Proceedings of the Fourth International Conference on Electronics and Sustainable Communication Systems (ICESC-2023)
IEEE Xplore Part Number: CFP23V66-ART; ISBN: 979-8-3503-0009-3

2023 4th International Conference on Electronics and Sustainable Communication Systems (ICESC) | 979-8-3503-0009-3/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICESC57686.2023.10193563

# Crime Hotspot Detection using Optimized K-means Clustering and Machine Learning Techniques

Janakiramaiah Bonam
*Department of CSE*
*PVP Siddhartha Institute of Technology*
Kanuru, Vijayawada, AP, India
bjanakiramaiah@gmail.com

Lakshmi Ramani Burra
*Department of CSE*
*Koneru Lakshmaiah Education Foundation,*
Vaddeswaram AP, India
ramanimythili@gmail.com

G Sai Venkata Naga Sudarshan Susheel
*Department of CSE*
*PVP Siddhartha Institute of Technology*
Kanuru, Vijayawada, AP, India
saisusheel03@gmail.com

Kadiyala Narendra
*Department of CSE*
*PVP Siddhartha Institute of Technology*
Kanuru, Vijayawada, AP, India
kadiyala.narendra7@gmail.com

Macharla Sandeep
*Department of CSE*
*PVP Siddhartha Institute of Technology*
Kanuru, Vijayawada, AP, India
sandeepmacharla01@gmail.com

Gali Nagamani
*Department of CSE*
*PVP Siddhartha Institute of Technology*
Kanuru, Vijayawada, AP, India
galinagamani2000@gmail.com

Corresponding Author: ramanimythili@gmail.com

***Abstract*— Over the world, crimes are becoming increasingly complicated and technologically advanced. It has become essential to prioritize actions targeting crime categories within every area. Nowadays, people are more concerned and stressed by the unprecedented increase in city crimes and violations. Large numbers of crimes are perpetrated frequently each day. Crime data analysts can aid law enforcement officials in discovering criminals more quickly. Crime analysis is an organized method of categorizing locations into crime hotspots and non-hotspot zones. The dataset utilized is the Kaggle-obtained UCI Crime and Communities Dataset. The algorithm can categorize the areas into hotspots and non-hotspot areas by examining crime records and associated variables. The suggested system uses data mining and machine learning techniques, such as optimized K-means clustering and the elbow approach, to explore datasets and analyze crimes committed. Three classification algorithms - Decision Tree Algorithm, Support Vector Machine, and Random Forest Algorithm are included in the model. They are trained on the dataset used to determine if an area is a hotspot based on the criteria that influence the occurrence of crimes in a region.**

***Keywords*— *K-Means Clustering, Elbow Method, Support Vector Machine, Random Forest, Decision Tree, UCI Crime dataset***

## I. INTRODUCTION

Crime is a pervasive problem worldwide, and India is no exception. In recent years, there has been a sharp increase in crime in India, making it more important than ever to develop effective crime analysis and prediction techniques. One such technique is clustering, which involves grouping similar data points based on certain characteristics. One popular clustering method is K-means clustering, which can help identify crime hotspots and aid in forecasting future crime. However, determining the optimal number of optimal clusters (known as the "K" value) is crucial for effective clustering. It can be a challenging task, particularly with large datasets. In this context, the elbow method is often used to identify the K value. This method can be used to select the ideal K value and choose better centroids to improve the quality of the clustering results.

This study uses K-means clustering to identify crime hotspots in India, by specifically focusing on three key variables: violent crime per capita, unemployment rate, and proportion of residents with only a high school diploma. We utilize a large dataset with 140 columns of various characteristics that can affect crime rates. After clustering the data into hotspots and non-hotspot areas, we construct a model using three classification algorithms: Decision Tree (DT), Support Vector Machine (SVM), and Random Forest (RF) [1-4]. We select features such as the percentage of illiterate people, per-capita income of various races in the area, percentage of graduates, employed individuals, and several others to train and test the model. We evaluate the performance of the three algorithms to determine which is best suited for the proposed model. Implementing the optimal K-means clustering in Python is done through the Anaconda software. The overall objective of this work is to develop a more effective crime prediction [5] model that can help identify crime hotspots and guide law enforcement agencies in their efforts to prevent and combat crime in India.

