Detecting and Predicting Crimes using Data Mining Techniques: Comparative Study

Samah Samir
Department of Computer Science
Faculty of computers and information,
Menofia University, Egypt.
samah.zaheran@med.menofia.edu.eg

Eman M. Mohamed
Department of Computer Science
Faculty of computers and information,
Menofia University, Egypt.
eman.mohamed@ci.menofia.edu.eg

Hamdy mousa
Department of Computer Science
Faculty of computers and information,
Menofia University, Egypt.
hamdy.mousa @ ci.menofia.edu.eg

Abstract— **Crime is a major problem in our society where the highest priority is concerned with individuals, society, and government. Thus, it seems important to study factors and relations between the occurrence of different crimes to avoid more upcoming crimes. Crime prediction is a method of trying to study the causes and motives of crime and predict the times and places of its occurrence to reduce the commission of crimes that are expected to occur in the future. Data mining is an important way to facilitate the solution of future crime problems by investigating hidden crime patterns and historical crime data. Therefore, this study aims to analyze and discuss the various factors affecting the commission of crimes and the methods that are applied to predict future crimes and analyze their results. In this study, the model of crime prediction is proposed which is based on some classification algorithms such as (NB, KNN, Decision Tree, random forest, Linear Regression, Logistic Regression, SVM), these classification algorithms are applied to four real data sets (Chicago dataset, Los Angeles dataset, Egypt dataset, United States dataset), Egypt data set was extracted primarily from the online website (Zabatak.com) and comparing between their scores. The experimental results showed that the Random Forest classifier achieves a high score on four data sets compared with other classifiers. Random Forest achieves %88 on the Los Angeles dataset, %92 on the Egypt dataset, %97 on the Chicago dataset, and 81.7% on the United States dataset.**

Keywords— **Crime prediction, data mining, classification, clustering, KNN, NB, SVM**

I. INTRODUCTION

Previously solving crimes was the responsibility of law enforcement and criminal justice authorities, But with the rapid development in the use of computerized systems to track crimes and identify their causes, locations and perpetrators of crimes, computer data analysts have begun to help police officers and investigators solve crimes quickly and identify criminals.Criminology is a branch of forensic science that studies crime as a phenomenon in the life of the individual and society, determines and explains the factors that led to its commission.

Day by day, we notice that the crime rate has increased rapidly, and yet the crime cannot be predicted because it is neither systematic nor random, so police stations must make records that contain all the data on the crimes committed to help data analysts to make predictions about crimes. Solving crimes and knowing their causes and their possible occurrence is one of the difficult tasks that require time, experience and human intelligence, and data mining is one of the most common techniques that can help solve crimes and better predict future crimes. Data mining technology can be applied to solve crimes that combine computer science and criminal justice. Crime is a disturbing social phenomenon that costs our society severely in many ways. Studies have shown that about 10% of criminals commit about 50% of crimes. Therefore, there is a need for a good understanding and analysis of the reasons for committing crimes to provide practical solutions to help reduce the incidence of crimes and predict the possibility and times of their occurrence.

Crime Analysis takes historical crime data to help predict when and where crime will be committed in the future. Forecasting future crimes is the process of discovering the crime rate changes from year to year and the reasons for committing the crime and anticipating these changes in the future. Crime can be predicted by studying the historical data of crimes and the reasons for their commission. There are many studies to help predict crime, such as environmental scanning and imagining the events of crimes to determine the future nature of crimes. Scientific methods can also be used to predict the future scope of the crime, the causes and places of its occurrence, and more specifically and clearly.

This study aims to help police stations to predict future crimes by applying several data mining techniques to real datasets. So, various classification algorithms such as (NB, KNN, Decision Tree, random forest, Linear Regression, Logistic Regression, SVM) are applied to four different real datasets and compared their outputs. Each dataset has its attributes that differ from the others. The experimental results showed that Random Forest classifiers achieve %88 on the Los Angeles dataset, %92 on the Egypt dataset, %97 on the Chicago dataset, and 81.7% on the United States dataset.

