Journal of Computer Science and Engineering Research (JCSER)
Volume 2, Issue 1, Date: 30 -June - 2025
ISSN: 2978-4328, DOI: 10.64820/AEPJCSER.21.21.26.62025

MapReduce-based Crime Data Analysis in Machine Learning

Dileep Kumar Kadali [1,*]
[1] Shri Vishnu Engineering College for Women, Bhimavaram-534202, India
ORCID ID: 0000-0003-4058-3215
*Corresponding Author

***Abstract*—Nowadays, digital evidence plays a vital role in criminal investigations and arraignments. Digital criminal Investigators can also use this as an opportunity if the vast amount of data is a current trial. Assess constructive and constructive data and advice from the defendant proof behind the crime in terms of issues. Identifying criminal or criminal activity is a big deal because it connects certain data sets. It set an innovative law framework to quickly and accurately solve problems within the law's boundary. In this regard, the machine learning approach Nav Bayes classification for digital criminology data sets is to identify criminals. The Naive Bayesian classification process is used for digital criminology data application. To approximate square estimate for data sets of digital criminology subgroups. Also, support the Hadoop Big Data System Understanding Map with Reduce programming with the Naïve Bayes classifier. The experiment result was a huge accumulated failure in the data quality. Based on these data, the estimation parameter of the statistical model is reached. The least-square estimate estimates the parameters that deal with the statistical model in the experimental result.**

Keywords—Digital Criminology, Big Data, Least Square Estimation, Map Reduce, Naïve Bayes.

I. INTRODUCTION

Nobody stresses that the contemporary domain is going through the Large Information age. Each individual gets hitched in different exercises on the Web utilizing web exchanges, web-based shopping, or other exercises. Because of this multitude of realities, they are incidentally creating immense measures of information consistently, with their optimistic effect on the Web. Alongside the information produced by many administrations, such as internet shopping destinations, a huge amount of information is also developing. The vulnerability of people in general on the Web has, thus, extraordinarily expanded the pace of cybercrime. Then, at that point, because of this reality, the calling of advanced criminological investigators is becoming increasingly difficult without convincing motivation to pool possible proof from the lake of Large Information. Notwithstanding, Enormous Information presents difficulties, yet advanced criminological examiners can involve it as an open door. Inspect developmental and unstructured information and the trouble of recognizing proof from the respondent's lake behind the wrongdoing. Once more, Large Information additionally gives expectations, for example, interconnecting various informational collections to distinguish some lawbreaker or crime. In this paper, we talk about Large Information for Computerized Criminal Specialists. Huge information is so colossal in volume that it can't be estimated regarding gigabytes or terabytes; all things being equal, it is pretty much as extensive as petabytes or zettabytes. Furthermore, the volume is as yet expanding at a quick rate with each second. Enormous information is a blend of organized and unstructured information. Five Versus group enormous information: variety, speed, volume, exactness, and worth. Computerized Criminal science is a part of Applied Science, which manages the recognition, assortment, association, security, and show of proof information that is permitted in an official courtroom. All the more, as of late, Computerized Criminological manages the assortment of proof from the Web. Computerised Criminological Security and Criminological Analysts can help investigate proof assembled from the Web. This sort of criminological investigation additionally manages cloud/haze and other dispersed conditions.

II. LITERATURE SURVEY

The kind of techniques and the reason behind the different wrongdoing expectation studies and applications shifted in the papers we gathered. A significant number of the wrongdoing expectation strategies were created for nonexclusive violations and circumstances, where various models were utilized and tried in wrongdoing forecast to decide the best one compared with the given dataset. For instance, an examination utilizing Machine Learning to foresee nonexclusive wrongdoing. Nonetheless, a few strategies have been produced for specific wrongdoing types or classes, for example, who demonstrated the Visa exchange succession of tasks utilizing a secret Markov model (Well). Different papers have zeroed in on playing out a near examination between the different learning model sorts; for example, two grouping calculations, to be specific, guileless Bayes and backpropagation, were looked at for anticipating the wrongdoing classification in light of a given dataset. Their investigation was performed utilizing 10-overlap cross-approval, and the discoveries show that guileless Bayes performed better than backpropagation for their wrongdoing dataset utilizing Weka. A couple of the papers discussed their goals. For instance, I worked with various learning models