## II. LITERATURE SURVEY

Numerous research publications have proposed several clustering techniques to deal with the prediction of crime occurrences.

Shiju Sathyadevan et al. present a methodology for analyzing and predicting crime using data mining. The authors collect crime data, select features, and build a classification model using a decision tree algorithm. They apply this approach to a dataset and demonstrate its effectiveness in classifying different types of crimes and predicting future incidents. Overall, the study highlights the potential of data mining in crime analysis and can help law enforcement agencies prevent and reduce crime [6]. Jesia Quader Yuki et al. propose using machine learning algorithms to predict crime based on location data and time. The authors applied this method to a dataset of crime incidents and demonstrated its effectiveness in predicting different types of crimes and identifying high-risk areas. Overall, the study highlights the probability of machine learning in crime prediction and can help law enforcement agencies prevent and reduce crime [7].

Peng Chen et al. present a methodology for crime pattern detection using a simple apriori algorithm. The method is applied to a dataset of crime incidents and demonstrates its effectiveness in identifying frequent patterns in the modus operandi of criminals. The study highlights the use of a simple apriori algorithm for crime pattern detection and can provide valuable insights for law enforcement agencies to prevent and solve crimes [8]. Jyoti Agarwal et al. recommended a methodology for crime analysis using the K-means clustering algorithm and provided a practical approach for implementing this technique in a real-world scenario. The authors demonstrate that their method can effectively identify crime patterns and clusters to help law enforcement agencies allocate resources and design effective strategies to prevent and reduce crime [9].

Sheng Li et al. suggested a method for predicting crime based on spatiotemporal data using a Long Short-Term Memory (LSTM) neural network. The authors collected crime data and corresponding spatiotemporal information in Beijing and constructed a dataset for the experiments. The LSTM network was trained using the dataset, and the results showed that the proposed method could accurately predict crime incidents based on spatiotemporal data. The study demonstrates the potential of using machine learning methods such as LSTM for crime prediction and highlights the importance of spatiotemporal data in crime analysis. It can be useful for law enforcement agencies in their efforts to prevent and reduce crime [10].

Kai Zhang et al. recommended a hybrid machine learning model for predicting crime using a mixture of a neural network and a decision tree algorithm. The authors collected crime data and relevant features in a city in China and trained and tested the proposed model using the dataset. The results show that the hybrid model outperforms individual machine learning methods, such as a neural network or decision tree algorithm, predicting crime incidents. The study demonstrates the potential of using a hybrid model for crime prediction. It provides insights into the importance of feature selection and model optimization for improving the accuracy of crime prediction [11]. Yan Jiang et al. recommended a deep learning-based method for crime prediction in smart cities. The proposed method, called Deep Crime, was evaluated on a large-scale dataset, and the results showed that it outperformed several baseline methods in predicting crime incidents. The authors collected crime data to extract spatiotemporal features and predict crime incidents. The features were extracted using CNN and LSTM networks. The study demonstrates the potential of using deep learning methods such as CNN and LSTM for crime prediction in smart cities. It highlights the importance of spatiotemporal features in crime analysis [12].
Md. Sabir Hossain et al. presented a machine learning-based method for predicting crime hotspots in Dhaka City, Bangladesh. The authors collected crime data and relevant features like population density and road network. They used several machine learning algorithms, including k-nearest neighbour (KNN), DT, RF, and SVM, to predict crime hotspots. The proposed method was evaluated using various performance metrics, and the results showed that the SVM algorithm outperformed other methods in predicting crime hotspots. The study demonstrates the potential of using machine learning methods [18-20] for crime analysis in developing countries, where data availability and quality can be limited. It provides insights into the importance of feature selection and model optimization for improving the accuracy of crime prediction [13].