The rest of the paper is organized as follows: some related works are discussed in section II. present materials and methods are used in this work in section III. produced experimental results and discussion in section IV. finally, the conclusion of the study is introduced in section VI.

II. RELATED WORK

A. Abdo in [1] Presents an approach for crime prediction with the help of data mining techniques, RapidMiner was used to compare between three algorithms (naive Bayes, decision tree, and deep learning). To find which classifier will be used with ada-boost. Real data from Egyptian forensic databases, specifically forensic databases in Alexandria, which contain 72175 records almost from 2014 to 2018. It also aims at helping to make strategic decisions to reduce crimes. RapidMiner software application was used to analyze their collected dataset. It achieves acceptable accuracy (about 98%), the simulator provided a simple, easy use, and real-time interface to crime prediction. A. Abdo proposed a method to extract age and gender from the Egyptian national number.

Mrinalini Jangra in [2] proposed a model that aimes to help police officails and detectives to identfy and understand crime issuesMrinalini introduces a data mining technique by applying the K-nearest neighbour (KNN) and naïve Bayes classifiers. This approach is applied to real data from Anaconda. The applied model uses Python's objective-oriented language. Then comparing the performance of the two classifiers (Naïve Bayes and KNN). The simulation results show that the Naïve Bayes algorithm achieved high accuracy (87%) and takes less execution time (.02 seconds).

Amit Gupta in [3] Suggest a crime forecasting solution to help various agencies, police departments to predict the rate of accident and develop strategies, plans, and preventive actions to reduce crime rate.this study is applied on areal dataset for the city of Denver. The data set is analyzed using six different classification algorithms (Bayes Net, Naive Bayes, J48, JRip, OneR and Resolution Table). The outputs result that outcome from this study is(correct rating, incorrect rating, true positive rate, false positive rate, accuracy, recall, and F-measurement). These outputs are analyzed by applying two different testing methods (k-fold validation and partitioning). The output results shows that the decision table and JRip achived the highest number of correct incidents (73.71%) and (73.66%), where the OneR rating achived the lowest number of correct incidents (64.95%). Although JRip is the most accurate classifier, it took a maximum time of (21.2) seconds to generate the model. The naive Bayes algorthim builds in the fastest time (0.57) seconds. Nevine Labib in [4] Identified that the main purpose of this work was to introduce a data mining model for predicting factors that affecting crime incidence by applying three classification techniques namely, naïve Bayes, decision trees, and association rules. The data set was collected from the Egyptian ministry of interior. Labib referred that dataset from (Alexandria, Egypt) were collected it presents the period from 1996 to 2012, and data consist of criminals' personal information including (age, profession, mental and educational level, social class, crime areas, and crime types). Finally, the accuracy results of the used algorithms are as follows Association Rules (98.11%), Naïve Bayes (97.81%), Decision Tree (92.07%).

J.Kiran in [5]Proposed a data mining model for crime Prediction in India. By applying two different algorithms namely, naïve Bayes and k- Nearest-Neighbour (KNN) classifiers. The proposed techniques are implemented in real data from Anaconda. The proposed technique is implemented using python language.the output results of the proposed technique are compared with existing techiques in two terms(execution time ,accuracy).The results showed that Naive bayes achives high accuracy(87%) as compared to KNN(77%). Prajakta Yerpude in [6] Present a model for predict features that affect the increase or decrease crime rates in a certain region.By applying different data mining classifiction algorithms namely(Decision Trees , Random Forest , Naïve Bayes , and Linear Regression).the model are applied on the chicago city crime dataset.datasets are avialble on UCI repository.the results showed that Random forest achived the high accuracy(83.39%),recall (84.86%),precision (88.30%) and F1(86.54%) .linear regression acheived low accuracy compared with other algorthims.

