International Journal of Engineering Research and Technology. ISSN 0974-3154, Volume 13, Number 6 (2020), pp. 1258-1265
© International Research Publication House. https://dx.doi.org/10.37624/IJERT/13.6.2020.1258-1265

Predicting Violent Crime Occurrence: An Evaluation of Decision Tree Model

Falade Adesola[1], Ambrose Azeta[2], Aderonke Oni[3], Felix Chidozie[4], Victor Azeta[5]

[1,2,3]Department of Computer and Information Sciences Covenant University, Ota Nigeria .
[4]Department of Political Science and International Relations, Covenant University, Ota, Nigeria.

[5]National Productivity Center, Calabar.

Abstract

Crime rate tends to be on the increase across the globe, and crime data analysis becomes imperative to aid predictive policing in tackling incidence of crime. In this paper data mining approach was applied to violent crime dataset for predicting next occurrence of violent crime. Previous researchers have used different supervised learning algorithms for crime prediction with accuracy results left to be improved upon. Consequently, this study particularly apply decision tree C5.0 algorithm on violent crime dataset in order to determine the probability of next occurrence of violent crime in Lagos metropolis. The data used was derived from Nigerian Police statistic department Obalende Lagos, pre-processed and applied on decision tree model built. The model was evaluated using the six violent crime types (murder, arm robbery, kidnapping, rape, non-negligent assault and man slaughter) dataset. The results obtained were evaluated using confusion matric and found to return an accuracy of 76.4% (percent). Based on this result, the model could be used by the Police authority to strategize and plan towards mitigating crime rate in the country.

Keywords: Decision Tree, Confusion matrix, Data Mining, Spatiotemporal, Supervised Learning, Machine Learning

1. INTRODUCTION

Crime has been an enemy of society from time immemorial that any responsible government must fight to a standstill. Urban development all over the world has introduced challenges such as housing problems as well as traffic and environmental problems which have resulted to rapid growth in population and this in turn has led to the increase in crime rate.

Additionally, crimes in urban areas have had negative effect on the lives of the populace and their properties. Hence, it is the government's constitutional responsibility to create a safer and ambient urban environment for everyone. The data available from the National Bureau of Statistics [1] shows that, over the last 30 years, crime rate have been on the increase by 3.4% on a yearly average. This is clearly worrisome and calls for immediate attention. Due to the increase in these dastard acts: murder, kidnapping, armed robbery, and rape, people's anxiety have increased tremendously; people can no longer sleep with their two eyes closed.

Crime prediction research has received a lot of attention in the past ([2], [3], [4], [5], [6]) because of its enormous attendance benefits to the country and citizenry at large. Through the use of Machine learning approaches, insights and prediction could be derived from the previously collected crime dataset to determine the next location of likely occurrence of violent crime. This will go a long way in helping Police in planning efficiently using their limited available resources in cubing the growing crime rate in the country. Crime prevention by strengthening Police patrols could be very costly in terms of human resources and finances. Therefore using this state-of-the-art method could show that Police patrols will only be undertaken depending on the location of predicted crime hot spots.

2. RELATED WORK

[7], applied Principal Component Analysis on the crime dataset collected from Katsina state Nigerian police in their research study to extrapolate crime rate, patterns in various hotspot areas in Katsina state. The results were made available to the Police authority for making proactive decisions and actions, but the approach is not efficient enough, as it takes time to get results and to make predictions.

In a bid at reducing sexual crime and criminalities in Korea, [8] developed an intelligent crime prevention system (ICPS) by collecting and analyzing crime data using Term Document and Inverse Document Frequency (TD-IDF) and Naïve Bayesian algorithm. The results from the system developed were then fed into IoT devices and sensors to predict dangerous areas so as to alert a woman with a wearable device when she passes a hotspot zone. The shortcoming of the system is that as the crime data becomes huge, noticeable reduction in overall performance of the system was noticed.

K-means algorithm data mining approach was applied by [9], in a paper titled "Machine Learning Approaches for Detecting Crime Patterns". The model developed was able to detect crime pattern detections and then make prediction on the likelihood of future crime occurrence. The crime dataset was clustered into various subsets and similarities for making prediction of crime occurrence. However, the limitation is that K –means algorithm usually failed to produce reliable result when the data is noisy and when the number of clusters are less.

[10], presented a study titled "Spatio-temporal prediction of crimes using network analytic approach" in which a network analytics technique was used on publicly available crime data in Chicago city to analyse and predict occurrence of crime. The authors discovered that as they add extra layers of data which represent aspect of the society, there is noticeable improvement in the quality of crime prediction. Their prediction model was able to determine total number of crimes for the whole Chicago city with a very good predictive accuracy. However, the developed model could not predict the time slot of crime occurrence.

