Proceedings on Big Data Analytics & Innovation (Peer-Reviewed), Volume 1, 2016, pp.53-66

Data Mining Application in Crime Analysis and Classification

OBUANDIKE GEORGINA N.
Federal University Dutsinma, Katsina State, Nigeria

JOHN ALHASAN & M. B. ABDULLAHI
Federal University of Technology, Minna, Niger State, Nigeria

ABSTRACT The analysis of crime data helps to unravel hidden trends that will aid in better understanding of crime pattern and the nature of those who commits such crimes. It also enables appropriate strategies to be put in place to control such crimes. Literature revealed that data mining has been successfully applied in crime analysis and control. In this work the crime data used was collected from selected Nigerian prisons. Exploratory analysis was performed on the crime data for better insight into the dataset before application of data mining algorithms. The exploratory analyses revealed that majority of the offenders are within the ages of 18 to 34 years old. The mining algorithm was limited to two classification algorithms. C4.5 algorithm was used to classify the data into vulnerable and non vulnerable groups. To verify the reliability of the C4.5 algorithm, Naïve Bayes algorithm was also used to classify the dataset. The result showed that C4.5 classified the data better with higher accuracy of 97% against 93% from Naïve Bayes. The rule generated by the C4.5 classifier revealed that those without educational qualification and not gainfully employed are the vulnerable groups. The authors are of the opinion that education is the way out of crime in Nigeria since most of the offenders either have low educational qualification or none.

Keywords: Data Mining, Crime Analysis, Naïve Bayesian, Tree Classifier

1. Introduction

Crime is a societal ill and its cost is usually enormous hence, the need for analysis of crime data to learn the factors that enhances crime and the nature of offenders (Wilson, 1963; Brown, 2003). Many people desire to live and operate in a secure environment; they want to be sure that their lives and that of their loved ones are secured. The main duty of any Government is to secure the lives and properties of its citizens through relevant policies and strategies. Nigeria is currently having serious security challenges ranging from the Boko Haram attacks in the North-East to armed robbery and kidnapping in other parts of the country. Data mining has gained recognition in crime analysis (Jawei et al, 2012). It is a field that cut across many other fields. There are many definitions of data mining according related literatures (Julio and Adem, 2009). The emergence of computing and communication technology has produced a society that extremely depends on information

(Witten and Frank, 2000). Development in technology has helped in collection and storage of large amount of data in many organizations' database. These usually contain hidden information. Most of these organizations gather these data for operational purposes after which they are dumped in data repositories or even thrown away or deleted. This type of data when mined can help in discovering of relevant information which can help the organization to increase productivity and can serve also as essential information to the society at large. Data mining has the capability to unravel the information that is usually hidden in such databases. Naisbitt (1986) is of the opinion that we are being choked with data but lack relevant information because the data are not mined to get relevant information. Data mining tools and algorithms are used to find relevant trends and to make necessary predictions and associations in data. Data mining has been successfully applied in virtually all areas of human endeavour which include banking, marketing, manufacturing, telecommunication, e-commerce and education (ZaoHui and Jamie, 2005). Data mining is an intelligent and potent data extraction technique that uses different types of data extraction algorithms. Data analysts explore large data repositories by using these data mining algorithms (Chen et al, 2004; Fayyad and Uthurusamy, 2002)

The analysis of crime data will help in cost reduction and reduction of training time of officer involved in crime control. It will also help in distribution of scarce resources to the appropriate quarters (Megaputer, 2002). The rest of the sections discussed about data mining techniques, data mining process; classification of the crime data using two popular classifiers and discussion of results.

2. Classification Techniques for Crime Analysis

Classification

Classification is a technique used to predict an unknown class label using a function. Classification as a method comes in two steps, the number one step involves the construction of the classification model (model training) while the second step involve using the model to predict class labels. An instance R of an R m-dimensional attribute vector can be represented as $R = (r_1, r_2 \dots r_m)$ each i $R = (r_1, r_2 \dots r_m)$ nstance belongs to a class of determined attributes $T_1, T_2 \dots . T_m$. When $T_1, T_2 \dots . T_m$ an attribute class is discrete value or unordered, it is said to be a categorical or nominal attribute and it serves as the category or fields of the records. The records that are used for the construction of the classification model is usually trainers and is portioned out from the dataset being used for analysis. The training model can be represented as a function $Z = f(r)$ which $Z = f(r)$ represents the used fields Z of a g Z iven record R (Jawei R et al, 2012).

