DOI: http://dx.doi.org/10.21123/bsj.2022.6310

A Crime Data Analysis of Prediction Based on Classification Approaches

Fatima Shaker Hussain [1]* *Abbas Fadhil Aljuboori* [2]

[1] Informatics Institute for Postgraduate Studies, Iraqi Commission for Computers and Informatics, Baghdad, Iraq.
[2] College of Engineering, University of Information Technology and Communications, Baghdad, Iraq.
*Corresponding author: abbas.aljuboori@uoitc.edu.iq
E-mails address: Ms201910521@iips.icci.edu.iq

Received 12/5/2021, Accepted 1/8/2021, Published Online First 20/3/2022, Published 1/10/2022

Abstract:

Crime is considered as an unlawful activity of all kinds and it is punished by law. Crimes have an impact on a society's quality of life and economic development. With a large rise in crime globally, there is a necessity to analyze crime data to bring down the rate of crime. This encourages the police and people to occupy the required measures and more effectively restricting the crimes. The purpose of this research is to develop predictive models that can aid in crime pattern analysis and thus support the Boston department's crime prevention efforts. The geographical location factor has been adopted in our model, and this is due to its being an influential factor in several situations, whether it is traveling to a specific area or living in it to assist people in recognizing between a secured and an unsecured environment. Geo-location, combined with new approaches and techniques, can be extremely useful in crime investigation. The aim is focused on comparative study between three supervised learning algorithms. Where learning used data sets to train and test it to get desired results on them. Various machine learning algorithms on the dataset of Boston city crime are Decision Tree, Naïve Bayes and Logistic Regression classifiers have been used here to predict the type of crime that happens in the area. The outputs of these methods are compared to each other to find the one model best fits this type of data with the best performance. From the results obtained, the Decision Tree demonstrated the highest result compared to Naïve Bayes and Logistic Regression.

Keywords: Crime, Crime Prediction, Decision Tree, Logistic Regression, Naïve Bayes

Introduction:

Crime is an offense against the society that is often pursuing and punishable by the law. Criminals have been known to commit crimes in a variety of locations and any manner. All over the planet, criminal activity has posed a threat to society. Law enforcement authorities generate a huge volume of crime data per year, and it is a major challenge for researchers to find an effective model or technique to manage such complicated data to implement decisions for preventing potential [1].

Geo-location services, combined with new approaches and techniques, can be extremely useful in crime investigation. It promotes a more holistic approach to criminal investigation, mapping, proactive decision-making, and crime prevention [2]. Machine learning provides powerful techniques and algorithms for this action. It is the science of instructing machines to make decisions without the use of humans. Machine learning is being used by law enforcement to better evaluate crime data and try to predict potential future events based on crime pattern recognition. Where predictive analysis is a statistical method for creating models that forecast future events[3, 4] .Typically, these predictive models are evaluated using a set of metrics. speech recognition [5] , industry [6], optical network[7] ,medical [8] are all examples of how machine learning has been used lately

The crime rate in Boston has risen significantly in recent years, especially in cases of property crimes such as burglary, theft, and vehicle jacking. Boston is the largest and most populous city in the United States, with several districts. As a result, the FBI's Uniform Crime Reports (UCR) currently rank it as the most dangerous city in the country [9].The aim is focused on comparative study between three supervised learning algorithms, which are decision tree, logistic regression, and Naive Bayes based on the result of the models to predict the type of crime in the area to be chance for police to take necessary actions and also hope to raise people’s awareness about the security in a certain area.

This contribution helps in obtaining better results in terms of time and effort instead of the manual traditional methods followed in the police stations themselves, and this case has prompted many crimes due to the lack of information available to this security cadre. A speedy implementation for crime problem can be provide or creates a great degree of safety for the citizen.

The questions that the research answers are how the crime type can be predicted from the available data? What are the theoretical concepts of modeling methods that applied in the field of crime prediction?

