Segmenting and Grouping Neighborhood Vulnerable to Ebola Virus Diseases

Michael Kumakech

(+ 256 706290029, mkumakech@gmail.com)

(March 2020)

Abstract

Most time it is very difficult for international tourists and others visitors to locate the countries vulnerable to Ebola Virus Disease (EVD). Support Vector Machine, Logistic Regression, Decision Tree, K- Nearest Neighbor Machine Learning (ML) algorithms were used to model EVD dataset from 1976 to 2018. The findings indicated that Zaire EVD is the most dangerous one found in many of the countries hit by the disease and that all EVD cases reported had high probability of deaths.

1. Introduction

This Section present a description of the problem and a discussion of the background.

1.1 Background

According to World Health Organization (WHO, 10th February 2020) news on her website:

- Ebola virus disease (EVD), formerly known as Ebola haemorrhagic fever, is a rare but severe, often fatal illness in humans. The virus 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, but has been documented. Reasons for this phenomenon are not yet fully understood. Studies of viral persistence indicate that in a small percentage of survivors, some body fluids may test positive on reverse transcriptase polymerase chain reaction (RT-PCR) testing for Ebola virus for longer than 9 months.

International tourists and other business people are unaware of the real locations of the six species of Ebola Viruses which has been identified as: Zaire, Bundibugyo, Sudan, Taï Forest, Reston and Bombali. Therefore it’s of merits to segment and group the location of these Ebola Virus species.

1.2 Problem Statement

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, Latitude and Longitude of location so that International tourists and business people are certain about real location of EVD in the world map. This project aimed at segmenting and clustering neighborhood vulnerable to Ebola viruses in the world map.

1.3 Significance of the Study

International main bodies will have to get interested in identifying the locations of each of these species of Ebola viruses.

Academia and other scientists may be motivated to do more research to identify why these Ebola species are common in such location and devise appropriate solutions to mitigate the Ebola outbreak in such neighborhood.

Helps the country to get organizations that are willing to support the victims of Ebola viruses and devise methods of controlling the outbreak of the disease through sensitization of the community.

1.4 Interest

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.

2. Data Acquisition and Cleaning

This section entails description of the data and how it will be used to solve the problem.

2.1. Data Source

Data was scrapped from WHO website: https://www.who.int/news-room/fact-sheets/detail/ebola-virus-disease . The table contains records with the following columns: Years of outbreak, Country, Ebola Virus Disease Species (EVD), Cases, Deaths and Case fatality. However, since this project deals with location data, the researcher used Foursquare data to locate the latitude and longitude of each country to be used in segmenting and clustering of neighborhood vulnerable to Ebola Virus diseases.

2.2. Data Cleaning

Data was downloaded from the WHO website and scrapped with detailed records of Ebola Virus Disease outbreak from 1976 to 2019. The dataset however, includes the Year, Country, Ebola Virus Disease (EVD) types, cases, deaths and case fatality recorded. The researcher used records from 1976 to 2018 since the records for 2019 was ongoing.

These were some of the problems with the dataset:

- The record for 2018-2019 for Democratic Republic of the Congo with EDV Zaire type cases were ongoing. Thus, the record was dropped out.
- The 2014 to 2016 EVD records for Zaire cases and deaths recorded for Sierra Leone and Liberia were in string data type. Thus, the string data types were changed to number values.

- Some records for Democratic Republic of the Congo Ebola cases, deaths and case fatality were not indicated in the table. Also, these row was dropped too.

2.3 Feature Selection

With feature selection, dependent variable was deaths and independent variables were cases, Zaire, Sudan, and Bundibugyo EVD types were used in the predictive modelling.

3. Methodology

Here statistical methods were employed to explore the data, create model development in to the data insights and used Machine Learning Algorithms to predict the deaths associated with the EVD cases.

3.1 Exploratory Data Analysis

Descriptive statistics and linear regression model were used in exploring the dataset.

3.3 Normalization of Data

Normalization of data was done using data standardization and processing techniques.

