study

- Academia and other scientists may be motivated to do more research.
- Helps the country to get organizations that are willing to support the victims of Ebola viruses.
- International main bodies (WHO, UNESCO, UN) will have to get interested in identifying the locations of each of these species of Ebola viruses.

Interest and Data Source

- The most target audience are the community, tourists, students, international health workers and other workers in the location.
- Governmental and non-governmental organization care about this problem of Ebola Virus Disease outbreak in such location.

- Data was scrapped from WHO website: https://www.who.int/news-room/fact-sheets/detail/ebola-virus-disease .
- With feature selection, dependent variable was deaths and independent variables were cases, Zaire, Sudan, and Bundibugyo EVD types were used in the predictive modelling.

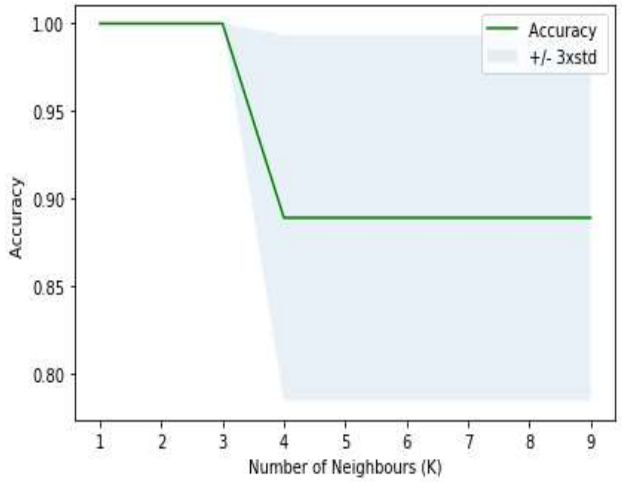
Methodology

- Descriptive Statistics
- Linear Regression
- Machine Learning Algorithms
- Support Vector Machine
  - Decision Tree
  - Logistic Regression
  - K-Nearest Neighbor.
- Location Data