[11], used social media text mining, KNN, logistic regression with other forecasting models in their research work to make crime prediction based on planned social events. The limitation however is that it depends on other models to work well and it is limited by the quality of crime data derived from social media platforms.

3. METHODOLOGY

A total of 8,234 violent crime dataset were extracted from the list of general crime dataset made available being the scope of this study. Among various interesting attributes in the dataset are: crime id, crime description, date crime was committed, time of crime, number of deaths involved, type of crime as well as crime location. Presented in Table 1 is the summary of the violent crime dataset extracted between June, 2014 and June, 2019 from Lagos Police headquarters:

Table 1: Summary of the total violent crime dataset collected

Crime type	Sum of Data collected	Percentage
Arm Robbery	4984	60.5%
Murder	2549	31.0%
Wilful Manslaughter	82	1.0%
Aggravated Assault	146	1.8%
Rape	273	3.3%
Kidnapping	200	2.4%
Total	**8,234**	

Also shown in Figure 1 is the summary of violent crime dataset use for this study in graphical form.

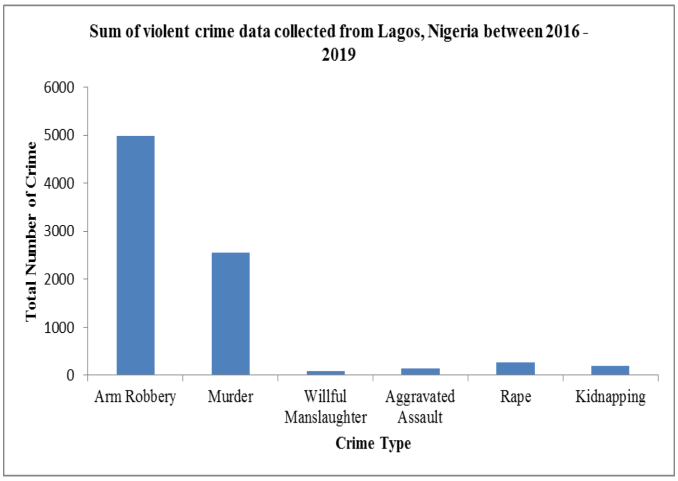


Figure 1: Summary of violent crime dataset analysis for this study

3.1 The Approach Overview

In order to predict occurrence of violent crime from a particular given location, the following tasks are formulated:

Given a location l and its feature vector space fi , a classification model is trained to output the probability of location l being a violent crime hotspot. However, feature vector space fi is extracted from the historical data which is presented in Table 1 and Figure 1 respectively. The collected violent crime dataset was further pre-processed through encoding of crime locations to construct a training dataset. The twenty four local government were encoded 1 – 24 as depicted in Table 2, and each local government was further divided into sub regions of 2.5km by 2.5km crime surveillance coverage according to [12]. The Table 2 shows the list of local government areas in Lagos state and the codes used for the local government areas and partitioned sub-regions which were tagged crime surveillance areas. In any classification problem formulation, there is usually positive and negative samples for the prediction. Here historical dataset collected which contained criminal information is regarded as the positive samples. Therefore there is need to get negative samples to avoid class imbalance during training. To be able to solve this problem, negative samples needed is generated randomly and uniformly from the map (Lagos state) as suggested by [13].

More specifically, evenly spaced points located on the map which are not coinciding with those positive samples are usually selected negative samples. However, it is also to be noted that sampling granularity determines the number of negative samples available. Therefore, to avoid getting class imbalanced (a case where positive/ samples becomes more than the negative samples or vice versa) which brings serious difficulty in most classification problem according to [12], the sampling granularity is then tuned in such a way that the number of negative samples is equal to that of positive samples. Consequently, the classification model was trained based on the training dataset constructed which was then used to evaluate the probability of occurrence of violent crime in any given location l in Lagos state.

Table 2: Codes used for regions and sub-regions in Lagos state during model building.