3.4 Machine Learning (ML) Algorithms

K-Nearest Neighbor, Support Vector Machine, Logistic Regression and Decision Tree algorithms were used in the prediction of deaths reported were cases of various types of EVD. Thus, this is a classification problem since for example the case can be for Zaire type of EVD or Not. Also, these types of Machine Learning algorithms have been commonly used in classification prediction problems.

3.5 Location Data

The researcher used geocode packages to get the latitude and longitude for each rows of the data frame for each country with EVD cases. This location data was used to come up with location which could be vulnerable to EVD.

4 Results and Discussion

Descriptive statistics and linear regression were used to explore the data set. Also, the following Machine Learning (ML) algorithms were used: K Nearest Neighbor (KNN), Decision Tree, Support Vector Machine and Logistic Regression to come up with best classifiers. Most of the Python libraries were used in the simulation of results (Wes McKinney 2018). The observations are entailed here.

4.1 Descriptive Statistics

out of 36 record of cleaned dataset, 24 were of Zaire type one, 7 for the Sudan, 2 are for Bundibugyo, 2 for Zaire type two and 1 for Taï Forest types of EVDs which affected the community greatly since 1976. Thus, Zaire EVD types were mainly in Democratic Republic of Congo and also was cited in Italy, UK, Spain and USA as well as some few West African countries. It appeared to be the top dangerous type in 2014. The Sudan EVD type was common in Uganda and Sudan only. The Bundibugyo EVD type however, was found in Uganda and Democratic Republic of Congo. Finally the Taï Forest EVD type was seen once in Côte d'Ivoire.

4.2 Linear Regression Model

In model development the researcher fit a linear regression model using the Cases feature and calculate the R^2 for Deaths and was found to be 91.3%. However the final estimated linear model of relationship between Cases and Deaths identified is expressed as:

$$\boldsymbol{Deaths = 56 + 0.35 \times Cases}$$

This implies that numbers of deaths associated with EVD is directly proportional to registered cases and people can die at a very faster rate if EVD is not control.

4.3 K-Nearest Neighbor (KNN) Algorithm

In order to come up with the best classifier, the researcher divided the data set to 75% as training set and 25% as test set and finally in 80% as training set and 20% as test set.

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

Here is a graphical representation of Set 1 and Set 2 classifier of KNN algorithms in figures and the researcher observed that the classifier performed well in Set 2.

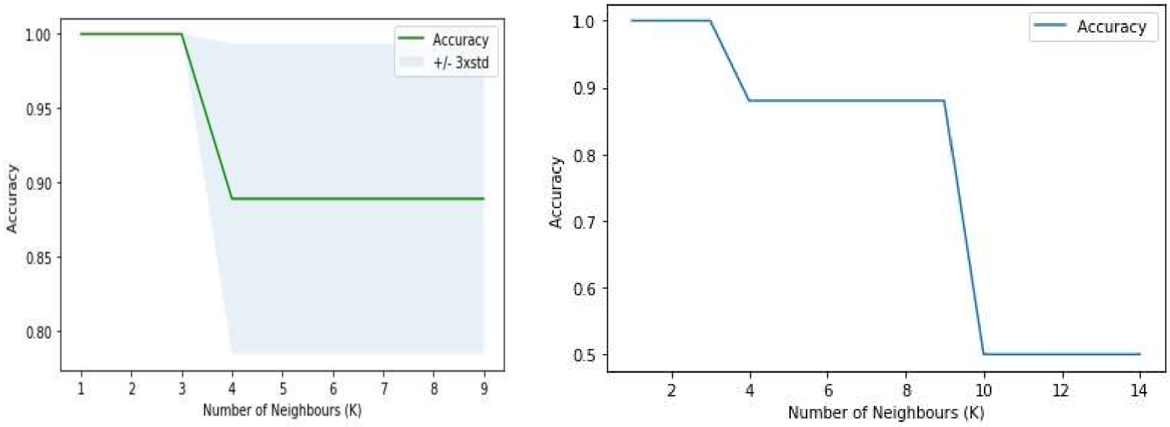


Fig. 1 Showing KNN classifier at different percentages of Training and Test set

4.4 Decision Tree (DT) Algorithm

In Decision Tree algorithm (IBM Data Science, 2019), the researcher observed that when you split dataset into 75% as Training set and 25% as Test set or 80% as Training set and 20% as Testing set, the classifier perform normally in similar way with accurate value of 1.0 and F1 score of also 1.0.