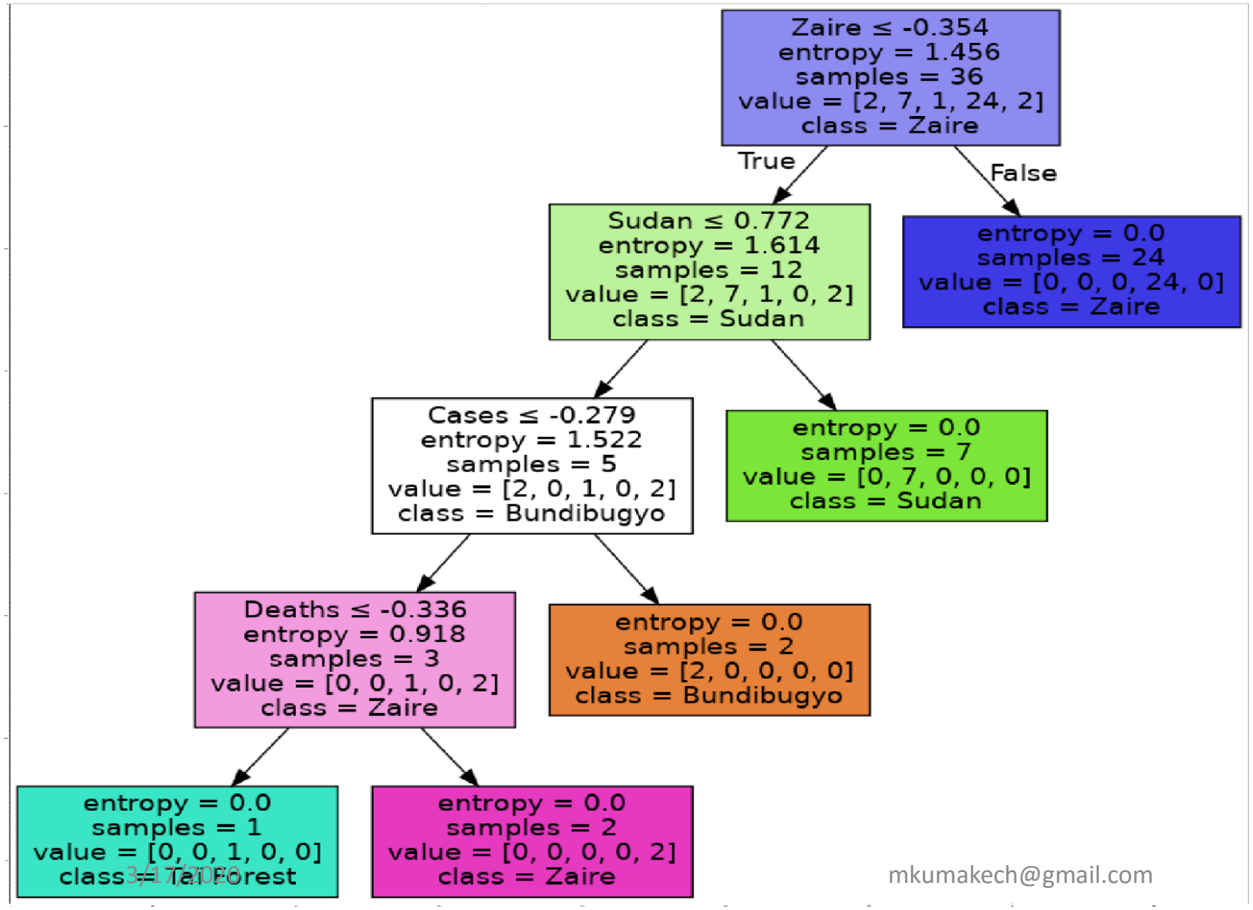
Results

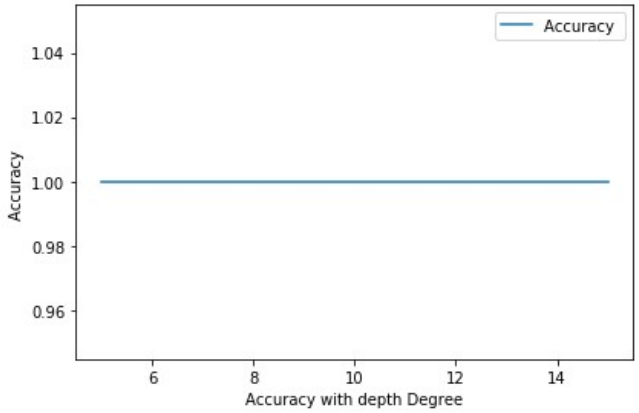
- **Descriptive Statistics:** Zaire EVD is top and commonest in most countries.
- **Linear Regression model**: $Deaths = 56 + 0.35 \times Cases.$
  - The R^2 for Deaths and was found to be 91.3%.
- **K-Nearest Neighbor (KNN) Algorithm**

Set 1: 75% Training and 25%Test data set			Set 2: 80% Training and 20%Test data set		
Best Accuracy	Value of K	F1 Score	Best Accuracy	K Value	F1 Score
0.888	1	0.63	1.0	1	1.0