Pratibha in [7] Present a model that can help in predicting the time,the palce and the type of crime has high probability of occurrence.they depend on some classifiction algorthims namely(K-nearest Neighbour,Support Vector Machine (SVM),Decision Tree ,Extra Tree and Artificial Neural Network (ANN)).classifiction helps in the process of extract the same feature and predict future trends of crime based on Historical crime data.this study apply the model on the crime data set of San Francisco city.This dataset is available on the online websit(kaggle.com in CSV format) .The result showed that Decision tree, KNN and the Extra tree are working best with optimal training and good accuracy in this dataset.but SVM takes long time in the step of training dataset.

III. MATERIALS AND METHODS

This section describes the crime datasets, the tool used, classification algorithms, research methodology used in this study.

A. Dataset Description

The crime data set used in the study is acquired from (https://data.world/datasets/open-data) website. The study contains four different datasets. The first dataset for the city of Chicago, this dataset represents the reported incidents that occurred in the city of Chicago for the year 2016. This data was extracted from the Chicago Police Department. It contains 17 attributes (ID, Case Number, Date, Block, IUCR, Primary type, Description, Location description, Arrest, Domestic, Beat, District, Ward, Community Area, FBI code, Latitude, Longitude) as shown in TABLE 1. The second data set for the City of Los Angeles, this data set reflects reported incidents in the City of Los Angeles that are in the year 2010. It contains six attributes (DR Number, Date Occurred, Crime Code, Crime Code Description, Premise Code, Premise Description) the description of these attributes is shown in TABLE 2. The third data set for Egypt, data was extracted primarily from (Zabatak.com). The dataset consists of about 2,000 crimes from different regions across Egypt. It contains five attributes (x, y, date, sub cat, cat) the description of these attributes is shown in TABLE 3. The fourth data set is for Communities within the United States. The data comprises a range of socio-economic data for societies in the United States from 1990 to 1995, it contains 128 attributes some of them shown in TABLE 4.

B. Tool Used

A software tool and platform that is used in the implementation are listed as follows.

- Anaconda python software editor for running python codes.
- Platform used is (windos8.1 pro, memory installed RAM 6.00GB,processor intel(r) core (TM)i5-5200uCPU @2.20GHz, system type 64bit OS)

- *Classification Algorithms*

Classification is a popular supervised learning technique of data mining where data was categorized into a certain number of categories or classes. Classification can be performed on structured or unstructured data. The input data is grouped into certain number of classes. Each class has a class label and number of attributes.Data can be divided in two types (training dataset and testing dataset).

TABLE 1 Chicago Crime Dataset Description.

Column Name	Type	Description
ID	Integer	identifier for the record in Dataset.
Case Number	String	It is a number that the Chicago Police records a number for each incident that is unique.
Date	Date /time	The date of the accident.
Block	String	The detailed address of the accident. It should be in the same block.
IUCR	String	It is a secondary define for a uniform crime reporting law ,it is subset of main description.
Primary type	String	Basic description of the IUCR code.
Description	String	It is a secondary description of the IUCR code, It is a subcategory of the main description.
Location description	String	It is a detailed description of the place where the crime take place.
Arrest	Boolean	Indication that the perpetrator of the accident has been arrested or not.
Domestic	Boolean	Indicates whether the incident was related to domestic violence.
Beat	Integer	Refers to the beating in which the accident occurred.
District	Integer	Indicates the police area where the accident took place.
Ward	Integer	It is the city council district where the crime take place.
Community Area	Integer	Represent for the community area where the crime takeplace.
FBI code	String	Refers to the classification of crime according to the national reporting system.
Latitude	Decimal	Refers to the latitude of the location where the crime occured.
Longitude	Decimal	Refers to the longitude of the location where the crime occured.

TABLE 2 LOS Angeles Crime Dataset Description.

Column Name	Type	Description
DR Number	integer	Represent number of records.
Date Occurred	Date time	Data where the incident occurs with the format (MM/DD/YYYY).
Crime Code	integer	Reprsent the crime committed.
Crime Code Description	String	Describe the Crime Code provided
Premise Code	Integer	Represent The type of vehicle, or location where the crime occurred.
Premise Description	String	Represent the Premise Code .

TABLE 3 Egypt Crime Dataset Description.