S/n	Local Government Areas	Population	Codes	No of crime coverage sub-regions
1	Agege	461,123	1	3
2	Ajeromi-Ifelodun	1,746,634	2	6
3	Alimosho	5,700,714	3	8
4	Amuwo-Odofin	318,576	4	3
5	Apapa	217,661	5	2
6	Badagry	241,437	6	3
7	Epe	181,715	7	3
8	Eti-Osa	287,958	8	4
9	Ibeju-Lekki	117,542	9	3
10	Ifako-Ijaiye	428,812	10	4
11	Ikeja	313,333	11	4
12	Ikorodu	535,811	12	5
13	Kosofe	665,998	13	4
14	Lagos Island	209,665	14	4
15	Lagos Mainland	317,980	15	4
16	Mushin, Lagos	633,543	16	5
17	Ojo	598,332	17	5
18	Oshodi-Isolo	621,789	18	5
19	Somolu	402,992	19	4
20	Surulere	504,409	20	5
	Lagos State	**16,506,023**		

3.2 Negative Sample Dataset

The negative data samples used for this study were derived by adopting the sampling method proposed by [13]. Using this method, negative samples were generated from the map with a certain distance interval from the positive samples as depicted in Figure 2.

3.3 Classification Model building

A violent crime dataset of 8,234 instances and 7 attributes were employed for this research based on features extraction with the target attribute being the crime_hot_spot. The attributes include crime id, crime description, date crime was committed, time of crime, number of deaths involved, type of crime as well as crime location. Python programming language was then used for violent crime using IBM cloud Watson studio. The classification was also executed in IBM cloud Watson studio [14] where cross validation and features extraction were done. Consequently the result yield an of 76.4% accuracy. Confusion matrix was used for evaluation in which the true positive rate and false positive rate are depicted in Table 5. The Positive predictive values as well as the false discovery rates of Decision Tree are shown in Figure 7. In addition, Figure 5 shows the ROC of the decision tree model. After using stratified cross validation [15], 82.04 % was rightly classified, and 17.96% was wrongly classified.

The pre-processed violent crime dataset was trained and tested based on five different violent crime type dataset in an attempt to measure and compare their performances in term of accuracy, precision, recall and F1 score metrics. Figure 3 show the model workflow of the Decision Tree model in Watson Studio.

