VIOLENT CRIME HOT-SPOTS PREDICTION USING SUPPORT VECTOR MACHINE ALGORITHM

FALADE ADESOLA[1], AMBROSE AZETA[2], ADERONKE ONI[3], A. E. AZETA[4], GREGORY ONWODI[5]

[1,2,3]Department of Computer and Information Sciences, Covenant University, Ota, Nigeria
[4]FIRRO; [5]Department of Computer Science, National Open University of Nigeria, Abuja.
{[1]adesola.falade, [2]ambrose.azeta, [3]aderonke.oni}@covenantuniversity.edu.ng
[4]aeeigbe@gmail.com, [5]gonwodi@noun.edu.ng

ABSTRACT

Accurate spatio-temporal violent crime hotspot prediction is a difficult and challenging task at this present time. Large amount of violent crime dataset are usually required for predicting future occurrence of violent crime in terms of location and time. Various data mining techniques have been applied in the previous studies on violent crime prediction with accuracy and other results that needed to be improved upon. In this paper, Support Vector Machine based spatial clustering technique for violent crime prediction was used. Firstly, historical violent crime dataset between 2014 and 2019 Lagos, Nigeria were collected and pre-processed through Principal Component Analysis, and then Support Vector Machine model built using IBM Watson Studio was applied on the six different violent crime dataset collected to determine violent crime hotspot locations for next day in Lagos Nigeria. The model was evaluated using real-life dataset of six violent crime types (murder, arm robbery, kidnapping, rape, non-negligent assault and man slaughter) dataset using confusion matrix. The results obtained found to return an accuracy of 82.12 percent which is good to be relied on for violent crime prediction. Based on this result, the model could be used by the Police authority to develop new crime control strategies and plan towards mitigating crime rate in the country.

Keywords: *Confusion matric, Data Mining, Support Vector Machine, Supervised Learning, Machine Learning*

1. INTRODUCTION

Crime is any act committed or omitted that violates the laws of the land and which results in penalty or imprisonment, according to [1]. Crime undoubtedly is a potential enemy of any individual or nation. A lot of issues relating to urban development such as housing, traffic, environmental problems have been reportedly emerging as a result of rapid growth in urban population and these has led to increase in crime rate. Migration of citizens from rural to urban areas has also increase the rate of crime in the country. Additionally, small as well as big crimes in urban areas have impacted negatively on our lives and properties. Therefore, it is imperative to create a safer urban environment for everyone to benefit from. National Bureau of Statistics reports that, rate of crime have been on the increase by 3.4% on a yearly average over the last 30 years. This is clearly worrisome and call for urgent attention. As a result of the increase in these heinous crimes: murder, kidnapping, armed robbery, and rape, people's anxiety have been intensified tremendously.

Crime occurrence prediction ([2], [3], [4], [5[, [6]) have received a lot of attention taking into consideration of its enormous benefits to the country and the citizenry at large. These authors have engaged several machine learning algorithms and data mining approaches to crime prediction by analyzing and modeling various previously collected violent crime dataset.

According to [2], there are different stages and analysis to crime prediction which includes: crime data collection, preprocessing, building of predictive model, training and testing of the model, prediction and evaluation. Also, data belongs to different types: ordinal, numerical and categorical. Therefore type of data usually determines the type of data mining approach to be adopted [2]. These data mining approaches includes clustering, regression and classification approach. Next is the trend and pattern identification phase of crime using a suitable algorithm. Ultimately, prediction, crime hotspot visualization, as well as evaluation phase are applied to study the results obtained and then analyse the behavior.

1.1 THE SUPPORT VECTOR MACHINE (SVM)

Structural risk management theory which Support Vector Machine premised on was discovered by [7]. Decision planes were used to define decision boundaries, used for separating groups of objects having diverse class memberships [8]. Support Vector Machine (SVM) works by creating an hyper plane using a linear model that are used for implementing non-linear class boundaries [9]. The SVM model have been used in diverse areas for prediction and detection tasks and has delivered outstanding performances, for example, in age estimation, pattern recognition, , telecommunications, system intrusion detection as well as face recognition according to ([10], [11], [12], [13). Figure 1 below is a diagrammatic representation of a simple support vector machine implementation.