## III. PROPOSED MODEL

The proposed methodology aims to develop a more accurate and efficient model for crime prediction using K-Means Clustering and classification algorithms. One major issue with existing models is their inability to produce optimal results. Our model is designed to work on unbalanced datasets and perform well across all datasets while requiring less computation time. Used the "Crime and Communities" dataset as the input data source to implement this model. Before analysis, we prepare data to remove null values in the dataset by replacing them with the respective row's mean values. After data preparation, we select several features from the dataset and employ the elbow approach to fix the ideal number of clusters for the K-Means Clustering algorithm.

Once the optimal number of clusters is identified, we group the data into hotspots and non-hotspot locations, as shown in Fig 1. Next, we train the model using the classification algorithms: DT, SVM, and RF to predict the hotspot regions. We can determine the best algorithm for our model using these algorithms. The proposed methodology is expected to provide a more accurate and efficient model for crime prediction, improving public safety efforts in the identified hotspot regions.

### *A. Algorithms*

- **K-Means Clustering Algorithm:**

The clustering process used by the unsupervised learning algorithm [14] divides the unlabeled input into multiple groups. Here, K determines the predefined clusters that should be produced as part of the process; for example, if K=2, there are two clusters. If K=3, then three clusters, etc.

This algorithm is an iterative approach that divides an unlabelled dataset into k separate clusters, each belonging to only one group that shares characteristics with the others. Each cluster is assigned a centroid as the algorithm is centroid-based. The main goal of this algorithm is to reduce the total distance between each data point in each cluster and its corresponding clusters.

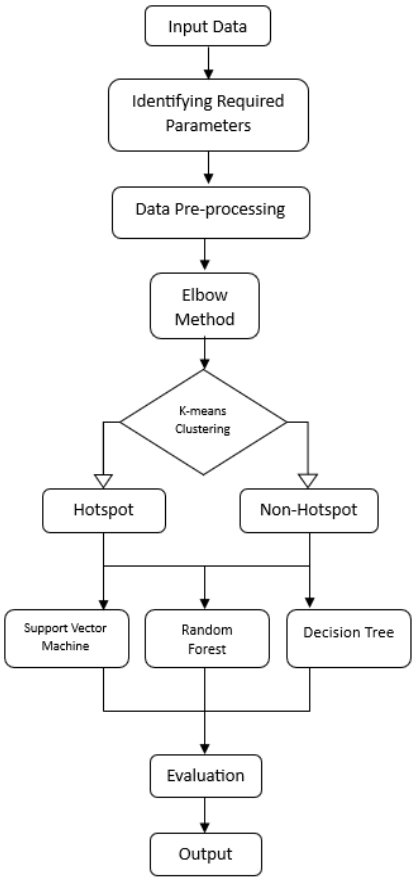


FIG. 1. FLOW PROCESS DIAGRAM FOR THE PROPOSED MODEL

- **Support Vector Machine Algorithm:**

Classification and regression issues are addressed using the Support Vector Machine algorithm [15], one of the most widely used supervised learning techniques. However, Machine Learning Classification issues account for the majority of its usage. The technique aims to specify the decision boundary that can categorize the n-dimensional space, enabling us to swiftly categorize new data points in the future. A hyperplane is the name of this optimal decision boundary. The hyperplane's extreme vectors and points are selected using SVM. The SVM method is built on support vectors that describe these extreme circumstances.

- **Random Forest Classifier Algorithm:**

The Random Forest algorithm [16], a supervised learning approach component, is the most widely used machine learning algorithm. It can be used to solve classification and regression-related machine-learning issues. The approach is built on ensemble learning, combining various classifiers to address challenging issues and enhance model performance. The Random Forest classifier, as its name suggests, employs numerous decision trees on variously divided portions of the input data set and calculates their average to improve the data set's prediction accuracy. The random forest algorithm takes predictions from each decision tree instead of just one and forecasts the result based on votes from most predictions.