Column Name	Type	Description
x	decimal	Latitude
Y	decimal	Longitude
date	Date time	Date when the incident occurred
sub cat	integer	Subcategory of crime. The data set has 26 subcategories.
Cat	integer	Category of crime. Dataset has 8 main categories.
x	decimal	Latitude

TABLE 4 United States Crime Dataset Description.

Column Name	Type	Description
state	nominal	Represent US state by number.
County	number	It is county number code.
community	number	Number code for the community in US.
Community name	string	Represent community name.
Fold	number	fold number for (10-fold cross-validation) it is useful for debugging, testing.
population	number-decimal	population for the community in US.
Household size	number-decimal	mean people per household in US.
Race pct black	number-decimal	Represent percentage of the population lives in African American.
Race Pct White	number-decimal	Represent percentage of the population lives in Caucasian.
Race Pct Asian	number-decimal	Represent percentage of the population lives in Asian heritage.
Race Pct Hisp	number-decimal	Represent percentage of the population lives in Hispanic heritage.

The main objective of a classification problem is to determine which category or class the new data will fall into. There are various classification algorithms such as (Support vector machine, K-Nearest Neighbor, Decision tree, Naïve Bayes, Random Forest Classifier, Linear Regression, Logistic Regression). They discussed as shown in TABLE 5. Then here I describe all the algorithms I utilized.

a) Naïve Bayes [12]

Naive Bayes classifier is a family of simple probabilistic classifiers based on an application of Bayes' theorem with strong (naive) independence assumptions between features. It is a supervised learning algorithm, which used for solving classification problems. Naive Bayes is one of the simple and most effective Classification algorithms which helps in building fast machine learning models that can make quick predictions. They are the simplest Bayes network models. But it can be combined with the estimation of the kernel density and achieve higher levels of accuracy.

b) The decision tree [17]

A decision tree is a supervised machine learning algorthim,it is used for classification and regression tasks.In decision trees, nodes represent the features of dataset and branches represent decision rules. To increase accuracy, multiple trees are used together in ensemble methods(Bagging,Random Forest classifier,Boosted trees,Rotation Forest).The goal of decision tree is to create a model that able to predicts the value of a target variable based on multible input data.

c) Random Forest [16]

Random forest is a supervised machine learning algorthim.It is an ensemble learning method for Classification and Regression problems. Random forest constructing a multitude of decision trees at training time,and takes the average to improve the prdictive accuracy of dataset.the output of the random forest is the class selected by most trees.

d) K-Nearest Neighbors [15]

The K-nearest neighbor algorithm falls under the category of supervised Machine Learning algorithms.It is used for classifiction and regression tasks.it assumes similarity between the new data and available data, then puts the new data into the category that is most like the available categories.it is classifies data based on similarity.this means when find new data it can be easy to classified into suitable category.Also it is used to calculate missing quantities data .

e) Support Vector Machine [13]

SVM is one of the most popular a supervised machine learning algorthim,that can be used for both classification or regression problems.But it is prefferrd for classifiction problems.In the SVM algorithm, we plot each data item as a point in n-dimensional space (where n is a number of features you have).so when new data point come it is easily put in the correct category.

f) Linear Regression

Linear regression is a statistical method used to interpret one variable over another. By showing a linear relationship between a dependent variable (X) and one or more independent variables (X), that is why it is called linear regression. Since linear regression is used to illustrate the linear relationship between two variables, it finds how the value of the dependent variable changes according to the value of the independent variable.

g) Logistic Regression [14]

Logistic regression is a statistical model that is used to model the probability of a certain class or event existing.it is used for predicting categorical dependent variable using another set of independent variables. Logistic regression outcome must be categorical or discrete value. Outcome can be either (0 or 1, Yes or No, true or false).it is useful for binary and linear classification problems. It gives probabilistic values which range between 0 and 1.