2 RELATED WORK

[14] applied k-means clustering to detect patterns of crime and speeding up process of solving crime. The author engaged semi supervised learning approach for knowledge discovery from past crime records in order to enhance predictive accuracy. They also came up with weighting scheme attributes to address the limitation of clustering technique,

A new hotspot optimization tool was developed by [15] for crime hotspot mapping that can be used to determine the differences between classes in a spatial dataset. The study was able to accurately map crime hotspots in the North-eastern city of United State of America.

[16], in their work, made comparison between three typical classification algorithms which includes: Decision Tree C4.5 algorithm, Naïve Bayes and KNN algorithm in order to get very accurate results. Their findings reveals that a better predictive accuracy can be gotten by using KNN and GBWKNN missing data algorithms which is premised on a grey relational analysis theory (GRA)

GIS and cluster analysis for crime hotspot detection was studied by [17]. Several optimization based problems on cluster analysis for crime hotspots prediction was outlined by the authors. It was further suggested that adoption of geometric techniques and existing statistical approaches should be followed.

[18] discussed their research on discovering temporal and spatial criminal hotspots by engaging statistical analysis on Los-Angelis crime dataset. Criminal pattern hotspots was then conducted using Apriori algorithm. In addition, the authors applied Decision Tree and Naïve Bayes classifier to predict the likelihood of crime types occurrence.

[19] summarizes the approached for spatio-temporal crime pattern analysis. The author threw more light on the knowledge that one can obtain from these patterns and which particular data mining approaches could be used.

[20], proposed a new spatio-temporal approach for data analysis to uncover abnormal spatiotemporal clustering patterns. The authors proposed a framework for quantitative evaluation and then used it to make comparison against a generally used space-time scan statistic-based method. They based it on a robust clustering engine to predict abnormal spots with inconsistent shapes more accurately than the space-time scan statistic-based technique.

[21], developed a system that can predict crime regions which has high probability for violent crime occurrence using Naïve Bayes algorithm that delivered 90% accuracy. The accuracy of the classification and prediction was tested based on different test sets. The system developed takes attributes/factors of a place and Apriori algorithm produces the frequent patterns of the place. The pattern was then used for building a model for the decision tree classification.

A paper was presented by Sharma [22] which employs decision tree-based classification approach to predict criminal activities. In their study the author found that good classification result for suspicious e-mail detection is possible with application of advanced decision tree classifier and feature selection method. The results from this experiment reveal that an advanced ID3 algorithm has better classification accuracy than the traditional ID3 Algorithm.

[23] developed a crime analysis tool using different data mining methods, including Apriori algorithm and MV algorithm. The tool developed has four main steps: data cleaning, clustering, classification, and outlier detection. The outlier detection step is used to predict crimes that might happen in the future. The results reveal that the tool is effective in terms of analysis speed, and future prediction.

[24], in their research titled “Effectiveness of Support Vector Machine for crime hotspots prediction” applied Support Vector Machine learning technique on crime dataset to predict crime hotspots. The authors presented a SVM based approach to predict crime location as an alternative to existing data mining and modeling approaches. The performance of the model built was evaluated on real-life dataset using two-class SVM and one-class SVM as well as Neural Network. The result obtained showed a slightly better performance when compared to the neural network model.

[25], in their paper titled “Crime data analysis using Principal Component method” used the data collected from Nigerian Police statistics department to extrapolate crime rate, patterns and various hotspot areas in Katsina state through the use of statistical and principal component analysis. The result of these were made available to the Police authority which they use in taking proactive decisions and actions, but their approach is not efficient enough, as it takes time to get results and to make predictions.

In the paper published by [26], [27], in their effort at reducing sexual crime in Korea, the authors developed an intelligent crime prevention system (ICPS) by collecting and analysing big data using Term Document and Inverse Document Frequency (TD-IDF) and Naïve Bayesian algorithm and then fed into IoT devices and sensors to predict dangerous areas and alert a woman with a wearable device as she passes a hotspot zone. The limitation is that as the data become huge, there is a noticeable reduction in overall performance of the system.

