IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 23, Issue 5, Ser. I (Sep. –Oct. 2021), PP 24-31
www.iosrjournals.org

Efficiency Measurement Application of Artificial Intelligence/Language for Predicting/Detecting Crime

Umidjon Makhmudov And Ma Bin (马斌)

School of Computer Sciences
North China University of Water Resource and Electric Power
Zhengzhou 450046, China

Abstract:
Purpose-*This research manuscript examines the dimension of machine learning methods for detecting crime and predicting future crimes using 2 data sets. ZeroR, Naive Bayes, Linear Regression, Boosted Trees, Decision Tree and K-Nearest are used for crime prediction on the basis of AUC values, K-Fold Cross Validation and F-measure metric.*
Design/methodology/approach- *To accomplish the rationale, study uses 2 data sets; Auto Theft and Rape Victims. Both the data set is key-in WEKA (Waikato Environment for Knowledge Analysis, developed at the University, Waikato, New Zealand). AUC values, K-Fold Cross Validation and F-measure metric are used to predict the estimation of six artificial intelligence/language application in order to measure the performance.*
Findings- *The study finds that all six methods are better than blue-collar methods for Crime prediction in the areas where the crimes already happened and out of the six methods accuracy of K-Nearest is uppermost and accurate to predict and detect the predict the crime of auto theft and rape.*
Originality/value- *This study offers immediate theory test six classifications of machine learning using three exclusive parameters. Proportional analysis is given to understand the performance of all methods.*
Social Implications- *Result obtained from the study can be implemented in assistance of police work and security agency as accurate information and precautions can be implemented in particular areas. This revolutionizes in crime prediction by cyber cell departments and help out many target victims to avail proper security and minimize threats in society.*
Key Words: *Naive Bayes, Linear Regression, Boosted Trees, Decision Tree and K-Nearest, Crime Prediction, Artificial Languages and Crime detection applications*

Date of Submission: 02-09-2021 Date of Acceptance: 15-09-2021

I. Introduction:

Technologies based on artificial intelligence (AI) and machine learning (ML) have seen dramatic increases in capability, accessibility and widespread deployment in recent years, and their growth shows no sign of abating. While the most visible AI technology is marketed as such (e.g. 'personal assistants' such as Amazon Alexa, Apple Siri and Google Home), learning-based methods are employed behind the scenes much more widely. From route-finding to language translation, biometric identification to political campaigning, and industrial process management to food supply logistics, AI saturates the modern connected world at many levels (Benaich and Hogarth 2019). These days artificial intelligence is in major demand for detecting crimes and solving crimes with similar precedence, this study measures different methods of artificial intelligence for performance check which emphasizes which application is more accurate one in predicting crimes of rape and auto theft.

A computer aided system and software that are used in different fields, generally consists of a information technology and a method for solving an intended problem. On the basis of the query posted to the system, it provides assistance to the police and agencies to detect crime immediately and react on it. Information and knowledge that already is posted on computer aided system for prediction purpose are from the instances of past crimes. Agile and aggressive research work on the ground breaking thoughts of Police Assistance System has proven computers are detecting and predicting crimes more efficiently then human minds. The importance of AIC as a distinct phenomenon has not yet been acknowledged. The literature on AI's ethical and social implications focuses on regulating and controlling AI's civil uses, rather than considering its possible role in crime (Kerr 2004). Furthermore, the AIC research that is available is scattered across disciplines, including socio-legal studies, computer science, psychology, and robotics, to name just a few. This lack of research

centred on AIC undermines the scope for both projections and solutions in this new area of potential criminal activity.