Here is a representation of decision tree in Figure 2 showing spreads of different EVD in the world. Thus, the most deadly EVD is the Zaire type followed by the Sudan one.

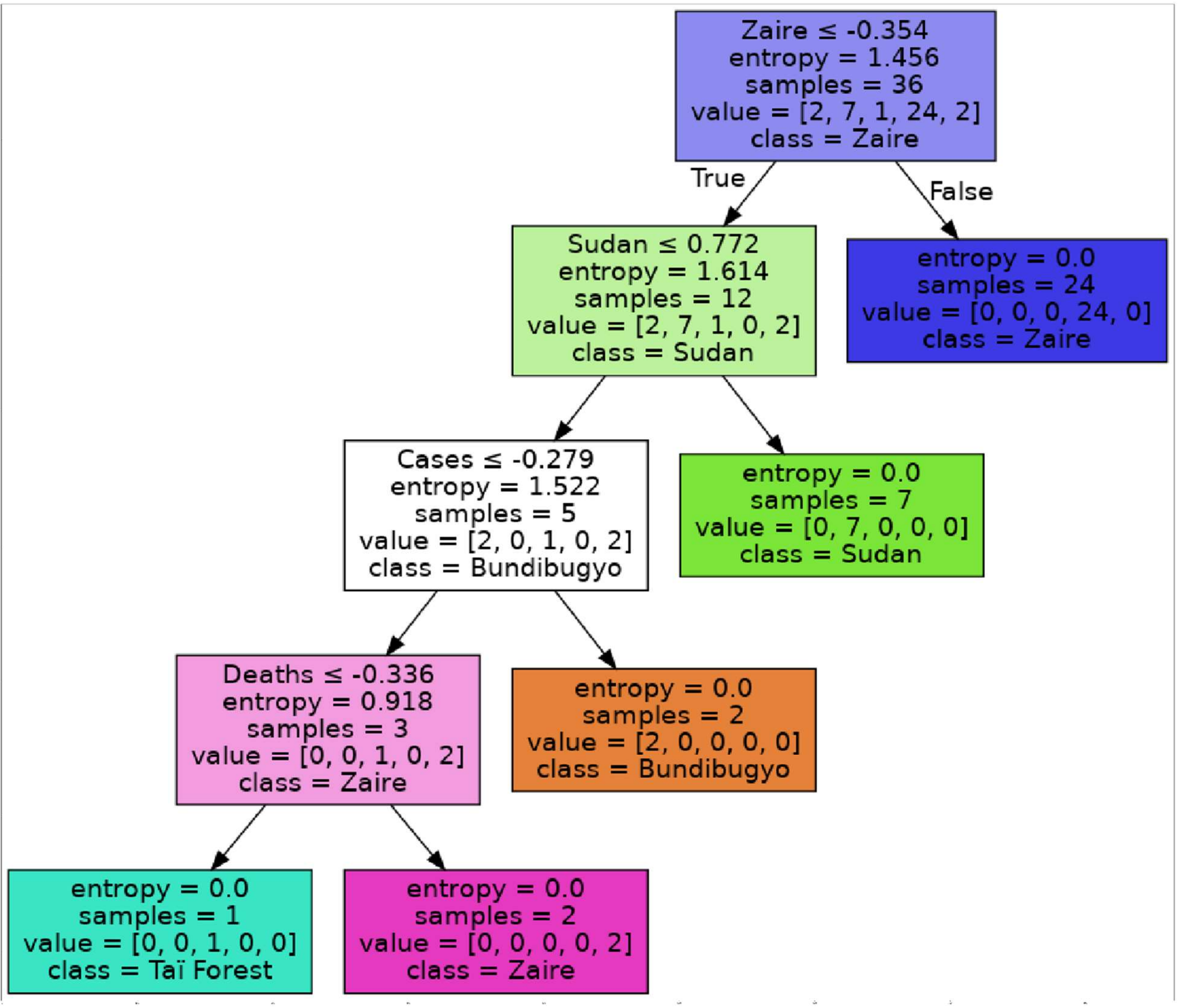
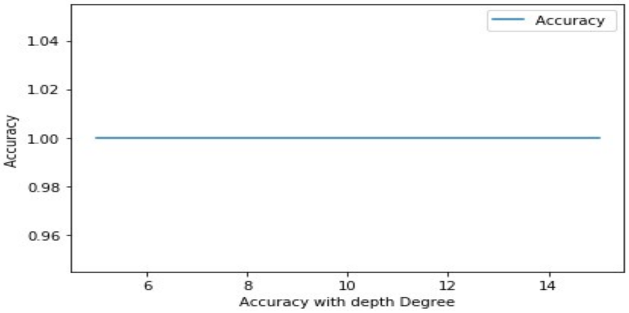