Another author [28], in a paper titled “Machine Learning Approaches for Detecting Crime Patterns” The research focused on use of k-means algorithm in crime pattern detection system. They used K-means algorithm to cluster the crime dataset to detect and make prediction of crime occurrence. The gap cited in their work is that K – means does not produce reliable result when the data is noisy and when the number of clusters are less.

[29], in a study titled “Spatio-temporal prediction of crimes using network analytic approach” used network analytics technique to analyze and predict occurrence of crime by using publicly available data fused with other social information sources in Chicago city. 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 predict 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.

[30], proposed a model for crime prediction using social media text mining, KNN, logistic regression as well as spatial statistics with forecasting models from crime prediction to predict occurrence of crime based on planned social events. The gap in their work is it depends on other model to work 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 which represent independent variables are: crime id, crime description, date crime was committed, time of crime, number of deaths involved, type of crime as well as crime location. The target attributes are crime_hotspot and predicted_time. Presented in Table 1 is the summary of the violent crime dataset extracted between June, 2014 and June, 2019. Also shown in Figure 2 is the summary of violent crime dataset use for this study in graphical form.

A violent crime dataset of 8,234 records and 7 field data attributes were engaged for the study based on important features extraction with the following target attribute: crime_hot_spot. The attributes include crime_id, crime_description, crime_date, crime_time, Number_of_deaths, crime_type as well as crime_location. The following are the description of thse data fields:

- Crime_id – crime identification number
- Crime_description - description of crime
- Crime_date - date crime was committed
- crime_time – time of crime occurrence
- number_of_dealths - number of deaths involved
- crime_type - type of crime
- crime_location – location of crime
- crime_hot_spot – location of next occurrence of crime
- predicted_time – likely time of crime occurence

Python programming language was then used for violent crime using IBM cloud Watson studio. The classification was also executed in IBM cloud Watson studio [31] where cross validation and features extraction were done. Consequently the result yield an of 82.12% accuracy. Confusion matrix was used for evaluation in which the false positive rate and true positive rate are depicted in Figure 2. The Positive predictive values as well as the false discovery rates of Support Vector Machine are shown in Figure 3. After using stratified cross validation [32], 89.04 % was rightly classified, and 10.96% was wrongly classified.

3.1 SUPPORT VECTOR MACHINE MODEL

SVM is known to be a general machine learning framework that has deliver outstanding performances in many predictive scenarios including crime prediction domain. To this end, we trained the SVM using Python in Watson Machine Learning Studio ([32]), which is an open source for training machine learning models. The result of training Support Vector Machine model in IBM Watson studio is a product of many processes as shown in the workflow in Figure 3.

During the empirical study, SVM classifier was trained using features extracted from convolutional base. In order to train this classifier, a conventional machine learning method was engaged. Consequently, error of the classifier was estimated using k-fold cross validation process [32]. Since the k-fold cross-validation was used, training and the validation sets were concatenated in order to enlarge training data used (as was done in the previous case the test set was kept untouched). The following Python codes shows how data was concatenated.

```
# Concatenate training and validation sets

svm_features = np.concatenate((train_features,
validation_features))
svm_labels = np.concatenate((train_labels,
validation_labels))
```

Finally, as stated before, SVM classifier has one hyper-parameter. This hyper-parameter is referred to as the penalty parameter C of the error term. In order to optimize this choice of hyper-parameter, there is need to use exhaustive grid search [33]. The following Python codes was used to build the classifier, while Figure 4 illustrates the learning curves.

```
 # Build model
 import sklearn
 from sklearn.cross_validation  import
       train_test_split from sklearn.grid_search
  import GridSearchCV from sklearn.svm
import LinearSVC
X_train, y_train =
svm_features.reshape(300,7*7*512), svm_labels
param = [{  "C": [0.01, 0.1, 1, 10, 100]    }]
svm = LinearSVC(penalty='l2',
loss='squared_hinge')
clf = GridSearchCV(svm, param, cv=10)
clf.fit(X_train, y_train)
```

4. RESULTS

During the experiment, the Support Vector Machine model built was used to predict the target columns (hot_spot and predicted_time), after the dataset was split into random training and test sets.