Received: 23-7-2024
Revised: 12-3-2025
Published: 30-6-2025

and calculations and tried them with multiple datasets. They inferred that it is basic to choose a model kind in light of the dataset given, as certain datasets are more viable with various model sorts. Then again, some new studies have investigated wrongdoing forecast strategies. For instance, it reviewed information-digging techniques for wrongdoing expectation in light of various forecast factors, like financial, spatial-transient, segment, and geographic characteristics. Moreover, an efficient writing survey was given in which the writers presented their commitment to identifying and forecasting spatiotemporal wrongdoing areas of interest. They gave the ML and information mining methodologies in the area of interest recognition, notwithstanding their viability, and illustrated the difficulties of building a spatiotemporal wrongdoing expectation model. Independently, I audited the different mechanical planning answers for wrongdoing expectations in brilliant urban areas. The creators considered a few unique portrayals of criminal portrayals and led a relative report. The creators accept that numerous thoughts and strategies have been laid out for wrongdoing expectations yet that field testing is fundamental for the ease of use of those methodologies. The creators center around fake brain organizations and convolutional network procedures for anticipating violations. Wellbeing and security are key significant viewpoints on personal satisfaction in metropolitan regions. The creators introduced an outline that summed up wrongdoing examination in metropolitan information, concentrated on a few kinds of criminal undertaking calculations, and talked about hypotheses on criminal science. Moreover, it gave a short, clear overview of the execution of techniques for wrongdoing expectations and the possibilities of further developing them later on. They utilized SVM, fluffy hypothesis, counterfeit brain organizations, and multivariate time series as Machine Learning techniques. Then again, they introduced a survey of the managed and unaided techniques for wrongdoing discovery, which they examined, as well as gauge violations.

III. PROPOSED WORK

A. *Naïve Bayes Classification on Crime Data:*

Naïve Bayes is one of the artless classifications and includes everything from the digital criminology application to the exception, especially the text classification shown in Fig-1. Record R is given to classify that the general procedure is to give that class C_i, whose probability is P (C_i | R). To the exact value of P (C_i | R), this classification naively assumes that the properties of R are independent of each other. It was once thought that the derivative is used to calculate P (C_i | R) as follows

$$P(C_i|\,R) = \frac{P\,(R \wedge C_i)}{P(R)} \quad (1)$$

$$= \frac{P\,(R \mid C_i)\,P(C_i)}{P(R)} \quad (2)$$

$$= P\,(R \mid C_i)\,P\,(C_i) \quad (3)$$

$$\propto P\,(A_i = x_i|C_i)\,P\,(C_i) \quad (4)$$

Here, the recorded R values apply to the values of A_i with x_i values. The denominator crosses P (R) because it is common to all classes. The final derivation (4) is obtained by gaining independence between properties.

B. *Crime Data Classification Process using Naïve Bayesian:*

The procedures of the Naïve Bayes Classifier enable the assignment of labels to objects. The labels in the classification are predetermined, where we find the structure and assign the labels. Classification problems are supervised learning methods. We start with a digital criminology training

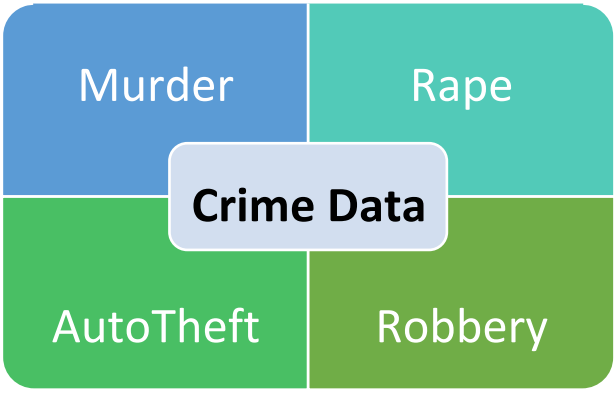


Fig. 1. Crime data attributes

set of pre-classified cases and assign class labels with knowledge of probability.