Recently, computer bases statistical analysis tools have made it easy for analyst to predict on the basis of econometric and statistical analysis in any field. WEKA and SPSS are most commonly used tools to predict series of data. Artificial Intelligence/Language and Machine Learning languages are used for the purpose of the analysis. Supervised learning is the most common form of machine learning scheme used in solving the engineering, clinical and other problems. It can be thought as the most appropriate way of mapping a set of input variables with a set of output variables. The system learns to infer a function from a collection of labeled training data. The training dataset contains a set of input features and several instance values for respective features. The predictive performance accuracy of a machine learning algorithm depends on the supervised learning scheme. The aim of the inferred function may be to solve a regression or classification problem. There are several metrics used in the measurement of the learning task like accuracy, sensitivity, specificity, kappa value, area under the curve etc. This study chooses six different classification models for classifying the nature of crimes and predicting that crimes, which separately are ZeroR, Naive Bayes, Linear Regression, Boosted Trees, Decision Tree and K-Nearest. These models are also classic algorithms from the area of machine learning. Then, the R programming language is employed as an essential tool to predict the nature of crime and prediction of that crime. One dataset from auto theft that was created by Patrício, and another dataset from rape victims that was created by Dr. William H. Wolberg are used in the study. Te wide range of legitimate AI applications includes systems for crime prevention and detection (Dilek et al. 2015; Li et al. 2010; Lin et al. 2017; McClendon and Meghanathan 2015), but the technology also has potential for misuse in the service of criminal activities (Kaloudi and Li 2020; Sharif et al. 2016; Mielke and Chen 2007; van der Wagen and Pieters 2015). Therefore it is essential to check the applications performance used in crime detection and prediction especially in auto theft and rape cases which are common in many countries.

II. Literature Review:

Data mining is part of the interdisciplinary field of knowledge discovery in databases (S. M. Nirkhi, 2012). Data mining consist of collecting raw data and, (through the processes of inference and analysis); creating information that can be used to make accurate predictions and applied to real world situations such as the stock market or tracking spending habits at the local Wal-Mart. It is the application of techniques that are used to conduct productive analytics. Data mining software packages such as the Waikato Environment for Knowledge Analysis (WEKA), the data mining software package used in this project, are used to conduct analysis of data sets by utilizing machine learning algorithms. The five tasks that these types of software packages are designed for are as follows: (i) Association - Identifying correlations among data and establishing relationships between data that exist together in a given record [(S. M. Nirkhi, 2012) and (E. W. T. Ngai, 2008). (ii) Classification - Discovering and sorting data into groups based on similarities of data (S. M. Nirkhi, 2012). Classification is one of the most common applications of data mining. The goal is to build a model to predict future outcomes through classification of database records into a number of predefined classes based on a certain criteria. Some common tools used for classification analysis include neural networks, decisions trees, and if-then-else rules (E. W. T. Ngai, 2008). (iii) Clustering - Finding and visually presenting groups of facts previously unknown or left unnoticed (S. M. Nirkhi, 2012). Heterogeneous data is segmented into a number of homogenous clusters. Common tools used for clustering include neural networks and survival analysis (E. W. T. Ngai, 2008). (iv) Forecasting - Discovering patterns and data that may lead to reasonable predictions (S. M. Nirkhi, 2012). It estimates the future value based on a record's pattern. It deals with continuously valued outcome. Forecasting relates to modeling and the logical relationships of the model at some time in the future (E. W. T. Ngai, 2008). (v) Visualization - Enabling researchers to rapidly and efficiently locate vital information that is of interest (S. M. Nirkhi, 2012).In the preparatory review phase of the project, examples were collected of existing or predicted interactions between AI and crime, with both terms interpreted quite broadly. Cases were drawn from the academic literature, but also from news and current affairs, and even from fiction and popular culture, which can be considered as a barometer of contemporary concerns and anxieties. Examples were organized into three non-exclusive categories according to the relationship between crime and AI:

- Defeat to AI—e.g., breaking into devices secured by facial recognition.
- AI to prevent crime—e.g., spotting fraudulent trading on financial markets.
- AI to commit crime—e.g., blackmailing people with "deep fake" video.