The result of next day crime prediction for the Support Vector Machine model is presented in Table 4 and the accuracy of the model was found to be 82.12%. This value is almost at per with the value reported in literature for SVM on crime dataset.

In Table 2 below is empirical result of SVM model trained in Python Jupyter Notebook. The column crime_location is representing the various crime prone zones in the entire Lagos state, Hot_spot and Predicted_time are the prediction results from the SVM model. Hot_spot value 0 means no likelihood of violent crime occurrence, while Hot_spot value 1 is predicting the location as violent crime hotspot area that should warrant

Police target. The last column Predicted time is showing the likely time of the violent crime occurrence.

The result of next day crime prediction for Support Vector Machine model is presented in Table 2, and the accuracy of the model was found to be 82%. Also presented in Table 3 is the various evaluation results of Support Vector Machine model for the different violent crime types. Support Vector Machine model returns an accuracy of 82.12%. Prediction results comparison for Support Vector Machine between violent crime types is also presented in Figure 5.

5. DISCUSSION

A good number of relevant Machine Learning models were considered as summarized in the related work. In the comparative analysis stage, Support Vector Machine was chosen because of its outstanding performance especially in quick adaptation to new dataset. SVM model was developed in IBM Watson Studio and used Python programming in Jupyter Notebook to train on the pre-processed violent crime dataset after splitting the dataset into training and test set. From the empirical results conducted, SVM predicted the unknown class labels to the accuracy of 82.12% which is good enough for a real system to be relied upon. This result showed an improvement on the work of [34] with SVM result accuracy of 79.2%.

Violent crime has impacted negatively on the socio-economy development 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 82.12% accuracy with a high true positive ratio. In addition 89.04% of the instances were correctly classified. As formerly presented in ([34], [35], [36]), hypothesis and testing in model formulation is not included in this study, rather SVM learning technique was engaged in the model formulation and prediction.

6 CONCLUSION

The research work has further demonstrated the versatility and efficacy of the Support Vector Machine algorithm which has delivered a prediction accuracy of 82.12% (percent) in this new context. A high accuracy was attained with a drastically reduced false positive rates. The usage of state-of-the-technology method to predict violent crime is displayed by this approach. In the future, other machine learning approaches may be combined together to see if it could deliver a better result.

7. ACKNOWLEDGEMENT

This is to thank and appreciate the support and sponsorship granted by Covenant University Centre for Research, Innovation and Discovery (CUCRID) for the success of this study.

REFRENCES:

[1] NSW Criminal Law Handbook, 2017, Available online: https://www.booktopia.com.au/nsw-criminal-law-handbook-2017/book/9780455500058.html

[2] Noora A. and Wala A. (2017), KNN classifier and Naïve Bayes classifier for crime prediction in San Francisco context, International Journal of Database management systems (IJDMS) Vol. 9, No 4, August, 2017

[3] Chen P, Yuan H, Shu X (2008). Forecasting Crime Using the ARIMA Model. In: Proceedings of the 5th IEEE International Conference on Fuzzy Systems and Knowledge Discovery. vol. 5; 2008. p. 627-630.

[4] 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.

[5] Wang P, Mathieu R, Ke J, Cai HJ (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.

[6] 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.

[7] 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

[8] Cortes, C., and Vapnik, V., (1995). Support vector network, Machine Learning, 20(3): 273- 297. DOI: https://doi.org/10.1023/A:1022627411411

[9] Chiu, N.H., and Guao, Y., Y. (2008). State classification of CBN grinding with support vector machine. Journal of Material Processing Technology, 201:601-605.

[10] Elmi, H.E., Sallehuddin, R., Ibrahim, S., and Zain, A.M. (2014). Classification of sim box fraud detection using support vector machine and artificial neural network, International Journal of Innovative Computing, 4 (2): 19-27.

[11] Guo, G., Fu, Y., Dyer, C.R., and Huang, T.S. (2008). Image-based human age estimation by manifold learning and locally adjusted robust regression, Transactions on Image Processing, IEEE, 17(7): 1178-1188.