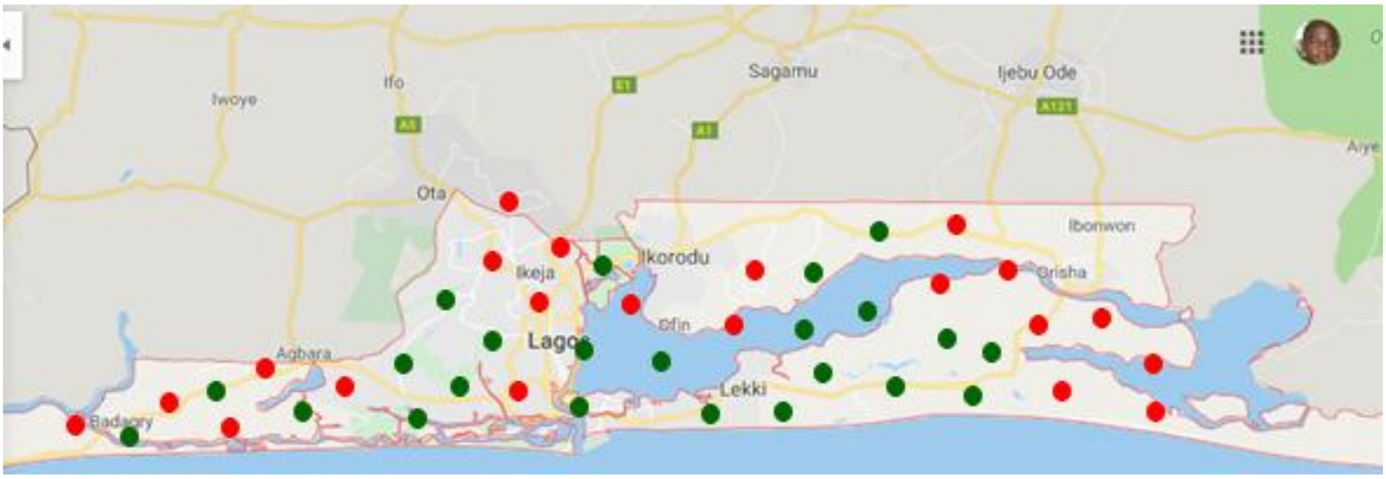


Figure 2: Lagos State boundary map for negative samples generation

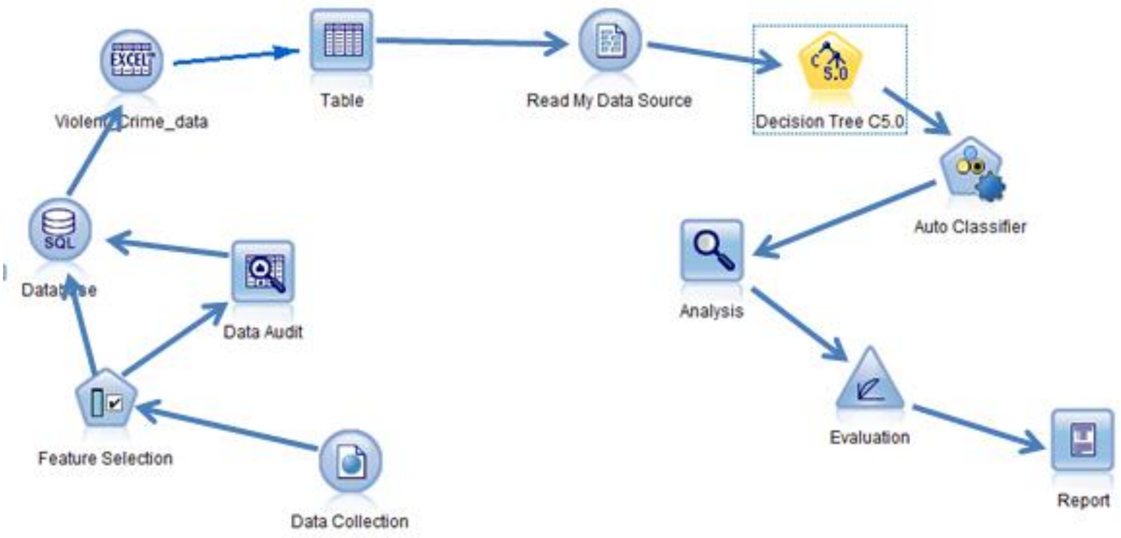
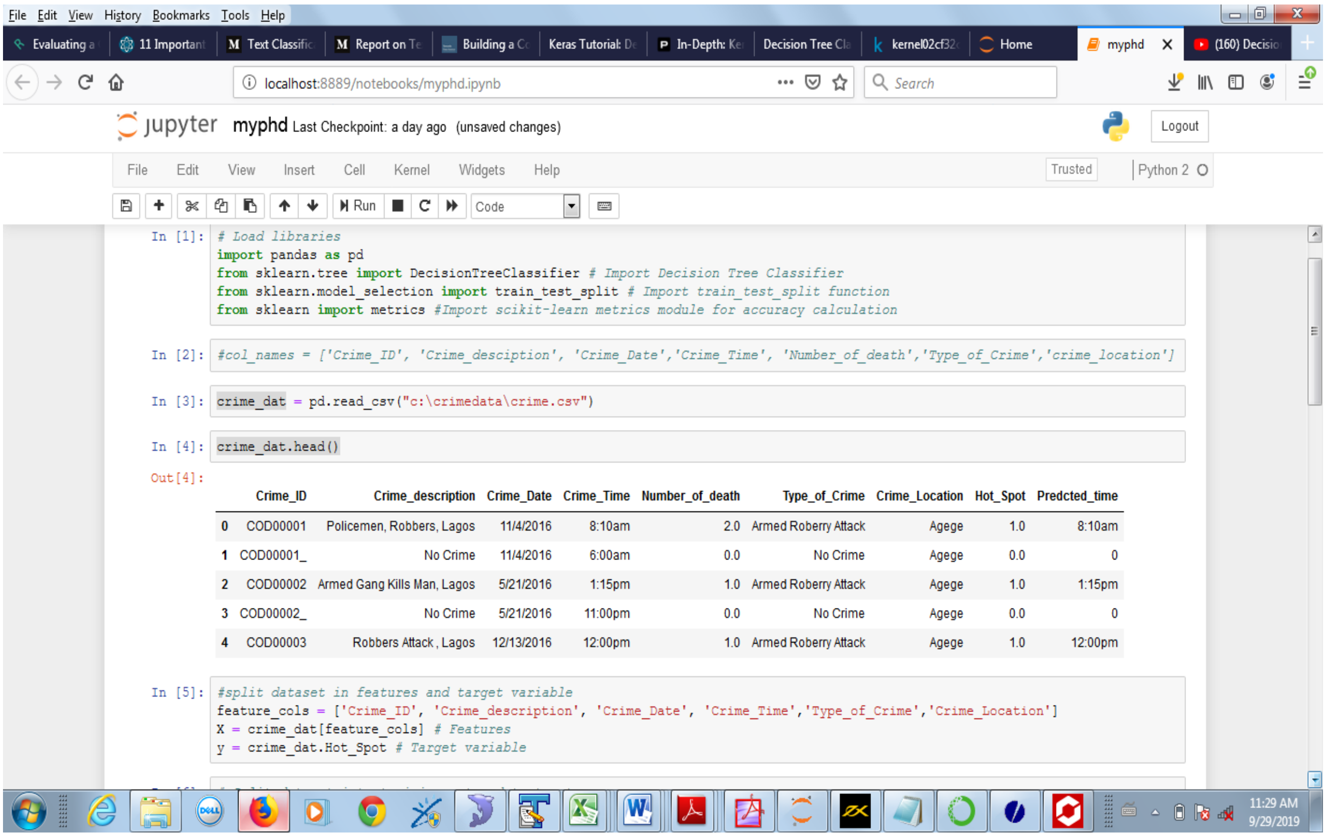