Machine learning is not new to crime prediction research. K-Nearest and decision trees have been used in crime detection and prediction for nearly 10 years (Dante, 2005). Today machine learning methods are being used in a wide range of applications ranging from detecting and classifying different levels of crimes before happening and providing provisional security to threats (Liotta, 2007) to the classification of crime rate, ratios

and nature (Zhou, 2015). In other words machine learning has been used primarily as an aid to crime prediction and detection (McCarthy, 2004). However crime detection is different than crime prediction, it may be considered artificial intelligence is also use in AI Crimes (Hagerty, 2005). The use of computers (and machine learning) in crime prediction is part of a growing trend towards personalized, predictive securities (Neil, 2004). (Ginesh, 2014)Proposed a new model for compose decision trees using interval-valued fuzzy membership values. Most existing fuzzy decision trees do not consider the concerned associated with their membership values; however, precise values of fuzzy membership values are not always possible. In this research, we implemented the Linear Regression, Additive Regression, and Decision Stump algorithms using the same finite set of features, on the communities and crime un normalized dataset to conduct a comparative study between the violent crime patterns from this particular dataset and actual crime statistical data for the India as auto theft and rape cases are more common in India. In (Alkesh Bharati, 2018), crime prediction is done on Chicago data set in which various machine learning models are used. Comparison of models like KNN, Naïve Bayes, and SVM is done this paper. It is seen that prediction varies depending upon the dataset and features that have been selected. The prediction accuracy found in (Alkesh Bharati, 2018) is 78% for KNN, 64% for Gaussian NB, 31% for SVC. Auto regressive integrated Moving average models were used in (E. Cesario, 2016) to make machine learning algorithms to forecast crime trends in urban areas. One of the major problems in crimes is detecting and analyzing the pattern of crimes. Understanding datasets is also an important concept in this case. We surely want to accurately predict so that we don't waste our resources due to false signals. In paper (M. V. Barnadas, 2016), Algorithms like KNN and neural networks are developed, tested and crime prediction is done on San Francisco. It is observed that many machine learning models are implemented on datasets of different cities having unique features, so predictions are different in all cases. Classification models have been implemented on various other applications like prediction of weather, in banking and finances also in security (M. V. Barnadas, 2016). Most of the research in crime prediction is finding the location of crimes and doing analysis based on proposed area-specific models using geographical data. Based on the review and studying previous work, KNN classification and decision tree models is shown to be giving high accuracy so we choose to use the same to predict crimes in Vancouver city.

STATEMENT OF PROBLEM:

More engaged and intense phenomenon of integration of computer aided intelligent crime prediction systems and security systems require better investigative approach as life and death situations are involved in criminal activities. For the just cause following problems are addresses in this study.

1. What machine learning algorithms are involved in detecting and predicting the level of the crimes as in auto theft and rape cases?
2. Extent to which the following algorithms imply to the predicting the criminal activities of auto thefts and rapes: ZeroR, Naive Bayes, Linear Regression, Boosted Trees, Decision Tree and K-Nearest
3. What are the techniques to evaluate the performance of above said methods and to what extent they are true?
4. To what extent following predictors can evaluate the performance check on Artificial Learning Algorithms: K-Fold Cross Validation, F-Measure Metric and AUC values

HYPOTHESES:

$H1_0$: There is no significant relationship amongstthe predictors K-Fold Cross Validation, F-measure metric and AUC values for performance evaluation of machine learning algorithms for predicting crimes.

$H2_0$: There is no significant relationship amongstthe methods of machines learning algorithms to detect and predict auto theft crimes and rape crimes.

$H3_0$: There is no variation in the performances of the machine learning algorithmic models used for predicting and detecting crimes.