Related Work

Tahani Almanie et al. [10] focused on finding temporal and spatial criminal hotspots using a set of real-world datasets of crimes include Los Angeles and Denver cities. Certainly, identifying ties between elements of crime will significantly help to predict possible hazardous hotspots at a clear point in the future. The strategy was therefore aimed at concentrating on three main elements of crime data, which are the kinds of crimes, the timing of incidences, and the place of crimes. Using the Apriori algorithm on datasets to classify all possible patterns of crime often regardless of the type of crime committed, then there was using the Naïve Bayesian Classifier and Decision Tree Classifier to construct two separate classification models, to forecast the possible form of crime in a particular place over a particular period in the future. It achieves an accuracy of 51 percent in Denver crime prediction concerning the Naïve Bayesian classifier, while it hits 54 percent for Los Angeles. On the other hand, with 42 percent for Denver and 43 percent for Los Angeles, the Decision Tree Classifier records less prediction accuracy.

Félix Mata et al. [11] emphasis was on designing mobile information systems in urban environments for routing and urban planning. It generates a hybrid solution using semantic analysis and classification algorithms to find safe routes based on data from social media and official police reports. The Bayes algorithm uses data submitted by the mobile application (origin and destination points) to return a path that avoids locations where crime has happened.

Jakaria Rabbi et al.[12] the linear regression model is used to predict Bangladesh's potential crime patterns. The actual crime dataset is compiled from various sources of the Bangladesh Force Police. The model of linear regression is trained on the real dataset. Crime forecasting for, robbery, murder, persecution of women and children, abduction, theft, and other crimes in Bangladesh's various regions is carried out after training the model. This work is beneficial for Bangladesh's police and law enforcement agencies to anticipate, prevent or address potential crime in Bangladesh.

Bhavna Saini et al. [13] developed module that offers an interactive image to navigate hither and thither the crime scene using Google Maps and can aid the analyst evaluates the protection of an area by showing locations that can be the focus for the nearer attacks. The methods of classification used to forecast crimes are K-Nearest Neighbor and Naïve Bayes to supports law enforcement agencies. The data is acquired from the official United Kingdom (U.K) website. The dataset used for the work is accurate, true, and credible, it includes 11 attributes in total.

Atharva Deshmukh et al.[14] provided an application that used high-level machine learning information at wholly different times of day and night to crime average during the zones of the city. With the aid of the latest crime data collection, the application will be able to predict new crime trends in the space. Predicting the crime hotspots primarily aimed at helping people to distinguish between safe and dangerous areas when traveling. Django Rest framework and React Native were used to implement the application.

Methodology

The dataset used in this analysis is a collection of records from the crime incident report database that spans half of 2015 to the first half of 2018 which classifying the sort of incident as well as providing details about when and where it occurred [15]. It’s in comma-separated values (CSV).
The following Tab.1 demonstrates the various machine learning methods applied to the Boston dataset.

Table1. Previous Study that Applied on Dataset

Author	Methods	Prediction Methodology	Remark
Sumaiya Tasnim, et al. [4]	Support Vector Machine, Decision Tree, Naïve Bayes and Binary Logistic Regression Decision tree and random forest	Predict Severity of Crime as High or Low	The geo-location factor is not supported
Jiarui Yi,et al [9]	Decision Tree and Random Forest	Predict Type of Crime	Low Result in both method

Pre-processing Phase

The data pre-processing stage is one of the model's most important steps. From the standpoint of machine learning, this phase is critical, as data pre-processing accounts for 60 to 80 percent of the entire analytical pipeline in a typical machine learning project [16]. Data is pre-processed from missing data by using the mean of all values of that attribute and then converted into a dimensionless shape using the normalization technique in the proposed solution. Using the feature scaling method, the raw data sets were normalized in a scale range of 0 to 1. The normalization of data is defined by the equation below [17]:

$$x' = (x_i - min_x)/(max_x - min_x)$$

Where

x_i is the raw value of the chosen sample in the corresponding data series x.

max_x is the raw data value with the highest value in the respective data series x.

min_x is the smallest raw data value in the given data series x.

Machine Learning

Model building is one of the most crucial tasks in a phase of machine learning methodology. To build the predictive model, Decision Tree(DT), Logistic Regression(LR), and Naive Bayes(NB) are trained and evaluated. For training and testing, the classification model used percentage split methods. In this method, where 80% of the data is used as training and the remaining 20% testing. This research focuses on prediction using:

Logistic Regression is a supervised learning method. It can be used to model and forecast continuous variables. When dealing with a classification problem, logistic regression is used. It generates a binomial result by measuring the likelihood of an occurrence occurring or not occurring based on the values of input variables. The following are some of the benefits of logistic regression: ease of implementation, computational efficiency, training efficiency, and regularization ease. Input features do not need to be scaled[18].

$$f(z) = \frac{1}{1 + e^{-(z)}}$$

$$Z = b_0 + b_1x_1 + b_2x_2 \ldots + b_kx_k$$

Where b_0 is intercept and b_1,$b_2$are a slopes against independent variable x_1-x_k .