C4.5 Classifier

It is a statistical classifier that is used to create a decision tree. It is carved out from ID3 method to overcome its methodological challenges by pruning the decision tree after construction and handling discrete and continuous dataset. Let D be data D set $d_1, d_2, \dots d_n$ with m $d_1, d_2, \dots d_n$ dimens m ional attributes $t_1, t_2, \dots t_m$ and $k_1, k_2, \dots k_i$ r $t_1, t_2, \dots t_m$ epres $k_1, k_2, \dots k_i$ ents class groups. At each point in the tree C4.5 algorithm usually pick an attribute that gave the best split of the dataset. The attribute with the highest normalized value is chosen for the split and it is placed at the root of the tree. C4.5 is a supervised learning method that is simple and easy to implement. It divides dataset into portions with different characteristics. The last leave of the tree usually depicts predictions while the in between nodes depicts various test on the attributes (from the root node to the leaf node). The normalized value is calculated using equation 1 and equation 2.

$$\text{Gain}\,(q) = f(inf(T) - inf[(q,T)]$$

.

$$\text{Gain}\,(q) = f(inf(T) - inf[(q,T)] \quad (1)$$

where

$$inf(q,T) = \sum_{i=1}^{n} q_i \times \text{Ent}\,(q_i)$$

.

$$inf(q,T) = \sum_{i=1}^{n} q_i \times \text{Ent}\,(q_i) \quad (2)$$

Naïve Bayes Classifier

Naïve Bayes classifier is a probability based classifier and has proved its effectiveness in many areas where it has been applied. It is fast and easy to use which made it popular in data mining field. Though usually criticized for its attribute independent assumptions but it still competes favourably with other higher classifiers. Naïve Bayes calculates the probability value and selects the class with the highest probability (Taheri et al, 2014). It is represented mathematically as shown in equation 3

$$.P(K_i \cap Y) = \frac{P(Y \cap K_i)P(K_i)}{P(Y)}$$

$$.P(K_i \cap Y) = \frac{P(Y \cap K_i)P(K_i)}{P(Y)} \quad (3)$$

For a database with high dimension the computational cost is usually high thus the application of Naïve Bayes.

$$P(Y \cap K_i) = \prod_{k=1}^{n} P(Y_k \cap K_i) \quad (4)$$

$$P(Y \cap K_i) = \prod_{k=1}^{n} P(Y_k \cap K_i)$$

Naïve Bayes Algorithm
1) Input attributes and the class of the instances
2) Compute the posterior value for each attribute against the class
3) Compute the value before the existing class
4) Multiply the results from 2 and 3 for all the classes
5) Choose the highest value as the classification

Source: Taheri et al, (2014)

3. Data Mining Process Models

Data mining process required following some basic outlined steps when mining data. These steps outline all the necessary procedures for data mining. This process was originally proposed by Kurgan and Musilek (2006) and since then, many other mining processes have been developed. One common thing about all the process models is that they all outline steps which usually include loops and iterations (Kurgan and Musilek, 2006). CRISP-DM is a popular mining methodology that is generally accepted by data mining experts and it is a leading methodology used by data miners (Kurgan and Musilek, 2006). CRISP-DM is the process model that has been chosen for this work.

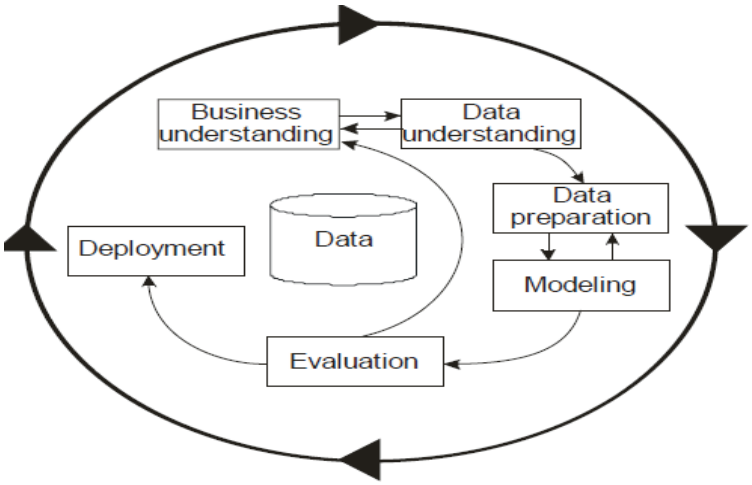

Figure 1: CRISP-DM Cycle

Source: (Gartner Group, 1995)

4. Methodology

The process followed in this research is as outlined in figure 2. The process started with business understanding which is all about understanding the problem domain and translating it to mining problem followed by the data collection stage where the required dataset was collected from selected Nigerian prisons. The exploratory analysis was done to get a better insight into the dataset before analysis. The classification stage was where the data mining proper was carried out using the two classifier namely C4.5 and Naïve Bayes classifiers. The evaluation stage is where the classification results from the two classifiers are discussed.