[12] Kumar, M., Ghani, R., and Mei, Z. S. (2010). Data mining to predict and prevent errors in health insurance claims processing. ACM 16th International Conference on Knowledge Discovery and Data Mining, pp. 65-74. Available from: http://dx.doi.org/10.1145/1835804.1835816

[13] Kirlidog, M., and Asuk, C. (2012). A fraud detection approach with data mining in health insurance. Procedia-Social and Behavioral Sciences, 62: 989-994. http://dx.doi.org/10.1016/j.sbspro.2012.09.168

[14] Anwar, S., Zain, J. M., Zolkipli, M. F., Inayat, Z., Khan, S., Anthony, B., and Chang, V. (2017). From intrusion detection to an intrusion response system: Fundamentals, requirements, and future directions. MDPI Algorithms, 10(2): 1-24, DOI: 10.3390/a10020039

[15] S. V. Nath, "<Crime Pattern Detection Using Data Mining.pdf>." International Journal of Database Management Systems (IJDMS) Vol.9, No.4, August 2017

[15] D. Wang, W. Ding, H. Lo, T. Stepinski, J. Salazar, and M. Morabito, "Crime hotspot mapping using the crime related factors—a spatial data mining approach," Applied Intelligence, vol. 39, pp. 772-781, 2012.

[17] C.-c. Sun, C.-l. Yao, X. Li, and K. Lee, "<Detecting Crime Types Using Classification Algorithms.pdf>."

[18] T. H. Grubesic, "<Detecting Hot Spots Using Cluster Analysis and GIS.pdf>."

[19] T. Almanie, R. Mirza, and E. Lor, "Crime Prediction Based on Crime Types and Using Spatial and Temporal Criminal Hotspots," International Journal of Data Mining & Knowledge Management Process, vol. 5, pp. 01-19, 2015.

[20] K. Leong and A. Sung, "<A review of spatio-temporal pattern analysis approaches on crime analysis.pdf>."

[21] W. Chang, D. Zeng, and H. Chen, "A stack-based prospective spatio-temporal data analysis approach," Decision Support Systems, vol. 45, pp. 697-713, 2008. S.

[22] Sathyadevan and S. Gangadharan., "<Crime Analysis and Prediction Using Data Mining.pdf>."

[23] M. Sharma, "Z-CRIME: A data mining tool for the detection of suspicious criminal activities based on decision tree," in Data Mining and Intelligent Computing (ICDMIC), 2014 International Conference on, 2014, pp. 1-6.

[24] A. Malathi and S. S. Baboo, "An enhanced algorithm to predict a future crime using data mining," 2011.

[25] Keivan K. and Reda A., (2008). Effectiveness of Support Vector Machine for Crime Hotspots prediction. Applied Artificial Intelligence An International Journal Volume 22, 2008 - Issue 5

[26] 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

[27] Sumanta D. and Malini R.C., (2016). A Geo-Statistical Approach for Crime hot spot Prediction.International Journal of Criminology and Sociological Theory, Vol. 9, No. 1, August 2016, 1-11.

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

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

[30] 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

[31] 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

[32] IBM Watson studio documentation and resources. Available online: https://dataplatform.cloud.ibm.com/docs/content/wsj/getting-started/overview-ws.html

[33] Fan R. E, Chang K. W, Hsieh C. J, Wang XR, Lin C. J. (2008). A library for large linear classification. The Journal of Machine Learning Research. 2008; 9:1871±1874

[34] Charles K. A, Jonathan A. Odukoya P, Ambrose A (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

[35] 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.

[36] 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.