TABLE 5
DESCRIPTION OF CLASSIFICATION ALGORITHMS THAT USED IN THIS STUDY

Classifier	Description
Naïve Bayes	This supervised machine learning technique is a probabilistic classifier and uses the statistical method .
Decision tree	This Supervised machine learning technique that mostly it is popular for solving Classification tasks.by producing a sequence of rules that used for classifying data
Random Forest Classifier	This supervised machine learning algorthim that is constructed from a decision tree. For classification tasks, the output is the class selected by most trees.
Linear Regression	Linear regression is a commonly used type of predictive analysis, it attempts to model the relationship between two variables by fitting a linear equation to the observed data.
Logistic Regression	Logistic regression is a statistical model belonging to linear regression models that enables the modeling of a binomial variable in terms of a set of expected random variables, whether numerical or categorical.
K-Nearest Neighbors	It assumes similarity between the new data and available data, then puts the new data into the category that is most like the available categories.it is classifies data based on similarity.
Support Vector Machine	Representing training data as points in space divided into categories by an apparent gap as wide as possible.

C. *Research Methodology*

This study is based on predicting future crime based on four different data sets.Different classification algorithms were applied and compared their output according to accuracy, precision, and recall. This study is divided into two parts, first part is applying four classification algorithms (K-Nearest Neighbors, Logistic Regression, Random Forest Classifier, Linear Regression) on the first three data sets (Los Angeles dataset, Chicago dataset, Egypt dataset) and comparing between scores for algorithms. The second part is applying five different classification algorithms (Decision Trees, Gaussian NB, SVM, Gradient Boosting, Random Forest) to the fourth data set (United States dataset) and comparing their scores achieved.

IV. RESULTS AND DISCUSSIONS

Our study was divided into two parts the first part is to apply four different classification algorithms (KNN, Random Forest classifier, Linear regression, and Logistic regression) into three different datasets (Los Angeles dataset, Chicago dataset, Egypt dataset) and compare between their scores. The feature of date was depended on in three different datasets.

The performance of KNN, Random Forest classifier, Linear regression, and Logistic regression are analyzed in terms of accuracy as shown in figure 1. Random Forest classifier was achieved a high score in three different data sets compared with other classifiers. Then note that Logistic Regression achieves a high score compared with K-Nearest Neighbors, Linear Regression has a low score compared with other classifiers.

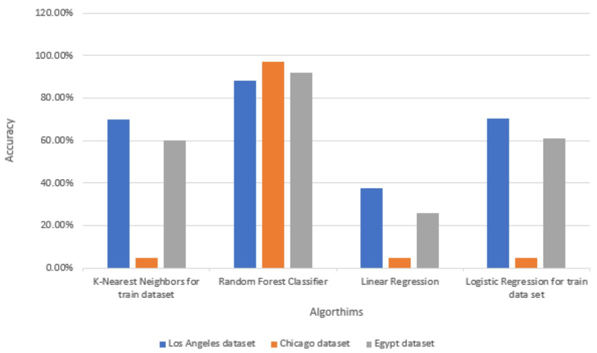

Fig. 1 Accuracy for Four Classification Algorithms for Three Datasets.

From the Linear Regression algorithm, the relations were found between month or week or day and crime code using the count of crimes in month or week or day. Here note that January month has a high number of crimes compared with other months, it has 1193 crimes. And note that September month has a low number of crimes, it has 881 crimes. As shown in figure 2.

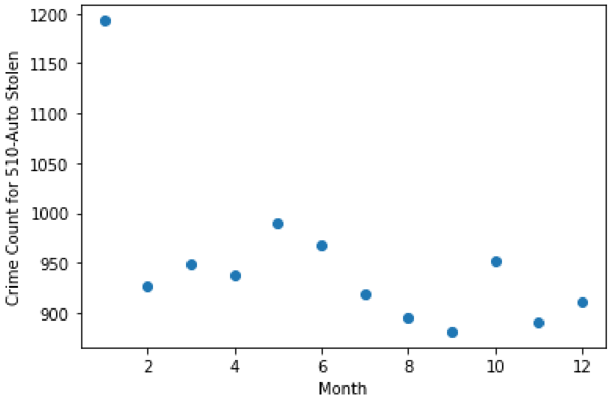

Fig.2 Relation Between A Month and Crime Code.