Figure 3: Model workflow for Decision Tree from Watson Studio

4. RESULTS

During the experiment, the decision tree learning was built to predict the target column, after the dataset was split into random training and test sets. Entropy and information gain as the splitting criteria was decided upon for splitting the dataset. Shannon Information Entropy Theory [16], defines the entropy of a variable as, , for the probability P of that variable taking values of 0,1,2, …, n and, the sum of probabilities of all variables is 1. [17] also stated in his paper that, the smaller the value of entropy, the better the system be described.

The following Figure 4 shows the decision tree generation as well as Python Jupyter Notebook classifier codes as snapshot during training:



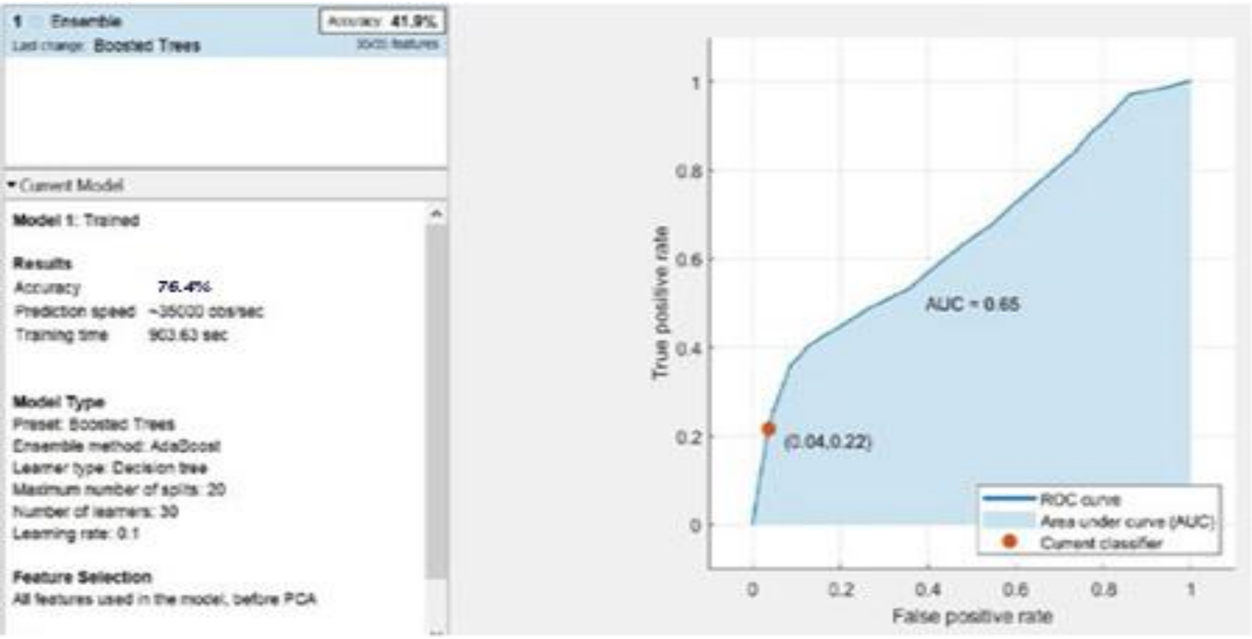
```
In [1]: # Load libraries
        import pandas as pd
        from sklearn.tree import DecisionTreeClassifier # Import Decision Tree Classifier
        from sklearn.model_selection import train_test_split # Import train_test_split function
        from sklearn import metrics #Import scikit-learn metrics module for accuracy calculation

In [2]: #col_names = ['Crime_ID', 'Crime_desciption', 'Crime_Date','Crime_Time', 'Number_of_death','Type_of_Crime','crime_location']

In [3]: crime_dat = pd.read_csv("c:\crimedata\crime.csv")

In [4]: crime_dat.head()
Out[4]:
```