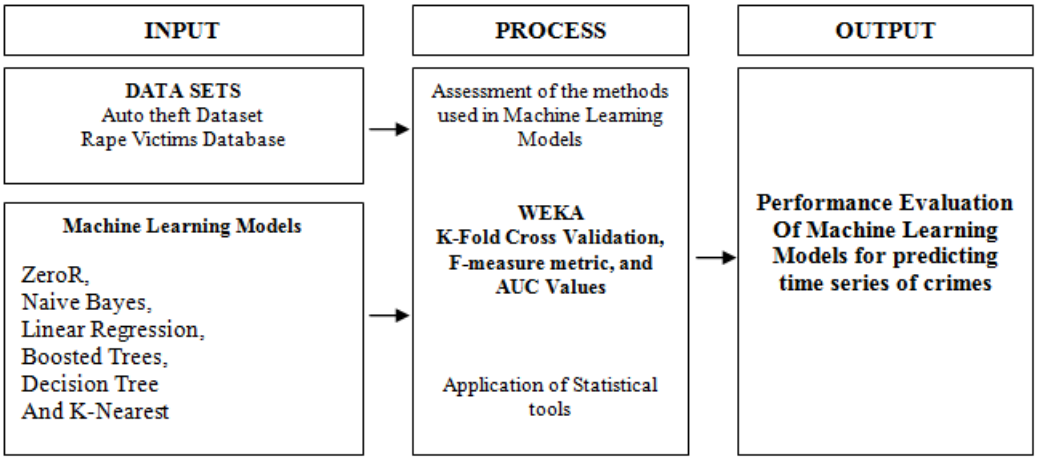

Figure 1: *Research Paradigm*

III. Data Set Source:

All the models introduced in this study are verified on two datasets, which are separately from AT and RV. The AT data is composed of 116 instances, reported by March 6th, 2019, with 10 attributes for each case. The independent attributes are: age, Indian cities and provinces where the auto theft occurred, all of which are anthropometric data and parameters. The dependent attribute is classification, which is presented by integer 1 and 2, where 1 stands for crime within one month and 2 stands for crime occurred above one month denoting time of actual crime. The RV data involves 699 instances, reported by July 15th, 2015. This dataset contains sample code number and 10 attributes for each case. The independent attributes are Indian cities and rape victim ages, ethnicity, religion, area range, profession and martial statuses. All of them are represented by the integer in the range of 1 to 10. The dependent attribute is class, which presented by integer 2 and 4, where 2 stand for forced rape and 4 stands prostitution rapes.

IV. Methodology And Design:

Study used six classifiers to evaluate the performances of prediction of crimes, in model comparing results of all six machine learning algorithms. Main focus here to map the attributes with values subjectively objects values are matched with the object attributes. The goal is to guarantee highly accurate and stable performance of the classification task. K-Nearest is operated by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes or means prediction of individual trees. We used K-Nearest at first to build the classification model to predict the class of crimes either occurred in rapid rates or not occurred in rapid rates in Indian States. Ratio of both sets are distributed in subsets such 70% to 30%. Study implies 70 percentages as training set and 30 percentages as testing data. The training data is applied to train the classification model by setting the parameters of K.N in order to better fit the model. For another, the test data is applied to test the predictability of the trained model through K.N. Moreover, the study validates the efficiency of this classification model. After obtaining the result of prediction accuracy, the study compares K.N with other machine learning models, such as ZR, N.B, LR, DT and BT.

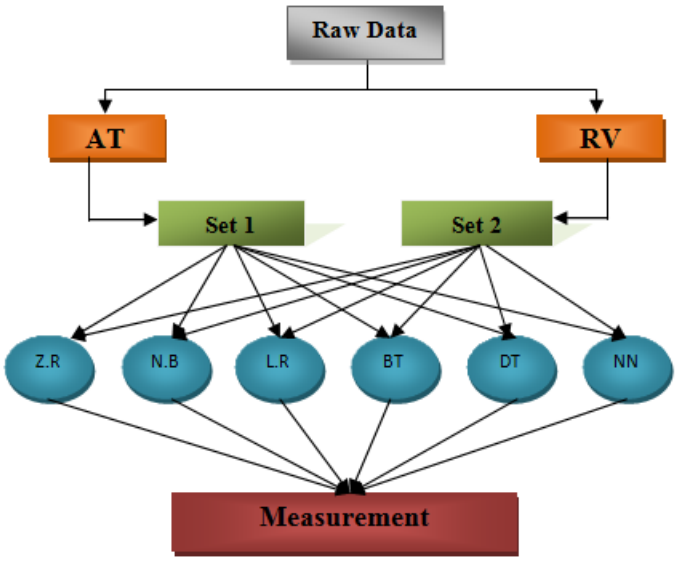

Figure 2 *Flowchart of Model*

EXPERIMENTS AND INTERPRETATIONS

This section focuses on the evaluation of comparative experiment based on five classification models, whose performances of over two crime datasets will also be presented. Data set AT classification has occurrence weight of 64 counts and non occurrence weight of 52 counts in total they are 116 count distributed in different classes and sets. It is intended to apply all six learning methods to analyze the database, which are ZR, N.B, LR, DT, BT and K.N. This study combines both accuracy and F measure metric as the index for choosing the primary analytic model. Accuracy emphasizes on the performance of the classifier and it calculates the proportion that true positive items occupy among the sum of true positive items and false positive items. The score of F-measure metric is the harmonic average of the precision and recall. The higher F-measure signifies the higher efficiency of the models, where 1 is the best value of F-measure while 0 is the worst. The F-measure metric values and the prediction accuracy of AT & RV data are shown in Table 1 and Table 2 below.