The second part of this study is to apply five different classification algorithms (Decision Trees, Gaussian NB, SVM, Gradient Boosting, Random Forest) to the United States dataset, the United States datasets have different features from the other three datasets it contains 128 attributes. The performance of five Classifier algorithms applied on the fourth data set (United States dataset) is analysed in terms of accuracy as shown in figure 3. Random forest was achieved a high score compared with others. Then SVM has a high score compared with Gradient Boosting and Decision Trees and Gaussian NB has a low score compared with other algorithms.

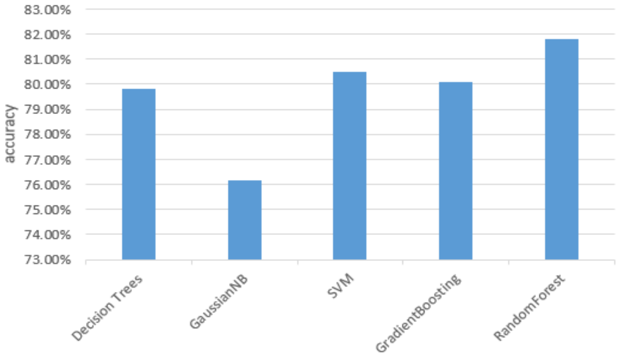

Fig. 3 Accuracy of Five Classification Algorithms.

The performance of five Classifier algorithms applied on the fourth data set (United States dataset) is analyzed in terms of precision as shown in figure 4. Naïve Bayes has achieved a high score comparing with others.

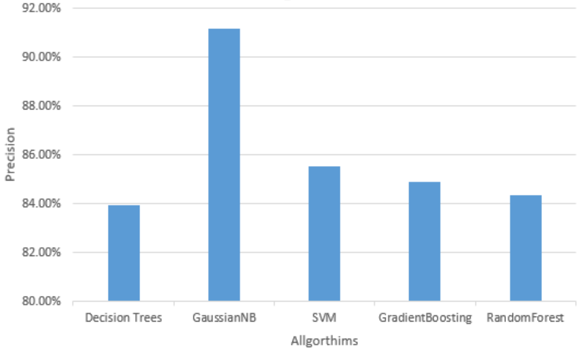

Fig. 4 Precision for Five Classification Algorithms.

The performance of five Classifier algorithms applied on the fourth data set (United States dataset) is analysed in terms of recall as shown in figure 5. random forest was achieved a high score compared with others.

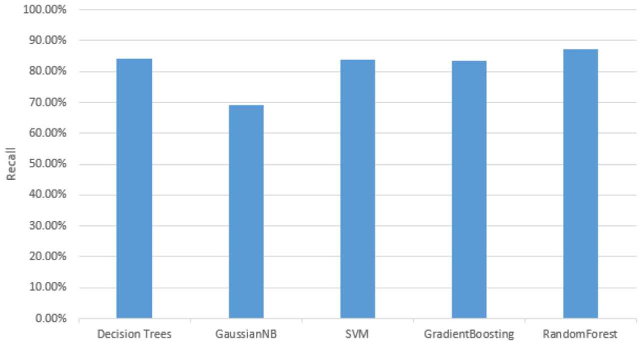

Fig. 5 Recall for Five Classification Algorithms.

From these results, the random forest was achieved a high score although in the second part the data set was changed, and its features also achieve a high score.

V. Conclusions

Crime prediction and data mining techniques have become the current trend in our society. It attempts to reduce crime occurrence by predicting future crime depending on the available crime dataset. This study is helpful for various agencies, police departments adding them to prevent future crime from occurring. In this study, different classification algorithms are applied to different four data sets and compared their scores. In the first part where four classification algorithms were applied on three different data set, the random forest classifier was achieved a high score on three data sets. Random Forest achieves %88 on the Los Angeles dataset, %92 on the Egypt dataset, and %97 on the Chicago dataset. In the second part where five classification algorithms were applied to the United States dataset, the random forest was achieved also a high score compared with other classifiers, which achieves 81.7%. Till now our crime prediction model was trained that there was a plan to include more factors to improve outcomes.