- **Decision Tree Classifier Algorithm:**

Using the supervised learning method known as a decision tree, classification and regression problems can be solved [17]. A tree-structured classifier was underlying the algorithm. The classifier uses the features of the given dataset to run the test or make the decision for the new data given. Each internal node of the tree produced for the provided data represents the dataset's features given to the algorithm. Branches are utilized for decision-making, and each tree leaf node is used for classification outcomes. A decision tree has two different sorts of nodes. Decision Node and Leaf Node are what they are. The decision nodes of the tree have numerous branches, as opposed to the leaf nodes, which are the results of decisions and have no more branches.

## IV. IMPLEMENTATION

The implementation of our proposed methodology begins by obtaining the UCI Crime and Communities dataset from Kaggle. In the feature selection stage, we choose the relevant features for our task. However, since the dataset contains null values, a data pre-processing step is applied to handle them. The null values are replaced with the mean values of the respective rows. After pre-processing, the data is prepared for the K-means clustering algorithm. The elbow approach is then utilized to determine the optimal number of clusters. The elbow technique applies K-means clustering to the data and divides it into the specified number of clusters after identifying the optimal clusters.

The data is labelled and separated following clustering into training and test sets. The RF, SVM and DT algorithms are trained using the training data. The performance of the trained algorithms is measured using accuracy metrics on the test data. The results of the algorithms are then compared, and the algorithm with the highest accuracy is identified. The output is presented as a model that accurately predicts crime hotspots while requiring less computation time, even on unbalanced datasets.

**Dataset Collection:**

The UCI Crime and Communities dataset, which can be publicly accessed through Kaggle, collects crime data in 2215 cities, and the dataset description is shown in Table 1. Each city is described by 147 attributes, which include information on the name of the community, state, nation

code, and community code, that provide specific details about each community. The dataset also includes 140 parameters that represent the number of murders, rapes, and assaults in each community, along with other factors such as median income, number of people living below the poverty line, number of arsons, and violent and non-violent crimes per population. These parameters can be utilized to construct a model for training and testing algorithms.

TABLE. 1. DATASET DESCRIPTION

| Dataset Name | UCI Crime and Communities dataset |
|---|---|
| Total Attributes | 147 |
| Total Tuples | 2215 |

**Data Pre-processing:**

Data pre-processing is a crucial stage in this project that aims to address inconsistent data and ensure more accurate and reliable results. This project's UCI Crime and Communities dataset contains missing values in certain attributes. As including these attributes with missing values would result in over 2,000 missing values, mean imputation is performed to replace the missing values with the mean values for the respective attributes. It is necessary as some attributes cannot have zero values, and omitting these attributes would result in incomplete and unreliable data. Therefore, mean imputation is performed to ensure all attributes are included in the implementation process.

**Elbow Method:**

We have used the elbow approach to calculate the ideal number of clusters for the K-means clustering algorithm. The within-cluster sum of squares (WCSS) metric is used to evaluate the effectiveness of clustering, where WCSS is the sum of squared distances between each point and the centroid of its assigned cluster. A larger number of clusters can result in a significant change in the WCSS value, making it essential to identify the ideal number of clusters for the model. By plotting the WCSS for various clusters, the optimal number of clusters is identified by looking for the "elbow point", where the WCSS begins to decrease at a slower rate.

**Clustering:**

We utilized K-means clustering on the dataset to divide each location into a hotspot or non-hotspot class. Each of our entries was given a class label (0 or 1) following the clustering process. The dataset is expanded to include the class labels produced by clustering. Before K-means clustering, it was found that the attributes were heavily correlated.

**Splitting the dataset:**

In this step, the dataset is split into training and testing datasets. The classification techniques employed in the model will be trained using the training dataset. The trained classification model will be tested on the test dataset. 80% of the dataset will be used as training data and 20% as testing data to prevent underfitting and overfitting. The complete dataset measures 2215 by 147 pixels, the training dataset contains 1772 photos, and the testing dataset has 443 photos.

**Model Building:**

A model is developed for forecasting hotspot areas at this stage using machine-learning classifier techniques like SVM, RF and DT. The Support Vector Machine employs a linear kernel and a zero-random-state technique. The Decision Tree and Random Forest algorithms use the Gini criterion, which effectively measures the distribution of values in the dataset. The Gini Index aims to minimize impurities from the root nodes at the top of the decision tree to the leaf nodes. This model is essential for accurately predicting and identifying hotspot areas, providing valuable insights for decision-makers in various applications, such as disaster management, disease control, and urban planning.