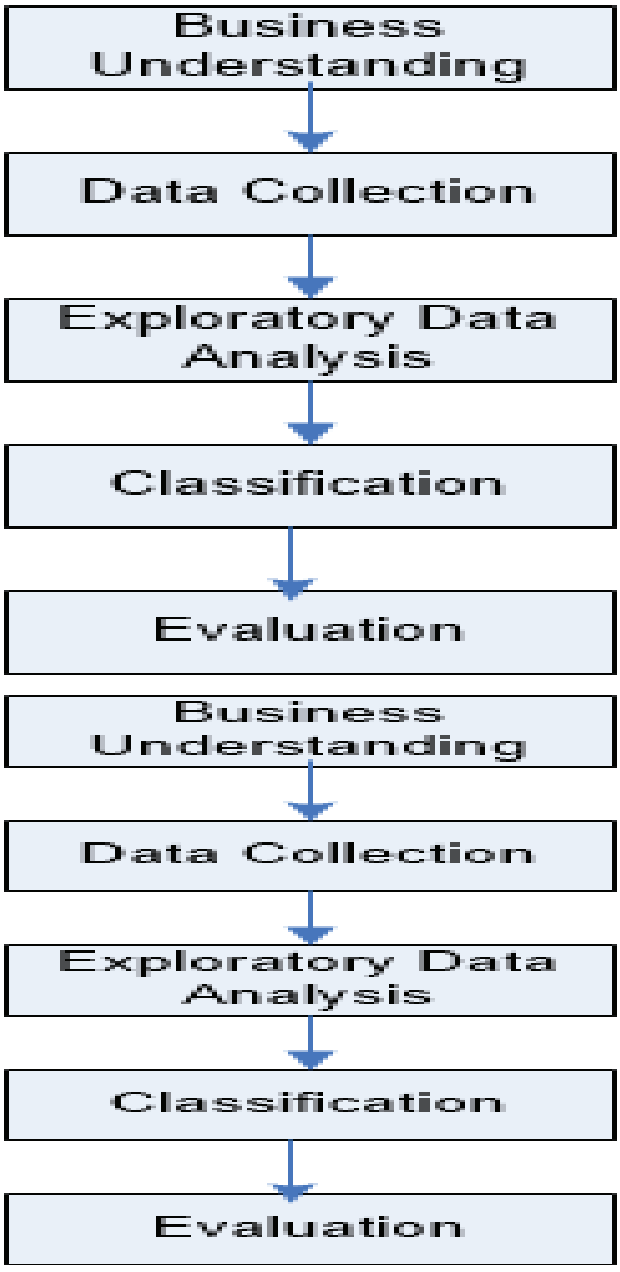

Figure 2: Work Methodology Flowchart

Dataset Description

The data used in this work were collected from selected Nigerian Prisons in Katsina, Kano, Kaduna, Niger and Abuja. Crime is broad term that covers range of unpleasant activities ranging from simple to complex. In this work crime was categorized into three groups as low, average and high crimes. The data consists of five fields and a class attribute categorized as shown below:

Education

- No Education: Implies not having any educational qualification
- Low Education: includes Primary and Secondary graduates
- Average Education: includes ND, NCE
- High Education: includes Degrees, PGD, MSC, MA,PhD

Occupation

- Unemployed: implies no work at all
- Self Employed: includes farmers, Apprentice, traders, artisans
- Employed: includes private employed and government employed

Age

- Early: ages (18 : 34)
- Middle: ages (35 : 50)
- Late: ages (51 : 150)

Crime

- Low Crime: Breach of Trust, Conspiracy, Assult
- Average Crime: Rape, Kidnapping, Drug
- High Crime: Homicide, Armed Robbery, Theft

Sex

- Male
- Female

Class

- Vulnerable
- Non Vulnerable

5. Performance Measures

The common measures for evaluating performance of data mining models are sensitivity, relevance, specificity, kappa statistics, area curve, time and accuracy.