Classification mode	**Z.R**	**N.B**	**L.R**	**B.T**	**D.T**	**K.N**
Accuracy	0.55	0.53	0.56	0.50	0.55	0.60
F-Measure Metric	0.71	0.52	0.68	0.52	0.70	0.72

Table 1 AT Accuracy and F-measure Metric values

Classification mode	**Z.R**	**N.B**	**L.R**	**B.T**	**D.T**	**K.N**
Accuracy	0.62	0.43	0.55	0.57	0.61	0.74
F-Measure Metric	0.77	0.56	0.62	0.51	0.62	0.79

Table 2 RV Accuracy and F-measure Metric values

Figure 3 is a histogram which is used to compare the accuracy and F-measure metric directly of six different models in AT dataset. Same with figure 6 histogram is used to compare the accuracy and F-measure metric directly of six different models in RV dataset.

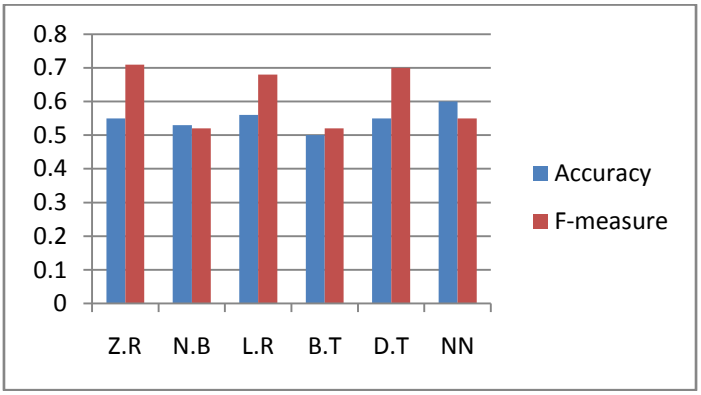

Figure 3 accuracy and F-measure metric of six classification models for AT data

As shown in both table and figure that K.N, K-Nearest has been the highest predicting method for AT data set that has comparatively predict highest number of accurate auto theft crimes with occurrence rate and non occurrence rate from 116 cases, same specifications apply RV figure 4 for accuracy and F measure metric.

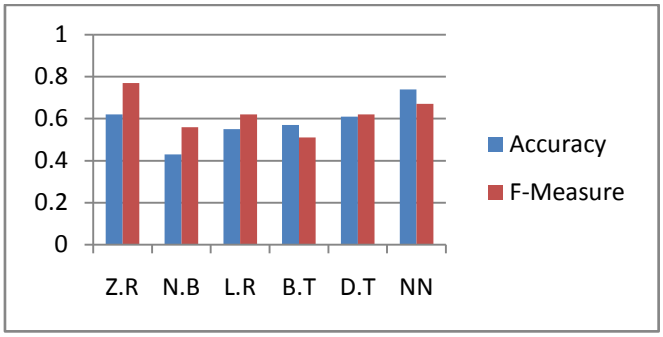

Figure 4 Accuracy and F-measure metric of six classification models for RV data

Figure 3 exhibits the values of RV data analysis of six methods of machine learning algorithm which implies that K-Nearest is the fittest way of predicting the predicting crimes amongst the six algorithms though accuracy of Decision Tree algorithm is almost the same in accuracy measurement but for Measure Metric K-Nearest is more efficient therefore overall it is concluded that K-Nearest and RF is better and efficient Natural Artificial intelligence pseudo code that can predict crimes of rape and auto theft better than other 6 methods which are: ZeroR, Naive Bayes, Linear Regression, Boosted Trees, Decision Tree and K-Nearest. For the purpose of verifying the performance of ensemble, the study performs prediction on different randomly split training and testing data 50 times.