Naive Bayesian classification is a probability classification based on Bayesian law and naive conditional independence assumptions. In simple terms, a Nav Bayes classification assumes that the presence or absence of a particular feature of a class/group is not related

to the presence or absence of other attributes. Input variables are usually discrete or categorical, but there are variations in algorithms that work with continuous variables. We only consider discrete input variables. However, weight can be regarded as a constant variable, and weight is classified as intervals to convert weight into a categorical variable. The output usually provides a probability score and class membership. Output form Most implementations assign a class lag probability score and a class label corresponding to the highest log probability score. Nav Bayesian classifiers are among the most successful approaches to learning to classify text data. Naïve Bayesian Classifiers are used to identify fraud. This application in digital criminology relies on feature-rich training data, whether or not we can categorize text data using least square estimation.

C. *Regression Analysis for Least Squares Estimation:*

Failure to produce, if any, is assumed to follow a linear model. However, it is essential to distinguish between failure data and non-failure data. It is also difficult to classify whether the failure is a real failure or a transmission delay caused by network issues and other issues related to the hardware. There may be some instances of misclassification

where data can be classified as an error area or vice versa. Therefore, it is essential to classify the data and identify the fault and error-free datasets. This paper uses the statistical method of Least Squares Estimation (LSE) for digital criminology data. It is also known as the least squares estimation used for small-size models. This approach considers the model parameters by fitting the functional relationship of the failure intensity to the mean value of one variable relative to the mean value of the other. Here, the data set coefficients of the equation Y = m X + c can be calculated by solving the general equation. The normal equations are represented by Regression equation of y on x:

$$\sum y = m \sum x + Nc \quad (5)$$

$$\sum xy = m \sum x^2 + c \sum x \quad (6)$$

Regression equation of x on y:

$$\sum x = m \sum y + Nc \quad (7)$$

$$\sum xy = m \sum y^2 + c \sum y \quad (8)$$

At this time, the values of a and b make it easy to compute the value of y for any given value of x or x for any given value of y. The values of a and b are found with the help of above normal equations.

D. Naïve Bayes Classifier with Map Reduce Approach:

The old-fashioned indoctrination languages have a sequential execution of analyzing the largest Digital Criminological data. MapReduce has parallel processing of Digital Criminological data with a set of mapper and reducer classes. Hadoop uses a MapReduce method to analyze the least square estimation of digital criminological data. Hadoop is intended for offline processing through read transactions, and from now on, analysis on analysis on huge Digital Criminological four-subgroup datasets has been made easy.

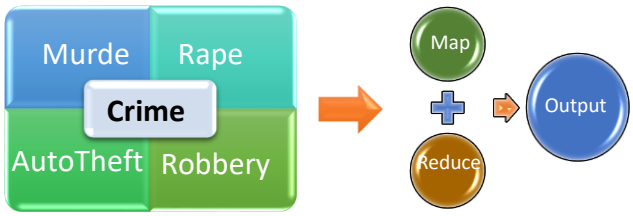


Fig. 2. Crime data classification using the map-reduce process

MapReduce is a Hadoop teaching approach that is utilized for dispersed-based least square assessment computation. In Fig-2, MapReduce parting the work conceded to by the client into little parallelized maps and decreased tasks. The client's piece is to determine a guide capability wherein the Mapper class proceeds as a key/esteem pair and makes a bunch of moderate key/esteem yield matches. The Minimizer class aggregates the middle key/esteem yield matches delivered past and creates a last key/esteem yield match. Hadoop provides a bunch of programming interfaces that are needed to make Mapper and Minimizer classes. The Mapper class accepts the contribution of each as a <Key, Value>pair, and the result is a <Key, Value>pair. The conceivable approach to making a mapper class is broadening a predefined Mapper class with indicated information and result designs. The usefulness of the mapper is clear in the guide capability. The expected approach to making a Minimizer class is expanding a predefined class named Minimizer with determined information and result designs. The usefulness of the minimizer is clear in the decreased capability. The contribution to an application is a bunch of keys and values handled by map capability, which creates a rundown of K_1, K_2, and V_1, V_2 values. The minimizer takes the K_1, K_2, and V_1, V_2 values as information, processes them, and produces a rundown of Keys and Values for everything. The Computerized Criminal Science application is utilized for a thoughtful MapReduce approach.

E. 4. Experimental Result

The Criminal data processed, which includes a huge number of criminal details, is used to control the Digital Criminological investigators in India. The statistical model depends upon the quality of failure data profoundly gathered. The prediction parameter of a statistical model is approximated based on these data, as shown in Table 1. The least-square estimation is used to approximate the parameters.