Naive Bayes is a simplistic probabilistic classifier that constructs a set of possibilities by counting the frequency and combinations of data values. The Bayes Theorem was used to estimate the likelihood that a given feature set belongs to a specific label [13]. Mathematically it can be stated

$$\mathrm{p}(h/x) = \frac{p\left(x/h\right) \cdot p(h)}{p(x)}$$

$\mathrm{p}(h/x)$ the probability of event **(h)** occurring if **(x)** is true.

$p\left(x/h\right)$ the probability of event (**x)** occurring if (**h)** is true.

$p(h)$ and $p(x)$ are of probabilities of observing of **(x)** and **(h)** independently of each other. **Decision Tree** is a supervised learning method of classification. In DT the dataset is divided into smaller parts, and the classification model creates a tree from it. Each leaf represents an outcome in a tree where each node symbolizes a feature, and each branch denotes a decision(rule). Low-importance features (attributes) are found in the lower levels of trees [19]. At each stage of the procedure, with the support of two functions, the DT selects a feature that best splits the data [20]:

- Gini impurity calculates the likelihood of incorrectly classifying a random sample.

$$\mathrm{GIN(p)} = 1 - \sum_{i=0}^{n} p_i(\mathrm{i})^2$$

- Information gain aids in deciding which feature to split next. The value of information gained can be calculated using entropy, which is defined as:

$$\mathrm{Eetropy(T)} = -\sum_{I=1}^{K} P_i \, log_2(p_i)$$

where P_i symbolize the percentage of each feature present in the child node after a split.

Crime Prediction

Security plays a main role in any society and should always be guaranteed to help people work in efficient and effective ways. Crime prediction process can be done using the old crime records where detection of the crime types are used to identify and analyze the crime that occurred in area, this process is used to provide the information to reduce those crimes [21].

Result and Discussion:

This research relied on the best performance obtained from the classification methods used here to obtain the final model to predict crime type. The methods of percentage split are used here in which the dataset is divided into two group the train and the test. Then, the evaluation parameters are computed to present the overall performance of the system. Evaluation is done by comparing the predicted class labels with the actual class labels are used to estimate a classifier's success. Evaluation measurements that used are (precision, recall and f1-measure). Performances of each classifier model are presented in Tab. 2.

Table 2. Classifier Performance on Dataset.

Methods	Precision	Recall	F1 score
Decision Tree	1.00	1.00	1.00
Naïve Bayes	0.95	0.94	0.95
Logistic Regression	0.93	0.89	0.90

The recall, precision, and F1score were calculated in the evaluation process, with TP denoting true positives, FP denoting false positives, and FN denoting false negatives. These classification metrics are as follows:

Precision=TP/(TP+FP)

Recall= R=TP/(TP+FN)

F1score= 2*Recall*Precision/(Recall + Precision)

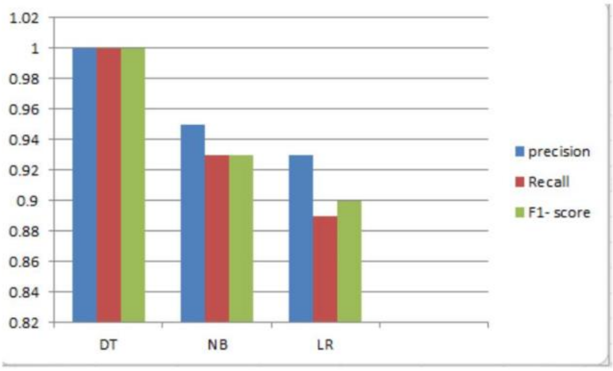

Figure1. Graph chart of performance measures

The graph chart in Fig.1 that could be created from Table- shows that DT algorithm has better results than other, which can be easily noticed as Precision, Recall and F1-measure values using DT algorithm are greater than other.