	Crime_ID	Crime_description	Crime_Date	Crime_Time	Number_of_death	Type_of_Crime	Crime_Location	Hot_Spot	Predcted_time
0	COD00001	Policemen, Robbers, Lagos	11/4/2016	8:10am	2.0	Armed Roberry Attack	Agege	1.0	8:10am
1	COD00001_	No Crime	11/4/2016	6:00am	0.0	No Crime	Agege	0.0	0
2	COD00002	Armed Gang Kills Man, Lagos	5/21/2016	1:15pm	1.0	Armed Roberry Attack	Agege	1.0	1:15pm
3	COD00002_	No Crime	5/21/2016	11:00pm	0.0	No Crime	Agege	0.0	0
4	COD00003	Robbers Attack , Lagos	12/13/2016	12:00pm	1.0	Armed Roberry Attack	Agege	1.0	12:00pm

```
In [5]: #split dataset in features and target variable
        feature_cols = ['Crime_ID', 'Crime_description', 'Crime_Date', 'Crime_Time','Type_of_Crime','Crime_Location']
        X = crime_dat[feature_cols] # Features
        y = crime_dat.Hot_Spot # Target variable
```

Figure 4: Decision Tree model training in Python Jupyter Notebook

Also presented in Figure 5 is the Area Under Curve for Decision Tree Learning.



1 Ensemble
Last change: Boosted Trees
Accuracy: 41.9%
30/30 features

Current Model

Model 1: Trained

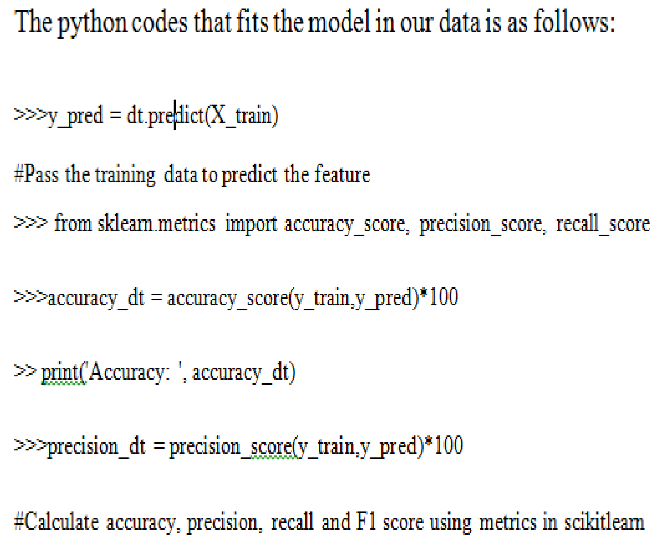
Results
Accuracy 76.4%
Prediction speed ~35000 obs/sec
Training time 903.63 sec

Model Type
Preset: Boosted Trees
Ensemble method: AdaBoost
Learner type: Decision tree
Maximum number of splits: 20
Number of learners: 30
Learning rate: 0.1

Feature Selection
All features used in the model, before PCA

Figure 5: AUC for Decision Tree Model

From Figure 5 above, it can be seen that the Decision Tree model attained an accuracy of 76.4% at a time of 82.63seconds.



The python codes that fits the model in our data is as follows:

```
>>>y_pred = dt.predict(X_train)

#Pass the training data to predict the feature

>>> from sklearn.metrics import accuracy_score, precision_score, recall_score

>>>accuracy_dt = accuracy_score(y_train,y_pred)*100

>> print('Accuracy: ', accuracy_dt)

>>>precision_dt = precision_score(y_train,y_pred)*100

#Calculate accuracy, precision, recall and F1 score using metrics in scikitlearn
```

The result of next day crime prediction for decision tree model is presented in Table 3 and the accuracy of the model was found to be 76.4%. This value is almost at per with the value reported in literature for decision tree on crime dataset.

Table 3: Sample of Empirical Results of Decision Tree model

Crime Locations	Hot_Spot	Predcted time
30	0	0
31	0	0
11	0	0
21	0	0
32	0	0
12	1	11.452
13	0	0
14	1	09.182
33	1	1.252
51	0	0
61	1	10.152
15	0	0
34	1	3.452
35	1	13.122
55	1	2.192
18	1	16.342
19	1	12.392
36	1	17.232
52	1	21.102
62	1	12.002
63	1	17.332

71	1	15.109
22	1	13.021
23	1	16.191
24	1	2.351
25	1	13.001
37	0	0
38	0	0
39	0	0

Also presented in Table 4.2 are the various evaluation results of Decision tree model for the different violent crimes.