**Evaluation:**

The prediction model's performance is measured using classification accuracy criteria in the evaluation phase. This measure provides insight into how well the model performs and allows for comparisons between different models. The model's accuracy is calculated as the ratio of the correctly predicted outcomes to the total number of predictions.

## V. RESULTS

The elbow method graph shows that two clusters are the ideal number to fit the model, as shown in Fig 2. It shows how the With-in Cluster Sum of Squares (WCSS) varies as the number of clusters changes. As seen in the graph, the WCSS abruptly changes when the number of clusters is set to 2. Hence, K = 2 will be considered while clustering the dataset to get the class label. After determining the class label, the data is divided into training and testing data. The models are trained first and then tested using the data to find the optimal algorithm for the model.

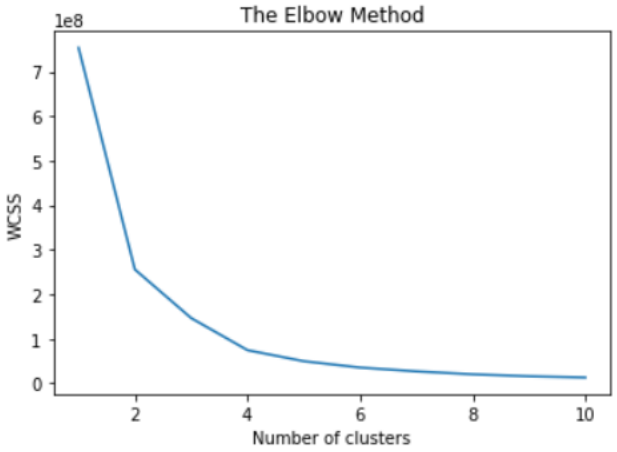


FIG. 2. ELBOW METHOD GRAPH

After testing the algorithms, the algorithms' accuracies in the data prediction are depicted in Table 2.

TABLE. 2. ACCURACIES OF ALGORITHMS

| ALGORITHM | ACCURACY |
|---|---|
| Support Vector Machine | 84.06 |
| Random Forest Classifier | 88.08 |
| Decision Tree Classifier | 85.55 |

## VI. CONCLUSION

The proposed system that utilizes data mining and machine learning techniques has shown promising results in identifying crime hotspots and non-hotspot areas. By analyzing the Kaggle-obtained UCI Crime and Communities Dataset, the model can accurately classify locations based on the criteria that influence the occurrence of crimes in a region. The classification algorithms, including Support Vector Machine, Random Forest Algorithm, and Decision Tree Algorithm, provide a comprehensive and reliable approach to identifying crime hotspots, with the Random Forest Classifier achieving the highest accuracy at 88.08%.

Identifying crime hotspots can aid law enforcement officials in discovering criminals more quickly and prioritizing their actions to reduce the occurrence of crimes in specific areas. The system can serve as a valuable tool for crime data analysts and law enforcement officials, providing insights and knowledge necessary to combat crimes effectively in an increasingly complex and technologically advanced world.

## VII. FUTURE SCOPE

In the suggested model, we have considered the percentage of low-skilled workers, unemployment, and violent crimes as population characteristics to determine whether or not an area is a hotspot. To identify the locations and increase the model's effectiveness, we will consider more data in the future, such as murders, assaults, average income, and other factors. We can consider logistic regression and other algorithms in addition to accuracy as the run time for the random forest method is large to discover the most effective technique that works with our model for hotspot prediction using a web application.