Figure 2 showing spreads of different EVD in the world.



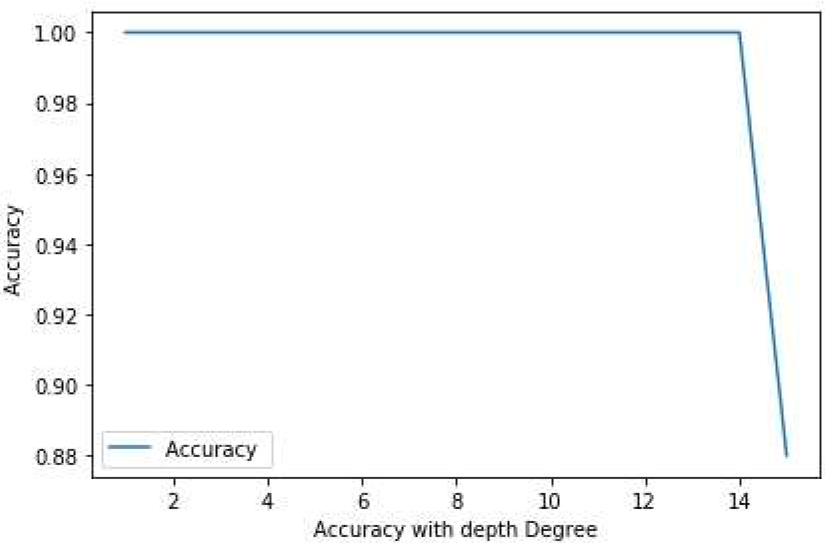
Also, when accuracy values are set from depth 5 -15, you get constant accuracy of 1.0 with F1 of 1.0. Here the performance of the classifier is better than in set 1.

4.5 Support Vector Machine (SVM) Algorithm

Using Support Vector Machine algorithms, the researcher found that with 80% training set and 25% test set produce better classifier than with 75% training data set and 25% testing data set. However, the second set provides better prediction results. See the summary in the table:

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

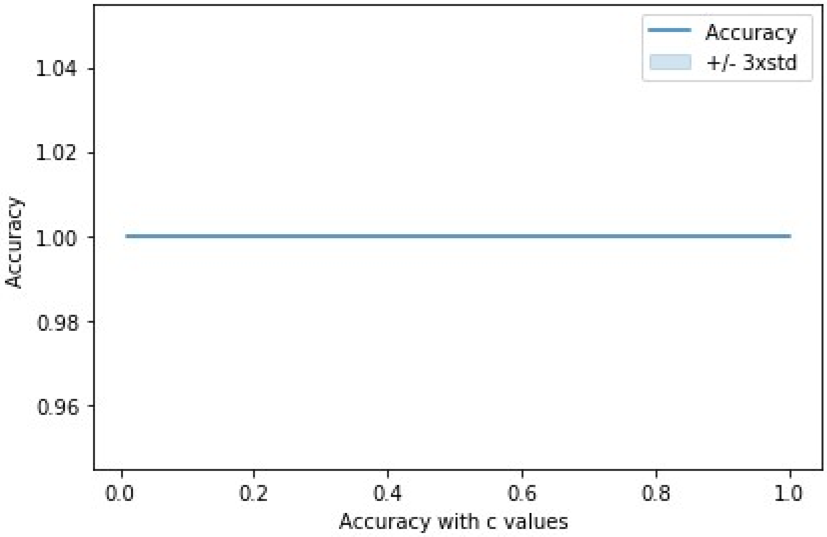
Here in set 2, if you adjust the degree of accuracy from 1 to 15, the researcher observed that the accuracy of the model remains constant at 1.0 with F1 score of 1.0 with give gives correct prediction of deaths once cases of EVD are registered.