The results reveal best option after experimenting with various modeling combinations, in which case will get a fairly robust tree that uses longitude and latitude. This is reasonable because the amount and type of crime are strongly linked to the location. This implies that only the crime scene's location can be used to create an ideal model.

Conclusion:

Three machine learning techniques namely DT, NB, and LR were applied to forecast types of crime. The results show that DT outperformed the other machine learning techniques as shown by its metrics values. These basic findings have encouraged us to relate crime with location factors more than ever and in near future, look forward to exploring their connection. For the time being, we sincerely hope that our findings will be a way to prevent crime or reduce the crime rate that occurs with a specific location, which will help police enforcement operations and thus maintain the safety of everyone.

In the future, other factors related to the crime can be adopted, for example determining the gender or identity of the offender.

Authors' declaration:

- Conflicts of Interest: None.
- We hereby confirm that all the Figures and Tables in the manuscript are mine ours. Besides, the Figures and images, which are not mine ours, have been given the permission for re-publication attached with the manuscript.
- Ethical Clearance: The project was approved by the local ethical committee in Iraqi Commission for Computers and Informatics.

Authors' contributions statement:

A. F. A. Contributed in suggest idea of crime prediction (conception) ،select dataset that used in the research and supervised the work.

F. S. Contributed in detect the related study and implement programming task (design). All authors discussed the results and contributed to the final manuscript)

References:

1. Yerpude P, Gudur V. Predictive Modelling of Crime Dataset Using Data Mining. IJDKP. 2017;7(4):43–58.
2. Chutia D, Santra M. Mapping of crime incidences and hotspot analysis through incremental auto correlation -A case study of Shillong city , Meghalaya , India. ISG. 2020;14(April). 61-70 .
3. Hassani H, Huang X, Silva ES, Ghodsi M. A review of data mining applications in crime. Stat Anal Data Min. 2016;9(3): 139–154 .
4. Tasnim S, Sarker P, Hossain A. A Classification Approach to Predict Severity of Crime on Boston City Crime Data. 7th Int Conf Data Sci SDGs. 2019;(December).405- 412 p.
5. Alkhatib B, Eddin MMK. Voice identification using MFCC and vector quantization. Baghdad Sci J. 2020;17(3):1019–28.
6. Ali BA, Gorgees HM, Kathum RI. Modeling human capital impact on the development of the iraqi oil

industry. Baghdad Sci J. 2019;16(4):1080–6.

7. Patwary MKH, Haque MM. A semi-supervised machine learning approach using K-means algorithm to prevent burst header packet flooding attack in optical burst switching network. Baghdad Sci J. 2019;16(3):804–15.
8. Alzubaidi L, Fadhel MA, Al-Shamma O, Zhang J, Santamaría J, Duan Y, et al. Towards a better understanding of transfer learning for medical imaging: A case study. Appl Sci. 2020;10(13).
9. Yin J, Michael IA, Afa IJ. Machine Learning Algorithms for Visualization and Prediction Modeling of Boston Crime Data. 2020;1(February):1–15.
10. Almanie T, Mirza R, Lor E. Crime Prediction Based on Crime Types and Using Spatial and Temporal Criminal Hotspots. Int J Data Min Knowl Manag Process(IJDKP). 2015;5(4):01–19.
11. Mata F, Torres-Ruiz M, Guzman G, Quintero R, Zagal-Flores R, Moreno-Ibarra M, et al. A Mobile Information System Based on Crowd-Sensed and Official Crime Data for Finding Safe Routes: A Case Study of Mexico City. Mob Inf Syst. 2016;2016:11.
12. Awal MA, Rabbi J, Hossain SI, Hashem MMA. Using linear regression to forecast future trends in crime of Bangladesh. 2016 5th Int Conf Informatics, Electron Vision, (ICIEV) 2016. 2016;(June 2020):333–8.
13. Toppireddy HKR, Saini B, Mahajan G. Crime Prediction & Monitoring Framework Based on Spatial Analysis. Procedia Comput Sci [Internet]. 2018;132(Iccids):696–705. Available from: https://doi.org/10.1016/j.procs.2018.05.075
14. Deshmukh A, Banka S, Dcruz SB, Shaikh S, Tripathy AK. Safety App: Crime Prediction Using GIS. 3rd Int Conf Commun Syst, Computing and IT Applications; 2020:120–124.
15. Crimes in Boston | Kaggle [Internet]. Available from: https://www.kaggle.com/ankkur13/boston-crime-data
16. Soni S, Shankar VG, Chaurasia S. Route-the safe: A robust model for safest route prediction using crime and accidental data. Int J Adv Sci Technol(IJAST). 2019;28(16):1415–28.
17. Ridzuan Khairuddin A, Alwee R, Haron H. A Comparative Analysis of Artificial Intelligence Techniques in Forecasting Violent Crime Rate. IOP Conf Ser Mater Sci Eng. 2020;864(1).
18. Ray S. A Quick Review of Machine Learning Algorithms. Proc Int Conf Mach Learn Big Data, Cloud Parallel Comput Trends, Prespectives Prospect Com 2019((Com-IT-Con),). 2019;35–9.
19. Razzaq Abdul Hussein R, Sadik Croock DM, Mahdi Al-Qaraawi DS. Improvement of Criminal Identification by Smart Optimization Method. MATEC Web Conf. 2019;281:05003.
20. Kim S, Joshi P, Kalsi PS, Taheri P. Crime Analysis Through Machine Learning. 2018 IEEE 9th Annu Inf Technol Electron Mob Commun Conf (IEMCON) 2018. 2019;415–20.
21. Sivanagaleela B, Rajesh S. Crime analysis and prediction using fuzzy c-means algorithm. Proc Int Conf Trends Electron Informatics, ICOEI 2019. 2019;(Icoei):595–9.