Sensitivity: It is a statistics that shows the records that are correctly labelled by the classifier. It can be defined as: $\text{Sensitivity} = {}^{TP}/_{N}$

$\text{Sensitivity} = {}^{TP}/_{N}$

Specificity: It is simply a report of instances incorrectly labelled as correct instances; it can be defined as: $\text{Specificity} = {}^{FP}/_{N}$

$$\text{Specificity} = {FP}/{N}$$

Precision: Simply measures exact relevant data retrieved. High precision means the model returns more relevant data than irrelevant data.

$$\text{Precision} = \frac{TP}{\text{TP} + FP}$$

Kappa: measures the relationship between classified instances and true classes. It usually lies between $[0, 1]$, the v $[0, 1]$ alue of 1 means perfect relationship while 0 means random guessing.

Accuracy: this shows the percentage of correctly classified instances in each classification model

Time: Implies time taken to perform the classification (Milan and Sunila, 2011; Hong etal, 2006)

6. Exploratory Analysis of Crime Data

In this work WEKA mining tool was applied. WEKA is a machine learning software that has gained recognition in data mining because it implements many different data mining algorithms and also has potent tools for data pre-processing and visualization. It is an open source and accepts its data in Attribute Related File Format (ARFF). The sample of the converted ARFF file for this work is shown in figure 4.

```
@relation 'Formated Prisons2-
weka.filters.unsupervised.attribute.Remove-R1-2,11-19
-weka.filters.unsupervised.attribute.Remove-R7-
weka.filters.unsupervised.attribute.Remove-R3-
weka.filters.unsupervised.attribute.Remove-R3'
@attribute Offence {low,high,average}
@attribute Age {early,late,middle}
@attribute Sex {M,F,'M '}
@attribute Edu-Qualification {low,average,NONE}
@attribute Occupation {'Self Em-
ployed',unemployed,employed}
@attribute Class {vulnerable,'non vulnerable'}
@data
low,early,M,low,'Self Employed',vulnerable
high,early,M,average,'Self Employed',vulnerable
low,early,M,average,'Self Employed',vulnerable
average,early,M,low,'Self Employed',vulnerable
average,early,M,low,'Self Employed',vulnerable
high,early,M,average,'Self Employed',vulnerable
average,early,M,low,'Self Employed',vulnerable
low,early,M,low,'Self Employed',vulnerable
high,early,M,low,'Self Employed',vulnerable
high,early,M,low,'Self Employed',vulnerable
low,early,M,low,'Self Employed',vulnerable
low,early,M,average,unemployed,vulnerable
average,late,M,NONE,employed,'non vulnerable'
low,early,M,average,'Self Employed',vulnerable
average,early,M,average,'Self Employed',vulnerable
```

Figure 4: A Sample ARFF for the Crime dataset

When preparing data for data mining seeing the data pictorially provides insight into what is happening and this insight can help improve model building. The data mining tool chosen for this work has the features for exploratory data analysis. The relative densities of the various attributes in the data set are as shown in figure 5.