4.6. Logistic Regression Algorithm

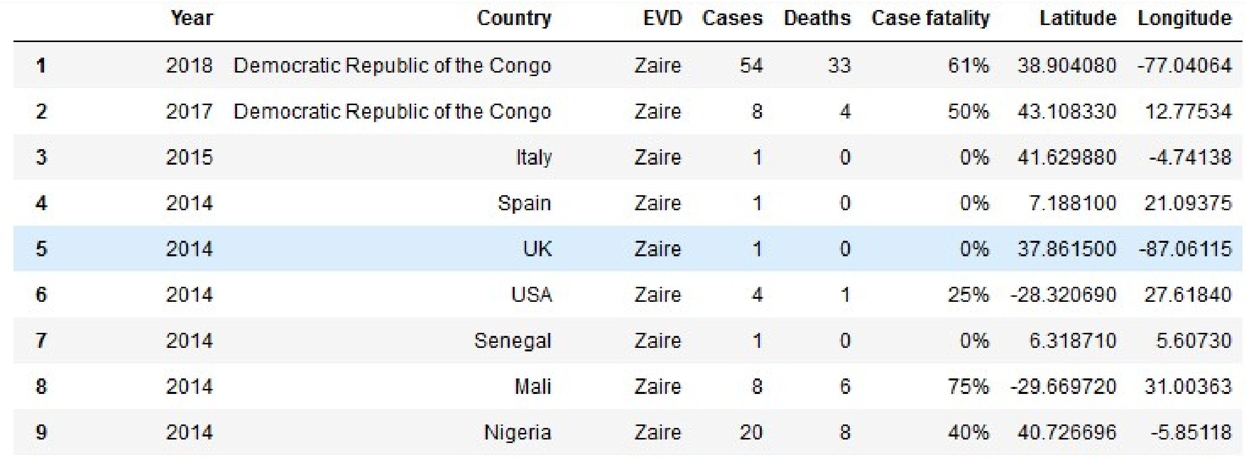
With Logistic Regression algorithm, the classifier performed excellently well in both separated percentages of training of either 75% or 80% with testing data set of also either 25% or 20% respectively with constant accuracy of 1.0, F1 Score of 1.0 and C values ranging from 0.01-

1.0. These values are indicating that probability of deaths is certain when cases of EVD is registered in a country as seen in the figure.