تحليل بيانات الجريمة للتنبؤ بناءً على مناهج التصنيف

فاطمة شاكر حسين[1] **عباس فاضل الجبوري**[2]

[1]معهد المعلوماتية للدرسات العليا، الهيئه العراقية للحاسبات والمعلوماتية، بغداد، العراق
[2]كلية الهندسة، جامعة تكنلوجيا المعلومات والاتصالات، بغداد، العراق

الخلاصة:

تعتبر الجرائم نشاطا غير مشروع بجميع أنواعه يعاقب عليه القانون ويؤثر على نوعية حياة المجتمع وتطوره الاقتصادي. مع الارتفاع الكبير في معدلات الجريمة على مستوى العالم، هناك ضرورة لتحليل بيانات الجريمة لخفض معدل الجريمة. وهذا يشجع الشرطة والأفراد على اتخاذ الإجراءات المطلوبة والحد بشكل أكثر فعالية من الجرائم. الغرض من هذا البحث هو تطوير نماذج تنبؤية يمكن أن تساعد في تحليل أنماط الجريمة وبالتالي دعم جهود منع الجريمة في قسم بوسطن. تم اعتماد عامل الموقع الجغرافي في نموذجنا ، ويرجع ذلك إلى كونه عاملاً مؤثرًا في عدة مواقف ، سواء كان السفر إلى منطقة معينة أو العيش فيها لمساعدة الناس في التعرف بين بيئة آمنة وغير آمنة. يمكن أن يكون الموقع الجغرافي، جنبًا إلى جنب مع الأساليب والتقنيات الجديدة، مفيدًا للغاية في التحقيق في الجرائم. يتركز الهدف على الدراسة المقارنة بين ثلاث خوارزميات تعلم تحت الإشراف. حيث يستخدم التعلم مجموعات البيانات للتدريب، واختبارها للحصول على النتائج المرجوة عليها. تم استخدام خوارزميات التعلم الآلي المختلفة في مجموعة البيانات الخاصة بجرائم مدينة بوسطن، وهي شجرة القرار ونايف بايز والانحدار اللوجستي المصنفات هنا للتنبؤ بنوع الجريمة التي تحدث في المنطقة. تتم مقارنة مخرجات هذه الطرق مع بعضها البعض للعثور على نموذج واحد يناسب هذا النوع من البيانات بأفضل أداء. من النتائج التي تم الحصول عليها، أظهرت شجرة القرار أعلى نتيجة مقارنة بـ نايف بايز والانحدار اللوجستي.

الكلمات المفتاحية: الجرائم، تنبأ الجرائم، شجرة القرار، الانحدار اللوجستي، نايف بايز.