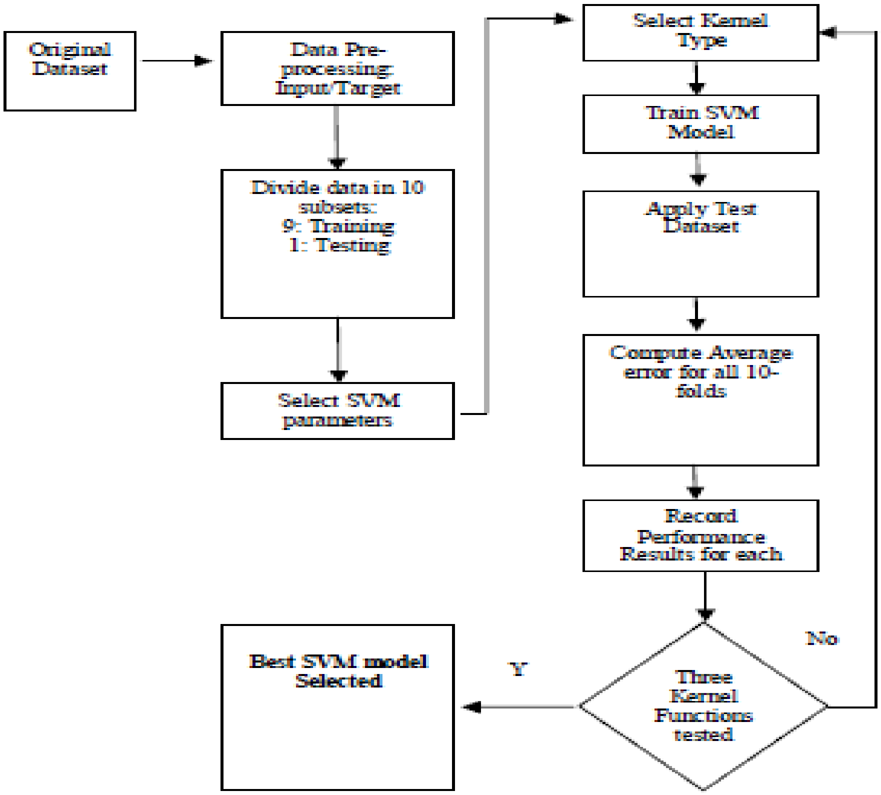
Original Dataset
Data Pre-processing: Input/Target
Divide data in 10 subsets: 9: Training 1: Testing
Select SVM parameters
Select Kernel Type
Train SVM Model
Apply Test Dataset
Compute Average error for all 10-folds
Record Performance Results for each
Three Kernel Functions tested
No
Y
Best SVM model Selected

Figure 1: A Sample Of SVM Implementation ([9]).