Classification mode	**Z.R**	**N.B**	**L.R**	**B.T**	**D.T**	**K.N**
Accuracy	0.43	0.60	0.61	0.58	0.43	.65

Table 3 AUC value of AT data

Expression is given in form:

K-Nearest Model>>Naïve Bayes Model>Boosted Trees Model>Decision Tree Model>ZeroR Model

Table 3 exhibits the values of ROC curve through AUC values and identifies the area under curve, it implies that K-Nearest area under curve is greater than other methods of machine learning such as ZeroR, Naive Bayes, Linear Regression, Boosted Trees, and Decision Tree. AUC values for the curves are given in table 4.

Classification mode	**Z.R**	**N.B**	**L.R**	**B.T**	**D.T**	**K.N**
Accuracy	0.493	0.788	0.65	0.36	0.37	.80

Table 4 AUC value of RV data

Expression is given in form:

K-Nearest Model> Naïve Bayes Model > Regression Model > ZeroR Model >Decision Tree Model>Boosted Tree Model

Table 4 exhibits for WBCD data set indicate the values of AUC implies that amongst all the machine learning languages and classifiers K-Nearest covers prediction of auto theft and rape crimes more efficiently and intelligently. Classes for both the sets are 1 for malignant and 2 for benign. Curves are sectioned with malignant classification and implementation.

SCOPE AND LIMITATIONS:

Study is limited as it used only 2 datasets and three parameters because of limited time and academic requirements as this research study is fulfillment of the Masters of Computer Science in North China University of Water Conservancy Management Sciences, China. Therefore due to limited time and less budget only few parameters were selected without any expert help except of the supervisor which I am thankful for the concluding and remarkable skill transformations. Study is limited and could be of more useful if all three supervised, unsupervised and reinforced machine learning methods of artificial intelligence are used which in this case was not possible due to time constraints. Study has scope of finding more efficient model based on the performance evaluation of the methods used in predicting the crimes of auto theft and rape cases. Six methods are applied for the performance check all with three parameter values which is not been explored before and has promising outcomes.

V. Conclusion And Findings:

SUMMARY OF FINDINGS:

Study found that all six machine learning methods are efficient in predicting the crimes as compare to the manual methodology and conventional police methods of solving crimes and predicting crimes.

Study found that Accuracy values by using cross validation method on each of the algorithm in k fold 10 gives the output which can easily predict the class 1 occurrences and class 2 non occurrences for both data sets AT and RV.

Study found that F Measure Metric values of six methods of machine learning has produced output comparatively far more comprehensive and authentic then manual methods

Study found that all six methods of machine learning are exactly efficient in generating output when comparing each other K-Nearest has produced more efficient results than other methods.

Study found that modules created can posses any ratio of set information has the same results as the K-Nearest the better method amongst the Six Languages.

VI. Conclusion

Concluding expression for the overall analysis is summed up in expression for AUC values and other attributes is suggesting that K-Nearest is comparatively more efficient method of machine learning than other six methods.

K-Nearest Model>>Naïve Bayes Model>Boosted Trees Model>Decision Tree Model>ZeroR Model

K-Nearest Model> Naïve Bayes Model > Regression Model > ZeroR Model >Decision Tree Model>Boosted Tree Model

Study applied six classifiers after converting nominal to numerical values of data set in AT and RV. Six methods are ZeroR, Naive Bayes, Linear Regression, Boosted Trees, Decision Tree and K-Nearest. Purpose of implementing machine learning algorithms to dataset was to evaluate the performance of natural language and analyze which method is more efficient in crime prediction. Classes were selected as 1 “Occurrence of Crime” and 2 “Non Occurrences of Crime”. Datasets were acquired by authentic source and used in several other studies to identify the case of Crime prediction. Yixuan Li, Zixuan Chen (2018) applied 5 methods to evaluate the performance of Support Vector Machine, Artificial K-Nearest, Decision Tree, K-Nearest and Logistic Regression and resulted in identifying that K-Nearest provides befitting results in predicting the cased of tumor patients in breast cancer. This study used the nature of data sets with same number of attributes for crimes in Indian cities, dataset AT with 116 volunteers and 9 attributes and RV with 699 volunteers and 11