Table 4: Evaluation Results of Decision Tree Classifier for different violent crime types

	Evaluation Results for Decision Tree model				
S/N	**Violent Crimes**	**TP Rates**	**FP Rates**	**Accuracy**	**Pre**
1	Armed Robbery	0.78	0.425	76.42%	77.
2	Kidnapping	0.77	0.389	77.25%	81.
3	Rape	0.79	0.478	75.83%	84.
4	Aggravated Assault	0.76	0.546	76.20%	73.
5	Murder	0.72	0,345	72.45%	73.
6	Ritual Killings	0.75	0.653	75.40%	81.

Prediction results comparison for Decision Tree between violent crime types is also presented in Figure 6.

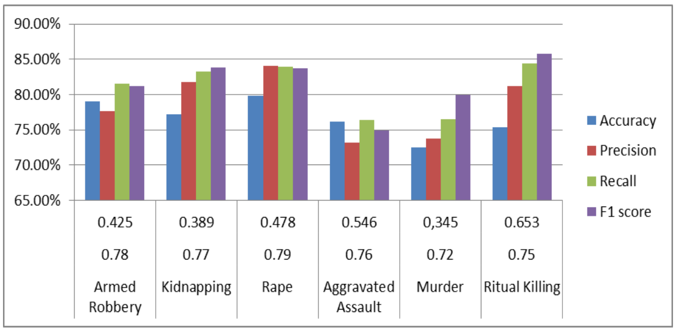


Figure 6: DT Prediction performance comparison between different violent crime types

The decision tree classification has an accuracy of 76.4% in its prediction and 82% of the instances correctly classified, while 18% of the instances were incorrectly classified. The scatter plot is depicted in Figure 7; showing the instances of the dataset against the crime location using the decision tree, and Table 5 showing it in a confusion matrix, the positive predictive values and false discovery rate.

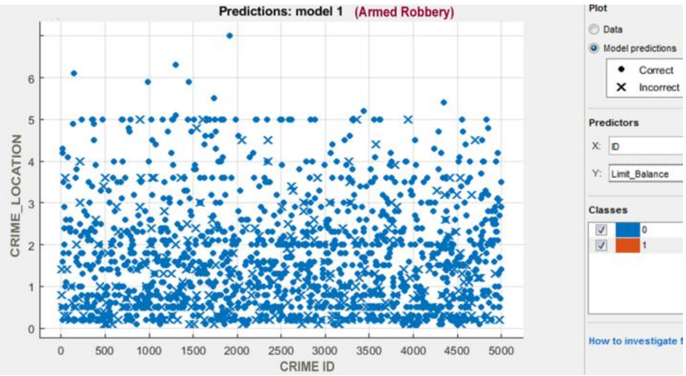


Figure 7: True Positive and False Negative rates of Decision Tree model

Presented in Table 4.3 is the result of Confusion matrix for decision tree.

Table 5: Confusion matrix result of Decision Tree model

True Class	Model 1	
	76%	18%
	24%	82%
Positive Predictive Value	76%	82%
False Discovery Rate	24%	18%
	Predicted class	

Accuracy = 76%

Figure 4.5 depicts the ROC curve of the decision tree prediction which is a function of the true positive rate and false positive rate of the prediction.

5 DISCUSSION

A number of relevant Machine Learning models were however considered as stipulated during the literature review and then compared in the analysis stage, out of which decision tree classifier model became a choice because of its outstanding performance especially in quick adaptation to the collected dataset. Decision classifier model was developed in IBM Watson Studio and use Python programming in Jupyter Notebook to be trained on the pre-processed violent crime dataset after splitting the dataset into training and test set. From the empirical results obtained, decision tree algorithm predicted the unknown class labels to the accuracy of 76.24% which is fair enough for a real system to be relied upon. This result showed an improvement on the work of [15] with decision tree result accuracy of 75.9%.

Violent crime has impacted negatively on the socio-economy of a nation and has increase the poverty rate of the citizenry. During the empirical study, cross validation was employed to avoid over-fitting during training and testing. This allows the model to work on a fraction of data not known before for the testing of the model. The training and testing yielded 76.4 % accuracy with a high true positive ratio. In addition 82.04% of the instances were correctly classified. As formerly presented in [18, 19, 20], hypothesis and testing in model formulation is not included in this study, rather decision tree machine learning technique was engaged in the model formulation and prediction.

6. CONCLUSION

The study has confirmed the efficiency of the decision tree algorithm with a prediction accuracy of 76.4 percent in a new context. A high accuracy was attained with a drastically reduced false positive and high true positives. The usage of state-of-the-technology technique to predict violent crime is displayed by this approach. In the future, other machine learning methods may be combined together to see if it could deliver a better result.