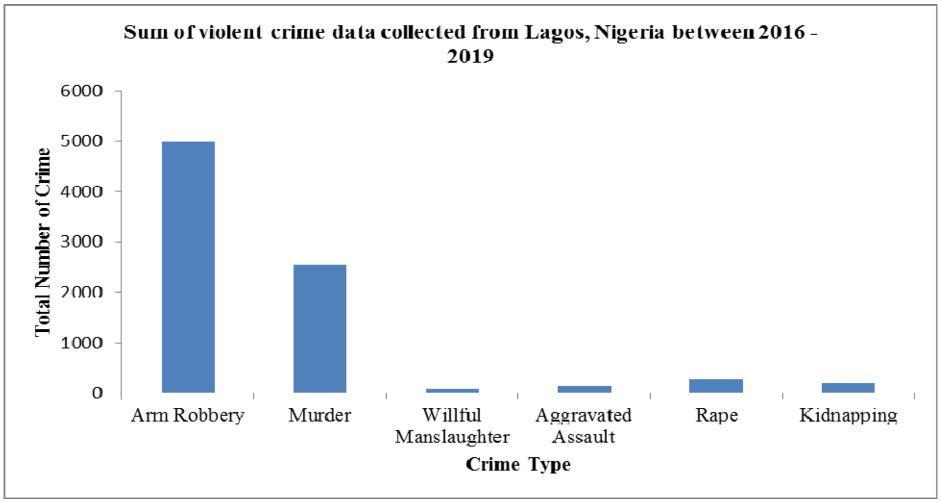

Figure 2: Summary Of Violent Crime Dataset Analysis For This Study

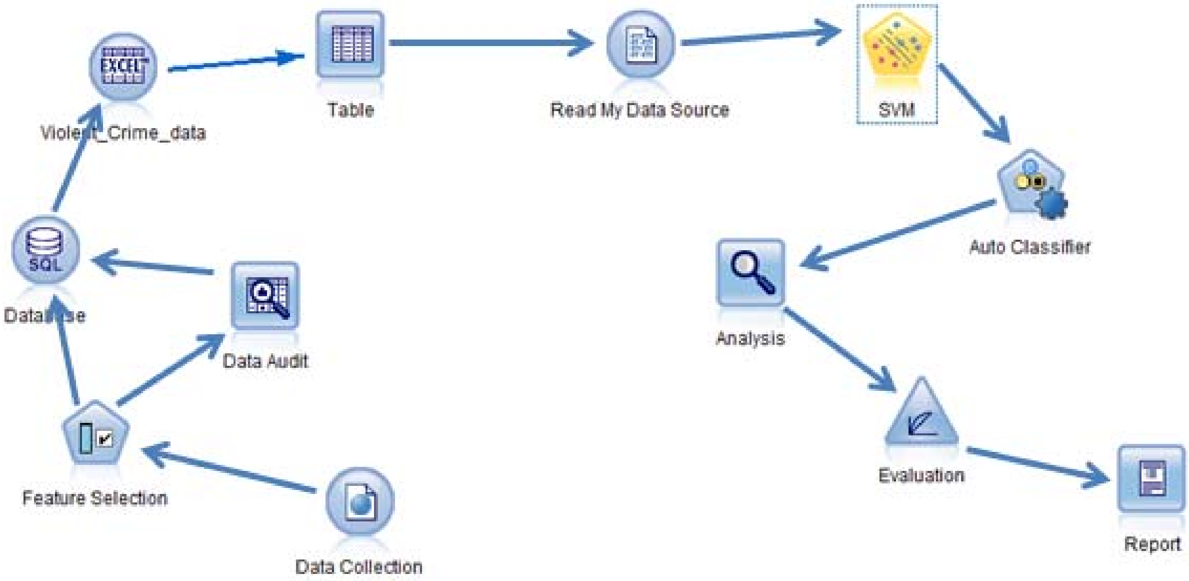

Figure 3: Model Workflow For Support Vector Machine In IBM Watson Studio

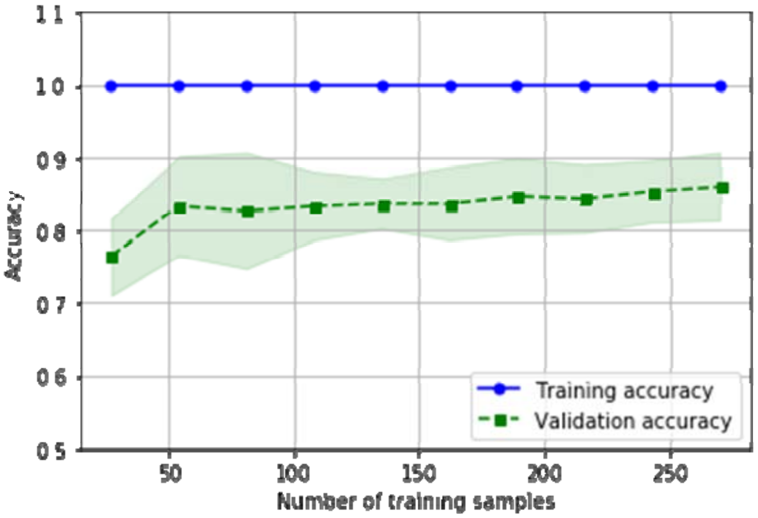

Figure 4. Accuracy Of The Support Vector Machine

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%
Willful Manslaughter	82	1.0%
Aggravated Assault	146	1.8%
Rape	273	3.3%
Kidnapping	200	2.4%

Table 2: Sample Of Empirical Results Of Support Vector Machine Model

Crime Locations	Hot_Spot	Predicted Time
10	0	0
11	0	0
12	0	0
13	1	8.101
14	1	16.101
30	1	11.101
31	1	8.101
32	1	14.101
51	0	0
52	1	6.101
53	0	0
54	0	0
32	0	0
34	0	0
35	1	19.122
60	1	9.122
61	1	10.122
62	1	6.251
55	1	9.122
63	1	11.201
64	1	9.201
81	1	15.201
82	1	9.201
83	0	0
90	1	9.201
91	1	9.201
57	0	0
92	0	0
93	0	0
84	1	10.001