attributes is used in this study produced result and compared six different algorithms: ZeroR, Naive Bayes, Linear Regression, Boosted Trees, Decision Tree and K-Nearest. Study used Cross validation, F-measure and AUC values to compare mean result through histogram and ROC curves. Though there are limitations in study like lack of indices and only 2 data sets but result produced are of greater benefit as it not only serve the cause of humanity but also highlights the importance of technology in Security systems and Police work.

Study concluded that machine learning algorithms are efficient way of producing the results of the predicting crimes which can also be applied to other societal issues. Study also compares the result and interprets that performance of evaluation of six methods which resulted in K-Nearest as the highest ranked method for detection and prediction of crimes of auto theft and rape cases.

References:

[1]. Predict Crime | Predictive Policing Software," PredPol. [Online]. Available: http://www.predpol.com/. [Accessed: 22-Jun-2017].
[2]. Aghababaei and M. Makrehchi, "Mining Social Media Content for Crime Prediction," in Web Intelligence (WI), 2016 IEEE/WIC/ACM International Conference on, 2016, pp. 526–531.
[3]. Chitrakar, C. Zhang, G. Warner, and X. Liao, "Social Media Image Retrieval Using Distilled Convolutional Neural Network for Suspicious e- Crime and Terrorist Account Detection," in Multimedia (ISM), 2016 IEEE International Symposium on, 2016, pp. 493–498.
[4]. E. Sawyer and C. W. Monckton, "'Shoe-fit'-a computerised shoe print database," 1995.
[5]. A. Alexander, A. Bouridane, and D. Crookes, "Automatic classification and recognition of shoeprints," 1999.
[6]. Kumar, N. R. Pal, B. Chanda, and J. D. Sharma, "Detection of fraudulent alterations in ball-point pen strokes using support vector machines," in India Conference (INDICON), 2009 Annual IEEE, 2009.
[7]. Vargas, A. C. Bahnsen, S. Villegas, and D. Ingevaldson, "Knowing your enemies: leveraging data analysis to expose phishing patterns against a major US financial institution," in Electronic Crime Research (eCrime), 2016 APWG Symposium on, 2016, pp. 1–10.
[8]. Y. Yu, X. Wan, G. Liu, H. Li, P. Li, and H. Lin, "A combinatorial clustering method for sequential fraud detection," in Service Systems and Service Management (ICSSSM), 2017 International Conference on, 2017, pp. 1– 6.
[9]. "Crimes - 2001 to present - Data.gov." [Online]. Available: https://catalog.data.gov/dataset/crimes-2001-topresent398a4. [Accessed: 24-Jun-2017].
[10]. Wang and M. S. Gerber, "Using Twitter for Next-Place Prediction, with an Application to Crime Prediction," 2015, pp. 941–948.
[11]. Chen, Y. Cho, and S. Y. Jang, "Crime prediction using Twitter sentiment and weather," in Systems and Information Engineering Design Symposium (SIEDS), 2015, 2015, pp. 63–68.
[12]. Munasinghe, H. Perera, S. Udeshini, and R. Weerasinghe, "Machine Learning based criminal short listing using Modus Operandi features," 2015, pp. 69–76.
[13]. Gorr, A. Olligschlaeger, and Y. Thompson, "Shortterm forecasting of crime," Int. J. Forecast., vol. 19, no. 4, pp.
[14]. Weifa, "A SVM Text Classification Approch Based on Binary Tree," 2009, pp. 455–458.
[15]. L. Youwen, X. Shixiong, and Z. Yong, "A SupervisedLocal Linear Embedding Based SVM Text Classification Algorithm," 2009, pp. 21–26.

Umidjon Makhmudov, et. al. "Efficiency Measurement Application of Artificial Intelligence/Language for Predicting/Detecting Crime ." *IOSR Journal of Computer Engineering (IOSR-JCE)*, 23(5), 2021, pp. 24-31.