Results- Decision Tree (DT) Algorithm

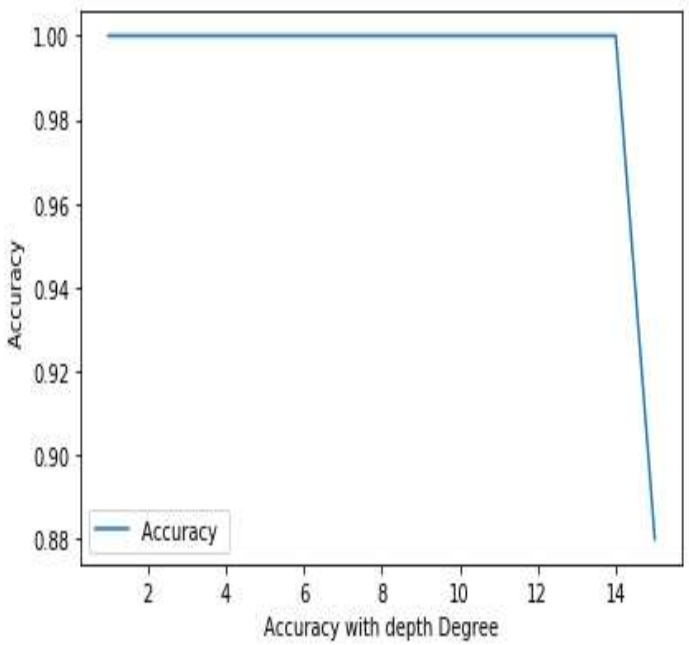




constant accuracy of 1.0 with F1 of 1.0.

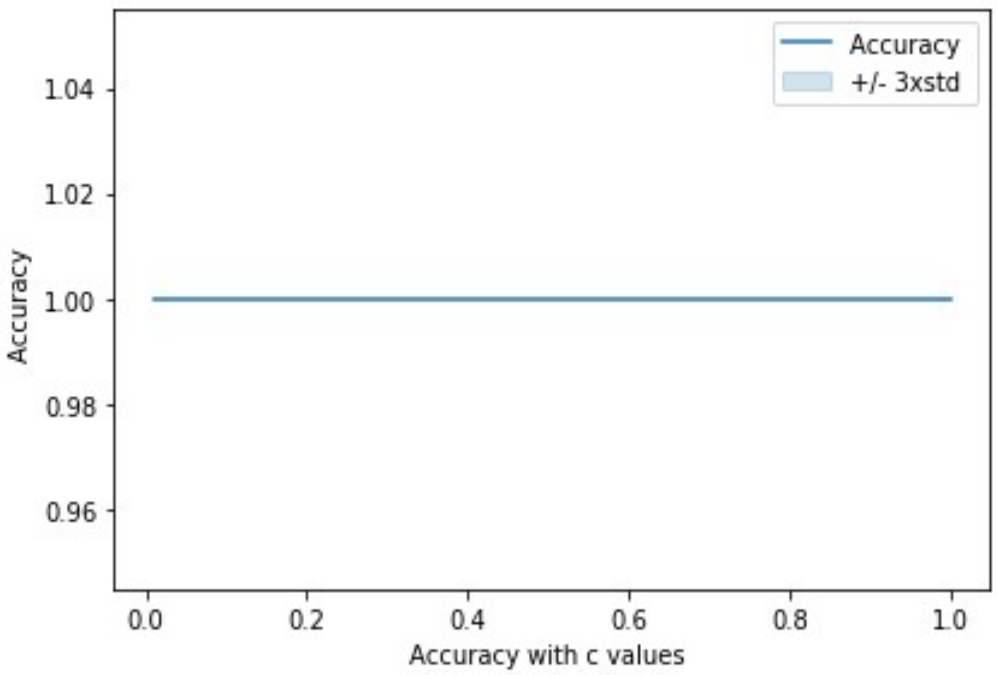
Results - Support Vector Machine (SVM) Algorithm

Set 1: 75% Training and 25%Test data set			Set 2: 80% Training and 20%Test data set		
Best Accuracy	Degree	F1 Score	Best Accuracy	Degree	F1 Score
0.972	1	0.7959	1.0	1	1.0