Table-1: Classified Crime Data Estimated and Predicted Population Values

Crime Data	Murder		Rape		AutoTheft		Robbery	
City Names	Population Value (Units)	Predicted Value (Units)	Population Value (Units)	Predicted Value (Units)	Population Value (Units)	Predicted Value (Units)	Population Value (Units)	Predicted Value (Units)
Bhimavaram	16.5	3.04	24.8	0.73	106	1260.42	494	2626.82
Narasapuram	4.2	2.05	13.3	14.64	122	984.15	954	4558.82
Tadepalligudem	11.6	0.23	24.7	0.78	340	620.82	645	150.42
Eluru	18.1	4.65	34.2	2.47	184	236.02	602	546.02
Vijayawada	6.9	0.54	41.5	11.94	173	331.35	780	510.42
Kakinada	13	0.7	35.7	3.83	477	3634.82	788	608.02
Tuni	2.5	3.5	8.8	24.88	68	2053.35	468	3360.02
Ravulapalem	3.6	2.52	12.7	15.85	42	2706.82	637	205.35
Rajahmundry	16.8	3.31	26.6	0.15	289	138.02	697	1.35
Amalapuram	10.8	0.07	43.2	15.16	255	8.82	765	350.42
Palakollu	9.7	0	51.8	37.39	286	120.42	862	1915.35
Visakhapatnma	10.3	0.02	39.7	8.94	266	33.75	776	464.82
Razole	9.4	0.01	19.4	5.07	522	5170.82	848	1612.02
Gudivada	5	1.5	23	1.75	157	498.82	488	2788.02
Gunturu	5.1	1.44	22.9	1.82	85	1674.82	483	2926.02
Annavaram	12.5	0.5	27.6	0.02	524	5245.35	793	673.35

The Predictable data are the difference between repetitions like Murder, Rape, AutoTheft, and Robbery are well-known in Digital Criminological. Using this approach, of above groups have been identified as Murder, Rape, AutoTheft, and Robbery. The initial selection of threshold value can affect the output in the groups. Hence, the procedure often runs many times with different initial circumstances to get a fair clarification of which group should be.