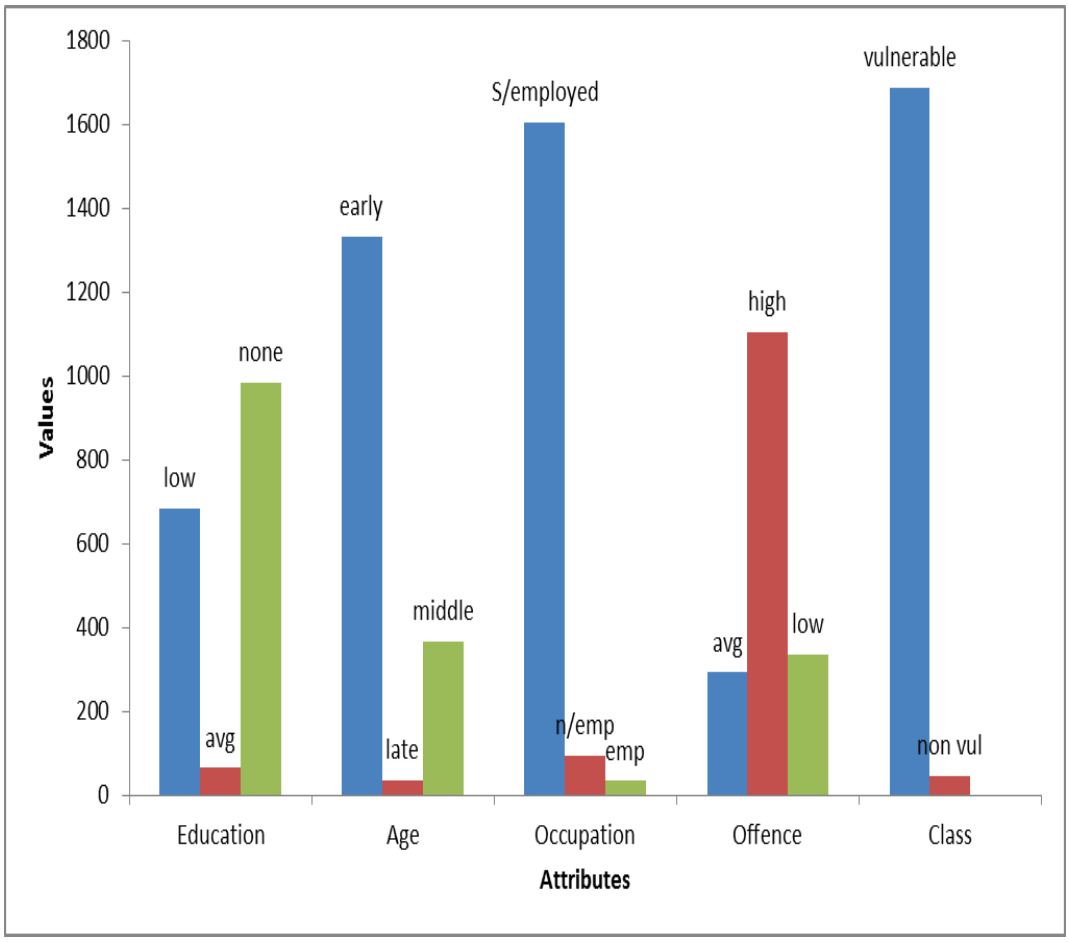

Figure 5: Densities of Attributes in the dataset

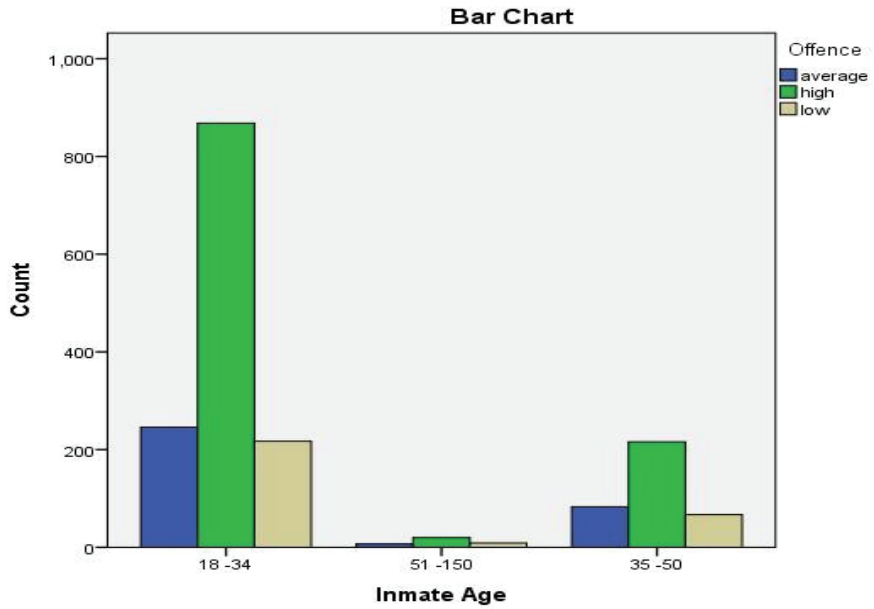

Figure 6: Association of Age and Offence

The visualization of the relationship between the age and offence reveals that majority of the offenders are within the ages of 18 -34 (early age) and commit high crime.