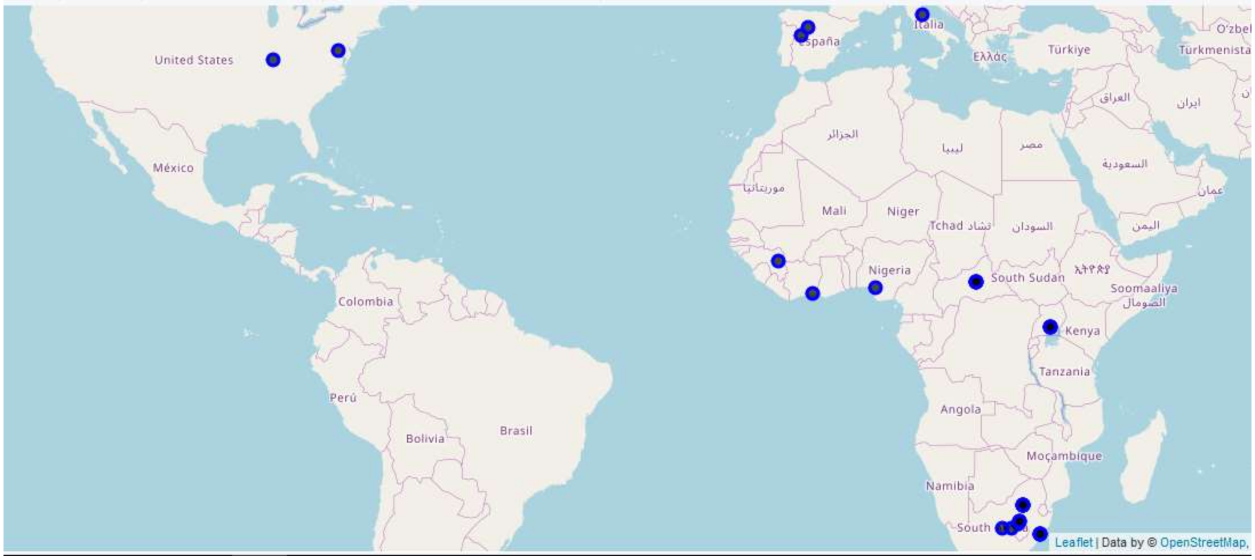
4.7 Location Data

It was significantly important to get the location data which include the geographical location such as Latitudes and Longitude of countries where EVD affected. Thus, the latitudes and longitudes were then got and put on data frame as shown in the table:



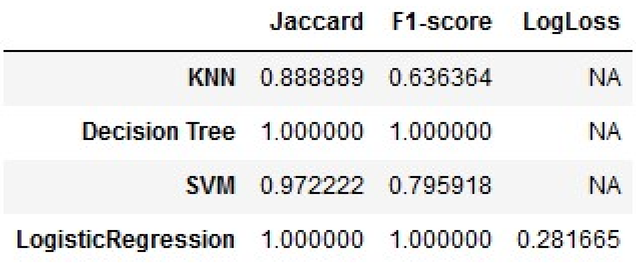
	Year	Country	EVD	Cases	Deaths	Case fatality	Latitude	Longitude
1	2018	Democratic Republic of the Congo	Zaire	54	33	61%	38.904080	-77.04064
2	2017	Democratic Republic of the Congo	Zaire	8	4	50%	43.108330	12.77534
3	2015	Italy	Zaire	1	0	0%	41.629880	-4.74138
4	2014	Spain	Zaire	1	0	0%	7.188100	21.09375
5	2014	UK	Zaire	1	0	0%	37.861500	-87.06115
6	2014	USA	Zaire	4	1	25%	-28.320690	27.61840
7	2014	Senegal	Zaire	1	0	0%	6.318710	5.60730
8	2014	Mali	Zaire	8	6	75%	-29.669720	31.00363
9	2014	Nigeria	Zaire	20	8	40%	40.726696	-5.85118

This location data was able to help the researcher to identify and locate the countries where EVD hit since 1967 as shown in the map of the world below. The blue shaded areas indicated the countries mostly affected with the disease.



4.8 Evaluation

Jaccard index, F1-scores and LogLoss were the performance metrics used for evaluating the accuracy of the models. When 25% of the dataset was used for evaluation, KNN, Decision Tree, SVM and Logistic regression performed as summarized in the table. Observations showed that Logistic Regression and Decision Tree algorithms outperformed KNN and SVM algorithms.



	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

However, when 20% of dataset was used for testing the jaccard index and F1-scores for these algorithms were all 1.0 indicating the accuracy of the models are excellent.

5 Conclusion

Zaire EVD was found to be the most dangerous types of Ebola which has affected most of the countries of the world: USA, Spain, Italy, Uganda, Democratic Republic of Congo and most of the West African countries. The top most affected county is Democratic Republic of Congo in 2014. The rate of reported cases is linear related with the deaths reported and all the four classification Machine learning algorithms models performs excellently well in prediction of deaths. Therefore the government, non-governmental organization, tourists, health workers, students, business men and community among others need to be aware of the danger of Ebola disease.

6 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.

References

1. *WHO (10th Feb. 2020),* *https://www.who.int/news-room/fact-sheets/detail/ebola-virus-disease.*
2. *https://eu-gb.dataplatform.cloud.ibm.com*
3. *Wes McKinney, Python for Data Analysis, 3nd Edn. 2018.*