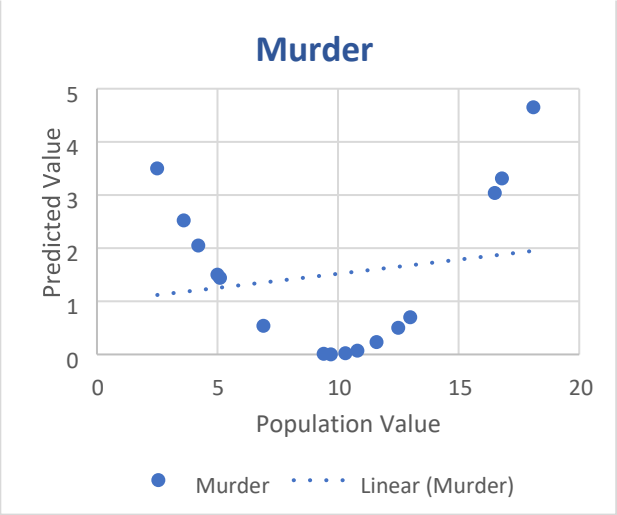


Fig. 3. Murder Data Statistical Evaluation

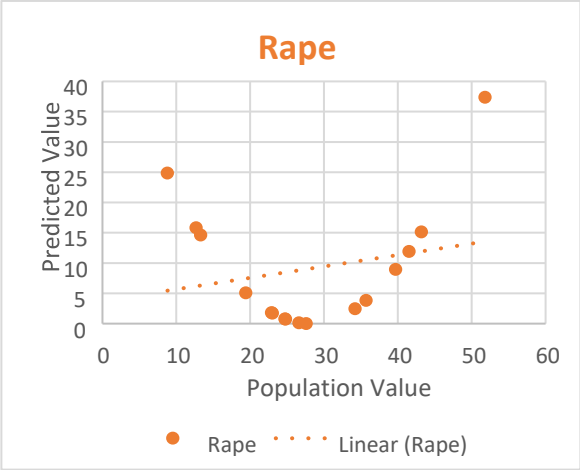


Fig. 4. Rape Data Statistical Evaluation

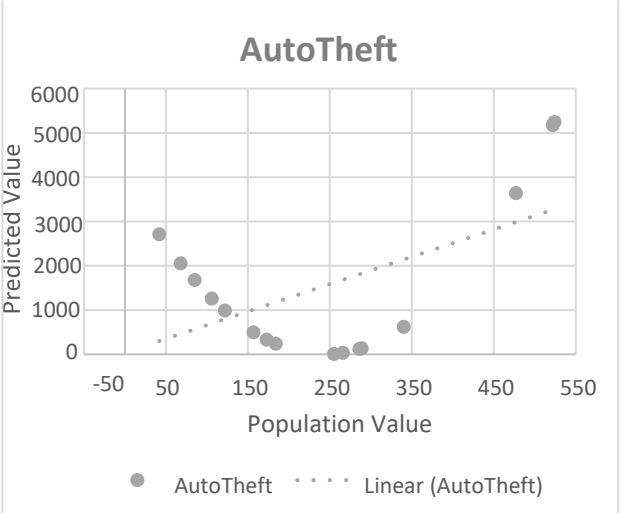


Fig. 5. AutoTheft Data Statistical Evaluation

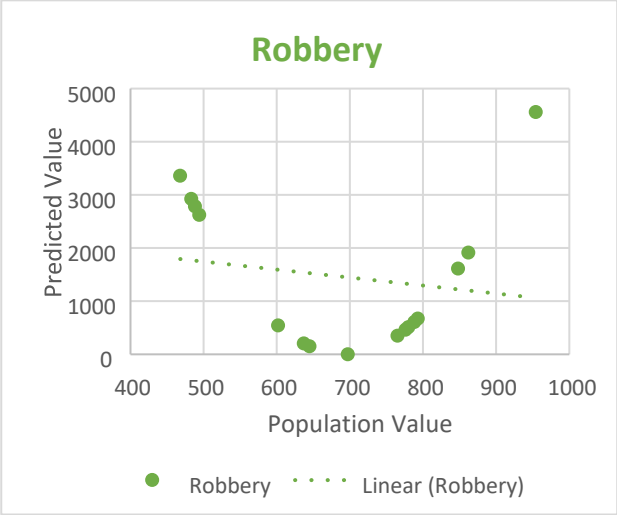
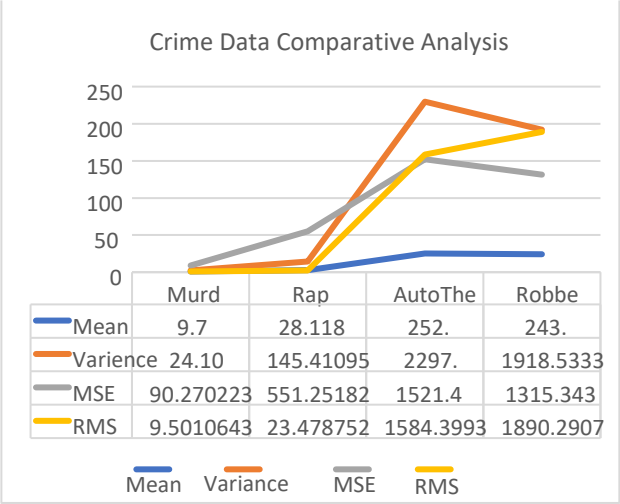


Fig. 6. Robbery Data Statistical Evaluation

For testing this proposed method, we conducted experiments with not the same target, specifically Murder, Rape, AutoTheft, and Robbery. This result is shown in Table 1, Mean, Variance, MSE, and RMSE. It can also clarify the relationship between the Mean, Variance, MSE, and RMSE.



	Murd	Rap	AutoThe	Robbe
Mean	9.7	28.118	252.	243.
Varience	24.10	145.41095	2297.	1918.5333
MSE	90.270223	551.25182	1521.4	1315.343
RMS	9.5010643	23.478752	1584.3993	1890.2907

Fig-7: Crime Date Comparative Analysis

The above graph shows some criminal data related in a curved fashion. Here, we have simple data containing two variables in each group. For each group, the first is the response variable, and the second is the predictor. It appears that the response variable increases and then decreases again. If you plot the data, then we can see the situation more clearly.