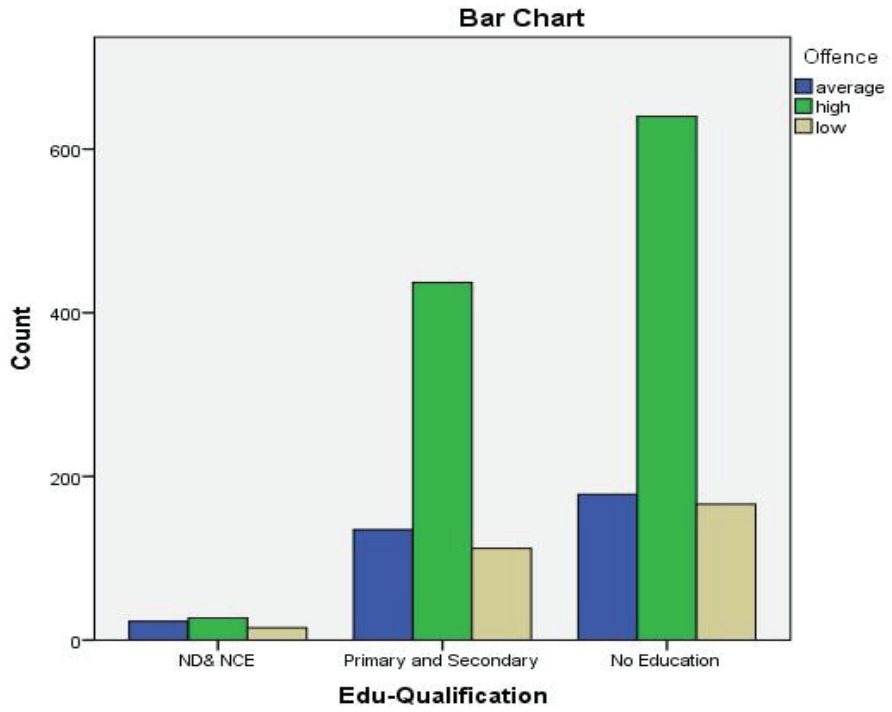

Figure 7: Association of Educational Qualification and Offence

The visualization of the offence versus educational qualification reveals that majority of the offenders have low education qualification (primary and secondary) or no education qualification at all and that they commit high crime and are vulnerable groups.

7. Classification of Crime Dataset

In this work k-Fold cross-validation method has been applied In order to ensure good performance of the classification model. The method was used to train and test the classifier. This method usually divides dataset into k folds; it trains the model with k-1 folds and tests the built model with the remaining k fold. It usually obtained k different results and takes the average to obtain the model accuracy. In this work, 10 fold cross validation was used. This method is better than the random sampling method because it takes care of the bias usually associated with random sampling method. The Classification was done using C4.5 classifier which is a decision tree classifier; Naïve Bayesian which is a probability based classifier de-

veloped to handle categorical dataset was used for reliability test. Table 1 is the tabulation of the result obtained from the two classifiers using WEKA mining tool.

Table 1: Tabulated Results

Evaluation Metrics	NB	C4.5
Time	0.05 secs	1.06 secs
Accuracy	93	97
TP Rate	0.935	0.971
FP Rate	0.067	0.027
Kappa	0.8696	0.9409
Precision	0.935	0.971
Recall	0.935	0.971
ROC curve	0.989	0.986

The result above revealed that the C4.5 classifier has better accuracy of 97 in comparism to the accuracy of Naïve Bayes. C4.5 though took more time of 1.06 seconds to build the model compare to 0.05 seconds taken by the Naïve Bayesian still handles the data better. In terms of classifier performance using the ROC curve the C4.5 classifier performed comparably well against Naïve Bayes on the dataset.

ROC curve is used to visualize classifiers performance. It is usually plotted using sensitivity at the y axis and specificity at the x axis. If the area under the curve is 1, it indicates perfect prediction while 0.5 implies random guess. The areas under the curve for the naïve Bayesian and C4.5classifiers are close to 1 which indicates the classifiers performed well.