## REFERENCES:

[1] Jangra, Mrinalini, and Shaveta Kalsi. "Naïve Bayes approach for the crime prediction in data mining." *International Journal of Computer Applications* 178.4 (2019): 33-37.
[2] Ivan, Niyonzima, et al. "Crime Prediction Using Decision Tree (J48) Classification Algorithm." (2017).
[3] Kim, Suhong, et al. "Crime analysis through machine learning." *2018 IEEE 9th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)*. IEEE, 2018.
[4] Kalyani, G., MVP Chandra Sekhara Rao, and B. Janakiramaiah. "Privacy-preserving classification rule mining for balancing data utility and knowledge privacy using adapted binary firefly algorithm." *Arabian Journal for Science and Engineering* 43 (2018): 3903-3925.
[5] Almanie, Tahani, Rsha Mirza, and Elizabeth Lor. "Crime prediction based on crime types and using spatial and temporal criminal hotspots." *arXiv preprint arXiv:1508.02050* (2015).
[6] Sathyadevan, Shiju, M. S. Devan, and S. Surya Gangadharan. "Crime analysis and prediction using data mining." *2014 First international conference on networks & soft computing (ICNSC2014)*. IEEE, 2014.
[7] Yuki, Jesia Quader, et al. "Predicting crime using time and location data." *Proceedings of the 7th International Conference on Computer and Communications Management*. 2019.
[8] Chen, Peng, and Justin Kurland. "Time, place, and modus operandi: a simple apriori algorithm experiment for crime pattern detection." *2018 9th International Conference on Information, Intelligence, Systems and Applications (IISA)*. IEEE, 2018.
[9] Agarwal, Jyoti, Renuka Nagpal, and Rajni Sehgal. "Crime analysis using k-means clustering." *International Journal of Computer Applications* 83.4 (2013).
[10] Sheng Li, Lihua Yang, Peng Zhang, and Lei Wang. "Crime prediction based on spatiotemporal data using LSTM neural network." *Neurocomputing* 441 (2021): 251-263.
Kai Zhang et al. "A hybrid model based on neural network and decision tree for crime prediction." *Journal of Ambient Intelligence and Humanized Computing*, vol. 12, no. 2, pp. 1553-1562, 2021.
[11] Kai Zhang et al. "A hybrid model based on neural network and decision tree for crime prediction." *Journal of Ambient Intelligence and Humanized Computing*, vol. 12, no. 2, pp. 1553-1562, 2021.
[12] Jiang, Y., Xu, C., Chen, Y., Chen, J., & Chen, H. (2021). DeepCrime: Crime Prediction Based on Deep Learning in Smart City. *IEEE Transactions on Industrial Informatics*, 17(4), 2485-2494.
[13] Hossain, M. S., Rahman, M. M., & Hossain, M. A. (2021). Predicting the crime hotspots using machine learning techniques: A case study in Dhaka city, Bangladesh. *PloS one*, 16(4), e0249774.
[14] Sinaga, Kristina P., and Miin-Shen Yang. "Unsupervised K-means clustering algorithm." *IEEE access* 8 (2020): 80716-80727.
[15] Suthaharan, Shan, and Shan Suthaharan. "Support vector machine." Machine learning models and algorithms for big data classification: thinking with examples for effective learning (2016): 207-235.
[16] Raza, Dewan Mamun, and Debasish Bhattacharjee Victor. "Data mining and Region Prediction Based on Crime Using Random Forest." 2021 *International Conference on Artificial Intelligence and Smart Systems (ICAIS). IEEE*, 2021.
[17] Hajela, Gaurav, Meenu Chawla, and Akhtar Rasool. "A clustering based hotspot identification approach for crime prediction." *Procedia Computer Science* 167 (2020): 1462-1470.

[18] Tumuluru, Praveen, et al. "DPMLT: Diabetes Prediction Using Machine Learning Techniques." *2022 International Conference on Electronics and Renewable Systems (ICEARS)*. IEEE, 2022.

[19] Lakshmi Ramani, B., et al. "Comparative Analysis of Customer Loan Approval Prediction using Machine Learning Algorithms." *2022 Second International Conference on Artificial Intelligence and Smart Energy (ICAIS)*. IEEE, 2022.

[20] Tumuluru, Praveen, et al. "APMWMM: Approach to Probe Malware on Windows Machine using Machine Learning." *2022 International Conference on Applied Artificial Intelligence and Computing (ICAAIC)*. IEEE, 2022.