REFERENCES

[1] A. Abdo, Hanan Fahmy, and Amir Abobaker Shaker" Mining Forensic Medicine Data for Crime Prediction "International Journal of Computer Science and Information Security (IJCSIS), Vol. 17, No. 6, June 2019.

[2] Mrinalini Jangra and Ms. Shaveta Kalsi "Crime Analysis for Multistate Network using Naive Bayes Classifier " International Journal of Computer Science and Mobile Computing (IJCSMC), Vol. 8, Issue. 6, June 2019, pg.134 – 143.

[3] Amit Gupta, Ali Syed, and Malka N. Halgamuge" A Comparative Study of Classification Algorithms using Data Mining: Crime and Accidents in Denver City the USA "International Journal of Advanced Computer Science and Applications · August 2016.

[4] Nevine Makram Labib and Brigadier-General Wael Kamal Arafa "A Proposed Data Mining Model to Enhance Counter- Criminal Systems with Application on National Security Crimes "International Research Journal of Computer Science (IRJCS) Issue 7, Volume 2 (July 2015).

[5] J.Kiran and Kaishveen. K"Prediction Analysis of Crime in India Using a Hybrid Clustering Approach" IEEE Xplore Part Number: CFP18OZV-ART; ISBN:978-1-5386-1442-6 I-(2018).

[6] Prajakta Yerpude and Vaishnavi Gudur "PREDICTIVE MODELLING OF CRIME DATASET USING DATA MINING" International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.7, No.4, July 2017.

[7] Pratibha, Akanksha Gahalot, Uprant, Suraina Dhiman, Lokesh Chouhan" Crime Prediction and Analysis "Conference Paper · February 2020.

[8] Rasoul Kiani, Siamak Mahdavi, Amin Keshavarzi" Analysis and Prediction of Crimes by Clustering and Classification" (IJARAI) International Journal of Advanced Research in Artificial Intelligence Vol. 4, No.8, 2015.

[9] H. Benjamin Fredrick David,A. Suruliandi "SURVEY ON CRIME ANALYSIS AND PREDICTION USING DATA MINING TECHNIQUES ICTACT JOURNAL ON SOFT COMPUTING, APRIL 2017, VOLUME: 07, ISSUE: 03".

[10] Ayisheshim Almaw, Kalyani Kadam "Survey Paper on Crime Prediction using Ensemble Approach "International Journal of Pure and Applied Mathematics Volume 118 No. 8 2018, 133-139.

[11] Falade Adesola, Ambrose Azeta, Isaac Odun-Ayo "Systematic Literature Review of Crime Prediction and Data Mining "journal home page: http//iieta.org/journals/rces Vol.6, No.3, September 2019, pp.56-63.

[12] Daniel Berrar"Bayes' Theorem and Naive Bayes Classifier"January 2018.

[13] Mariette Awad, Rahul Khanna" Support Vector Machines for Classification"January 2015.

[14] Sandro Sperandei" Understanding logistic regression analysis" Article in Biochemia Medica · February 2014.

[15] Viswanath Pulabaigari, Hitendra Sarma T"An Improvement to k-Nearest Neighbor Classifier" Conference Paper · September 2011.

[16] Adele Cutler, David Richard Cutler, John R Stevens"Random Forests" January 2011.

[17] Lior Rokach, Oded Maimon" Decision Trees" January 2005.

[18] UCI Machine Learning Repository. [online] https://archive.ics.uci.edu/ml/datasets/communities+and+crime, available at 30/6/2020.

[19] Chicago data portal [online] https://data.cityofchicago.org/Public-Safety/Crimes-2016/kf95-mnd6/data, available at 30/6/2020.

[20] Data.world[online] https://data.world/search?context=community&q=los+&type=all, available at 30/6/2020.