7. Acknowledgement

We are pleased to express our deep appreciation for the support and sponsorship granted by Covenant University Centre for Research, Innovation and Discovery (CUCRID).

REFERENCES

[1] National Bureau of Statistics, Nigeria (2018), Available online at http://nigeria.opendataforafrica.org/cgijuze/crime-statstics-reported-offences-by-type-and-state-2018

[2] Falade A, Ambrose A, Aderonke A. Oni, Felix Chidozie. (2019) Forecasting Violent Crime Hotspots Using a Theory-Driven Algorithm International Journal of Engineering Research and Technology. ISSN 0974-3154, Volume 12, Number 12 (2019), pp. 3130-3135.

[3] Liao R, Wang X, Li L, Qin Z. (2010). A novel serial crime prediction model based on Bayesian learning theory. In: Proceedings of the 2010 IEEE International Conference on Machine Learning and Cybernetics. vol. 4; 2010. p. 1757-1762.

[4] Wang P, Mathieu R, Ke J, Cai H.J (2010). Predicting Criminal Recidivism with Support Vector Machine. In: Proceedings of the 2010 IEEE International Conference on Management and Service Science; 2010. p. 1±9.

[5] Mohler G, Short M, Brantingham P, Schoenberg F, Tita G. (2011). Self-Exciting Point Process Modeling of Crime. Journal of the American Statistical Association. 2011; 106(493):100-108.

[6] Alves L., Ribeiro H., Lenzi E., Mendes R. (2013). Distance to the scaling law: a useful approach for unveiling relationships between crime and urban metrics. PLoS One. 2013; 8(8):1±8. https://doi.org/10.1371/journal.pone.0069580 PMID: 23940525

[7] Shehu G., Dikko H., and Yusuf B. (2014). Analysis of Crime Data using Principal Component Analysis: A case study of Katsina State, CBN Journal of Applied Statistics Vol. 3 No.2

[8] Jeon J and Jeong S (2016), Designing a Crime-Prevention System by Converging Big Data and IoT, Journal of Internet Computing and Services(JICS) 2016. Jun: 17(3): 115-128

[9] Waduge N. (2017), Machine Learning Approaches for Detecting Crime Patterns, web url – https://www.reseachgate.net/publication/319465093

[10] Saroj K, Dash I, Safro R., and Sakrepatna S. (2018). Spatio-temporal prediction of crimes using network analytic approach, arXiv:1808.06241v1 [stat.AP] 19 Aug 2018, Online acess - https://www.researchgate.net/publication/327134012_Spatio-temproal_prediction_of_crimes_using_network_analytic_approach

[11] Ristea A., and Leitner M. (2018). Integration of Social Media in Spatial Crime Analysis and Prediction Models for Events, AGILE 2018 – Lund, June 12-15, 2018

[2] Wang B., Zhang D., Brantingham P., Andrea L. (2018), Deep Learning for Real Time Crime Forecasting, eprint arXiv:1807.03340, NOLTA, 2018, 2018arXiv170703340W

[13] Gerber M (2014). Predicting crime using twitter and kernel density estimation. Decision Support Systems 61:115–125

[14] Lian D, Tao H, En C, Jianfeng Z, and Chao G., (2017). Deep Convolutional Neural Networks for Spatiotemporal Crime Prediction. Int'l Conf. Information and Knowledge Engineering | IKE'17 |

[15] Ahishakiye E., Elisha O., Danison T., Ivan N., (2017), Crime Prediction Using Decision Tree (J48) Classification Algorithm, International Journal of Computer and Information Technology (ISSN: 2279 – 0764) Volume 06 – Issue 03, May 2017

[16] Xing L. (2007), Tang Hua, Data mining algorithm based on genetic algorithm and entropy, Journal of Computational Information Systems, Volume 3, May 2007

[17] Dingsheng W. (2008), Data Mining Algorithmic Research and Application Based on Information Entropy, 2008 International Conference on Computer Science and Software Engineering

[18] Nicholas-Omoregbe O S Azeta A A Chiazor I A and Omoregbe N 2017 Predicting the adoption of e-learning management system: A case of selected private universities in Nigeria. Turkish Online Journal of Distance Education-TOJDE 18(2) 106-121.

[19] Azeta A A Misra S Azeta V I Osamor V C 2019 Determining suitability of speech-enabled examination result management system. Wireless Networks 1-8.

[20] C. K. Ayo, Jonathan A. Odukoya, Ambrose Azeta (2014), “A Review of Open and Distance Education and Human Development in Nigeria”. International Journal of Emerging Technologies in Learning. Volume 9 Issue 6, 2014. pp. 63-67. ISSN: 1863-0383