8. Deductions from C4.5 Classification Tree

i. Offence = high implies vulnerable
ii. Offence = low and Education Qualification = high implies non vulnerable
iii. Offence = low and Education Qualification = average implies non vulnerable
iv. Offence = low, Education Qualification = NONE, Age = early, and Sex = male implies vulnerable
v. Offence = low, Education Qualification = NONE, Age = middle implies non vulnerable
vi. Offence = low, Education Qualification = NONE, Age = late implies non vulnerable

vii. Offence = low, Education Qualification = NONE, Age = early, Sex = F and Occupation = employed or unemployed implies vulnerable
viii. Offence = low, Education Qualification = low, Age = middle implies non vulnerable
ix. Offence = low, Education Qualification = low, Age = early and Sex = M implies vulnerable
x. Offence = low, Education Qualification = low and Age = late implies non vulnerable

9. Conclusion

Data mining has the capability that makes it simple convenient and suitable for data extraction from large databases. It employs different mining algorithms for its work. Many agencies gather data for its operational purposes, such data can be mined to discover some relevant patterns that can aid in decision making. The analysis of crime data will help to unravel crime pattern and nature of those who commits such crime so that appropriate strategies and rules will be put in place to control such crimes. The work reveals that the majority of the inmates that commit crime are between the ages of 18 to 34 and have low or no educational qualification and are either self employed or not doing anything at all. The classification result reveals that 98 percent of these groups of people are threat to the society. Thus, the researchers are of the opinion that government should encourage education and our youths should be gainfully employed.

Correspondence
Obuandike Georgina N.
Department of Mathematical Sciences and IT
Federal University Dutsinma
Katsina state, Nigeria

References

Barnett, V., Lewis T., Outliers in Statistical Data. John Wiley, 1994.

Brown, D. (2003). The Regional Crime Analysis Program (RECAP): A Framework for Mining Data to Catch Criminals. http://vijis.sys.virginia.edu/publication/RECAP.pdf

Chen, H., Chung, W., Xu, J.J., Wang, G., Qin, Y., and Chau, M. (2004). Crime Data Mining: A General Framework and Some Examples. *Computer,* 37(4), 50-56.

Fayyad, U.M. and Uthurusamy, R. (2002), "Evolving Data Mining into Solutions for Insights", Communications of ACM*,* 45(8), 28-31.

Gartner Group (1995), Gartner Group Advanced Technologies and Applications Research Note http://www.gartner.com.

Hawkins, D., Identification of Outliers, Chapman and Hall, 1980.

Hong, H., Jiuyong, L., Ashley P., (2006) "A Comparative Study of Classification Methods for Microarray Data Analysis", published in CRPIT, Vol. 61.

Jiawei, H., Micheline, K., and Jian P. (2012)"Data mining: Concept and Techniques" 3rd edition, Elsevier,

Johnson, R., (1992) "Applied Multivariate Statistical Analysis", Prentice Hall.

Julio, P. and Adem, K. (2009) Data Mining and Knowledge Discovery in Real Life Applications, ISBN 978-3 -902613-53-0, pp. 438, I-Tech, Vienna, Austria.

Kurgan, L. and Musilek, P. (2006); "A survey of Knowledge Discovery and Data Mining process models", The Knowledge Engineering Review. Volume 21 Issue 1, pp 1 - 24, Cambridge University Press, New York, NY, USA.

Megaputer Intelligence, Inc. (2002). Crime Pattern Analysis: Megaputer Case Study. http://www.elon.edu/facstaff/mconklin/cis230/cases/crime_pattern_case.pdf

Milan, K., Sunila, G.,(2011) "Comparative Study of Data Mining Classification Methods in cardiovascular Disease Prediction", IJCST, Vol. 2, Issue 2, pp. 304-308,.

Naisbitt J. (1986) " *Megatrends"*, 6th ed., Warner Books, New York.

Otto, G. and Ukpere, W. I., (2012) "National security and development in Nigeria", African Journal of Business Management Vol.6 (23), pp. 6765-6770.

Taheri, S., Year Wood, J., Mammadov M. Seifollahi S. (2014) "Attribute Weighted NaïveBayes Classifier Usinga Local Optimization", *Neural Computing and Application*, Volume 24, Issue 5, pp. 995–1002.

Williams, G. J., Baxter, R. A., He H. X., Hawkins S., Gu L.,(2002) "A Comparative Study of RNN for Outlier Detection in Data Mining," IEEE International Conference on Data-mining (ICDM'02), Maebashi City, Japan, CSIRO Technical Report CMIS-02/102.

Wilson , O.W. (1963). *Police Administration*. USA, McGraw Hill Company.

Witten, I. and Frank, E. (2000). *Data mining: Practical Machine Learning Tools and Techniques with Java Implementations.* San Francisco: Morgan Kaufmann publishers.

ZhaoHui, T. and Jamie, M. (2005), "Data Mining with SQL Server 2005",Wiley Publishing Inc, Indianapolis, Indiana, 2005.