IV. CONCLUSION

Large subgroups of data sets are on display in front of digital criminology. It highlights the need for a well-trained probability assessment module to identify, collect, preserve, and securely analyze big data evidence. A class of knowledge to deal with future work uncertainty data is uncertain. To solve the problem of one-class learning and concept summarization practice on uncertain one-class data brooks. Wide trials on uncertain data brooks prove that our future uncertainty one-class learning method works better than others, and our concept summarization method can capture user emerging interests from parts of history.

REFERENCES

[1] Almaw, A., & Kadam, K. (2018). Crime data analysis and prediction using ensemble learning. 2018 Second International Conference on Intelligent Computing and Control Systems (ICICCS).

[2] Tutak, M., & Brodny, J. (2023). A smart city is a safe city: Analysis and evaluation of the state of crime and safety in Polish cities. Smart Cities, 6(6), 3359–3392. https://doi.org/10.3390/smartcities6060149

[3] Vengadeswaran, Binu, D., & Rai, L. (2024). An efficient framework for crime prediction using feature engineering and machine learning. In Advances in Data and Information Sciences (pp. 49–59). Springer Nature Singapore.

[4] Wang, J., Hu, J., Shen, S., Zhuang, J., & Ni, S. (2020). Crime risk analysis through big data algorithm with urban metrics. Physica A, 545(123627), 123627. https://doi.org/10.1016/j.physa.2019.123627

[5] Wu, J., Frias-Martinez, E., & Frias-Martinez, V. (2020). Addressing under-reporting to enhance fairness and accuracy in mobility-based crime prediction. Proceedings of the 28th International Conference on Advances in Geographic Information Systems.

[6] Araujo, A. P. de B., & Costa, A. P. C. S. (2024). A decision support system for predictive crime analytics and a patrol system. In Social Science Research Network. https://doi.org/10.2139/ssrn.4769478

[7] Bandekar, S. R., & Vijayalakshmi, C. (2020). Design and Analysis of Machine Learning Algorithms for the reduction of crime rates in India. Procedia Computer Science, 172,

[8] 122–127. https://doi.org/10.1016/j.procs.2020.05.018

[9] D. K. Kadali, R. Mohan, and M. C. Naik, "Enhancing crime cluster reliability using neutrosophic logic and a Three-Stage model," Journal of Engineering Science and Technology Review, vol. 16, no. 4, pp. 35–40, Jan. 2023, doi: 10.25103/jestr.164.05.

[10] Biswas, A. A., & Basak, S. (2019). Forecasting the trends and patterns of crime in Bangladesh using machine learning model. 2019 2nd International Conference on Intelligent Communication and Computational Techniques (ICCT).

[11] Borges, J., Ziehr, D., Beigl, M., Cacho, N., Martins, A., Araujo, A., Bezerra, L., & Geisler, S. (2018). Time-series features for predictive policing. 2018 IEEE International Smart Cities Conference (ISC2).

[12] Borowik, G., Wawrzyniak, Z. M., & Cichosz, P. (2018). Time series analysis for crime forecasting. 2018 26th International Conference on Systems Engineering (ICSEng).

[13] Calderon, M. H. H., Palad, E. B. B., & Tangkeko, M. S. (2020). Filipino online scam data classification using decision tree algorithms. 2020 International Conference on Data Science and Its Applications (ICoDSA).

[14] D. K. Kadali, D. Raju, and P. V. R. Raju, "Cluster query optimization technique using Blockchain," in Cognitive science and technology, 2023, pp. 631–638. doi: 10.1007/978981-99-2742-5_65.

[15] Cichosz, P. (2020). Urban crime risk prediction using point of interest data. ISPRS International Journal of Geo-Information, 9(7), 459. https://doi.org/10.3390/ijgi9070459 15. do Rêgo, L. G. C., da Silva, T. L. C., Magalhães, R. P., de Macêdo, J. A. F., & Silva, W. C. P. (2020). Exploiting points of interest for predictive policing. Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Advances in Resilient and Intelligent Cities.

[16] Jha, G., Ahuja, L., & Rana, A. (2020). Criminal behaviour analysis and segmentation using K-means clustering. 2020 8th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO).

[17] Kadar, C., Maculan, R., & Feuerriegel, S. (2019). Public decision support for low population density areas: An imbalance-aware hyper-ensemble for spatio-temporal crime prediction. Decision Support Systems, 119, 107–117. https://doi.org/10.1016/j.dss.2019.03.001.