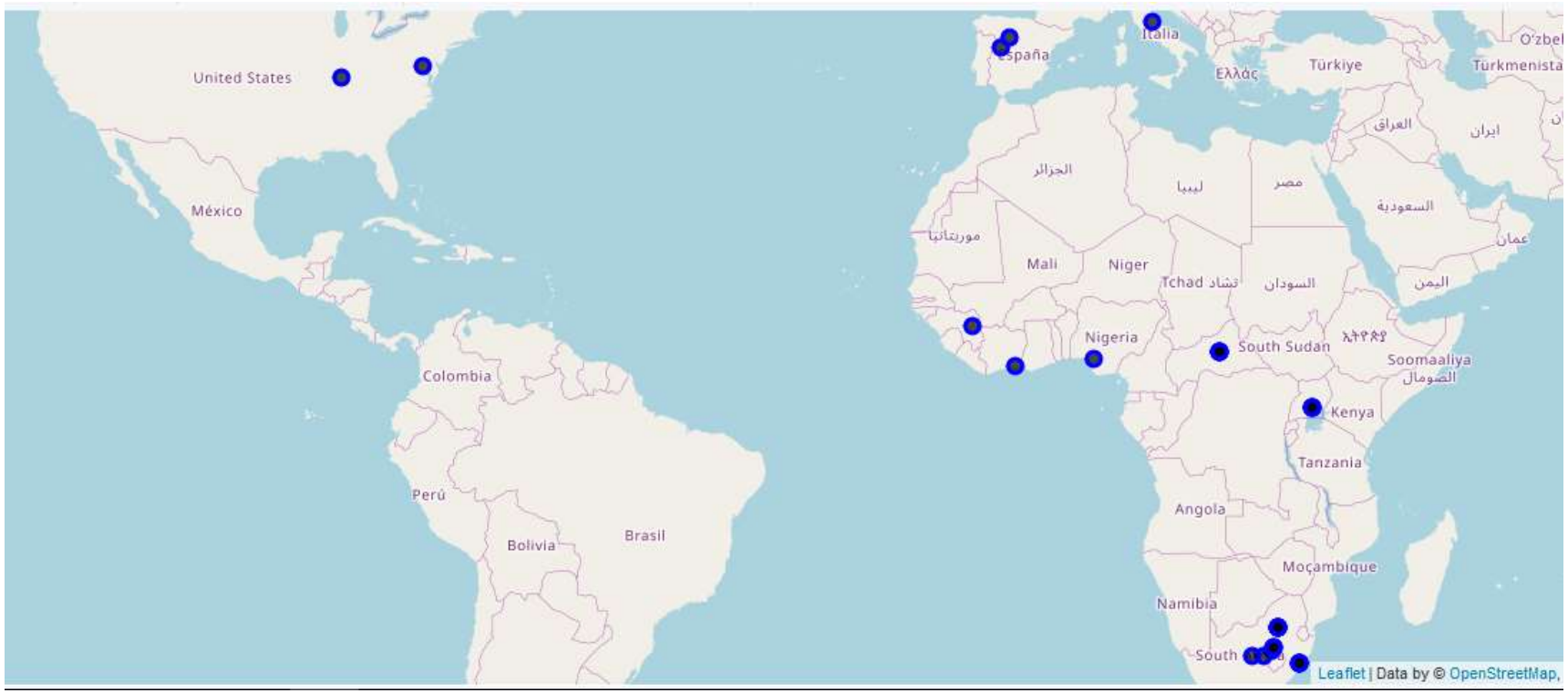
In Set 2 the researcher observed that the accuracy of the model remains constant at 1.0 with F1 score of 1.0 with give gives correct prediction of deaths once cases of EVD are registered.

Results - Logistic Regression Algorithm



- With either 75% or 80% with testing data set. Also either 25% or 20% . Result showed constant accuracy of 1.0, F1 Score of 1.0 and C values ranging from 0.01-1.0.
- Indicating that probability of deaths is certain when cases of EVD is registered in a country.

Results - Location Data hit by EVD



Evaluation & Conclusion

	Jaccard	F1-score	LogLoss
KNN	0.888889	0.636364	NA
Decision Tree	1.000000	1.000000	NA
SVM	0.972222	0.795918	NA
LogisticRegression	1.000000	1.000000	0.281665

- Zaire EVD was found to be the most dangerous types of Ebola which has affected most of the countries of the world.
- The top most affected county is Democratic Republic of Congo in 2014.
- DT and Logistic Regression Machine learning algorithms models perform excellently well in the prediction.
- Tourists and vistors are be aware of the danger of Ebola disease and Location epecially in Africa.

Future Work

- Use of Foursquare is necessary to zero the cases to regional, district and towns' level so that K- means algorithm may be used to visualize the data into smaller clusters.
- Currently, Data about Ebola Virus Disease seem not to be available in the Foursquare database.

- **Thank You**
- **Michael Kumakech +256706290029**