[18] R.N.V.Jagan Mohan, Dileep Kumar Kadali and Y. Vamsidhar, "Similarity based Query Optimization on Map Reduce using Euler Angle Oriented Approach", International Journal of Science and Engineering Research (IJSER), ISSN:2229-5518, Volume-3, Issue-8, Page No.2, August 2012 Edition Statistics, Impact Factor:1.4. (Peer-Reviewed)

[19] K. N. Remani, V. S. Naresh, S. Reddi, and D. K. Kadali, "Crime data optimization using neutrosophic logic based game theory," Concurrency and Computation: Practice and Experience, vol. 34, no. 15, Mar. 2022, doi: 10.1002/cpe.6973.

[20] D. K. Kadali and R. Mohan, "Shortest route analysis for High-Level Slotting using Peer-to-Peer," in Apple Academic Press eBooks, 2022, pp. 113–122. doi: 10.1201/9781003048367-10.

[21] D. K. Kadali, M. C. Naik, and K. N. Remani, "Estimation of data parameters using cluster optimization," in Lecture notes on data engineering and communications technologies, 2022, pp. 331–342. doi: 10.1007/978-981-19-2600-6_23.

[22] Abdul Jalil, M. @. M., Mohd, F., & Mohamad Noor, N. M. (2017). A comparative study to evaluate filtering methods for crime data feature selection. Procedia Computer Science, 116, 113–120. https://doi.org/10.1016/j.procs.2017.10.018

[23] Aldossari, B. S., Alqahtani, F. M., Alshahrani, N. S., Alhammam, M. M., Alzamanan, R. M., Aslam, N., & Irfanullah. (2020). A comparative study of decision tree and naive Bayes machine learning model for crime category prediction in Chicago. Proceedings of 2020 6th International Conference on Computing and Data Engineering.

[24] Alkhaibari, A. A., & Chung, P.-T. (2017). Cluster analysis for reducing city crime rates. 2017 IEEE Long Island Systems, Applications and Technology Conference (LISAT).

[25] Li, Z., Zhang, T., Jing, X., & Wang, Y. (2021). Facial expression-based analysis on emotion correlations, hotspots, and potential occurrence of urban crimes. Alexandria Engineering Journal, 60(1), 1411–1420. https://doi.org/10.1016/j.aej.2020.10.061

[26] D. K. Kadali, R. Mohan, N. Padhy, S. C. Satapathy, N. Salimath, and R. D. Sah, "Machine learning approach for corona virus disease extrapolation: A case study," International Journal of Knowledge-based and Intelligent Engineering Systems, vol. 26, no. 3, pp. 219– 227, Dec. 2022, doi: 10.3233/kes-220015.

[27] Massarelli, C., & Uricchio, V. F. (2024). The contribution of open source software in identifying environmental crimes caused by illicit waste management in urban areas. Urban Science, 8(1), 21. https://doi.org/10.3390/urbansci8010021

[28] Momtaz, M., Padela, J., Leslie, R., & Quader, F. (2024). Developing predictive models for smart policing based on Baltimore's crime and product price correlation. In Intelligent Sustainable Systems (pp. 551–566). Springer Nature Singapore.

[29] Rumiantsev, T., van der Rijst, R., & Admiraal, W. (2023). A systematic literature review of collaborative learning in conservatoire education. Social Sciences & Humanities Open, 8(1), 100683. https://doi.org/10.1016/j.ssaho.2023.100683

[30] Saravanan, P., Selvaprabu, J., Arun Raj, L., Abdul Azeez Khan, A., & Javubar Sathick, K. (2021). Survey on crime analysis and prediction using data mining and machine learning techniques. In Lecture Notes in Electrical Engineering (pp. 435–448). Springer Singapore.

[31] Shukla, S., Jain, P. K., Babu, C. R., & Pamula, R. (2020). A multivariate regression model for identifying, analyzing and predicting crimes. Wireless Personal Communications, 113(4), 2447–2461. https://doi.org/10.1007/s11277-020-07335-w

[32] Tamilarasi, P., & Rani, R. U. (2020). Diagnosis of Crime Rate against Women using K-fold Cross Validation through Machine Learning. 2020 Fourth International Conference on Computing Methodologies and Communication (ICCMC).