Comparative Study of Approaches for Identification of Crime Prone Areas

Mr. Ashish Bishnoi [1], Aashvi Jain [2], Divyanshu Jain [3]
Teerthanker Mahaveer University, Moradabad
[1] ashish.computers@tmu.ac.in

Abstract— **To find or locate the areas famous for criminal activities in a specific city and the using the forces available affectively to eliminate it are both essential exploratory techniques. Different algorithms and techniques had been designed and developed to identify crime hotspots, but few research works have rigorously compared how well they function, particularly when it comes to identifying complex-shaped crime hotspots. Maintaining peace and order in any society requires knowing where crime is likely to occur. Using k-means clustering and random forest classification, this research study describes an approach for detecting crime-prone locations. The model predicts whether new areas are likely to experience crime depending on new input values for latitude and longitude. The approach involves clustering crime data using the k-means algorithm, then training a random forest classifier on the clustered data to forecast crime incidence in various locations. We train the proposed model using the Support Vector Machine (SVM) and Decision Tree algorithms and compare it with other models. We tested this methodology on an authenticated dataset of Lucknow's crime incidence that we collected from 112 hotline numbers. The results of the proposed model show that the methodology used is highly accurate in detecting crime-prone locations.**

***Keywords*: Crime Prone, hotspots.**

I. INTRODUCTION

Crime is a serious societal issue that exists in all communities. It is crucial to pinpoint the places where crime is most likely to occur since with this knowledge, crime can be reduced through practical techniques. The manual study of crime data and expert judgment are traditional techniques for locating crime-prone locations. These techniques take a lot of time, though, and they might not be reliable. Machine learning has made it possible to create approaches for identifying crime-prone locations or areas more effectively and precisely. In environmental criminology to identify the crime prone areas in a particular location had draw the focus of researchers in this field. Hotspots for crime are even seen to be the most fundamental and significant measure of how unevenly committed crimes are. The identification of crime hotspots is crucial for both crime prevention and prediction. The study done for this research paper, aids in locating high-risk areas that demand additional care or perhaps more forces than other areas to maintain peace and security. Such studies support in finding or locating the crime hotspots in a particular city and it could aid in illuminating the factors that in turn affects the occurrence of crime. This may inform the allocation of resources and the creation of policies. Crime hotspots in a particular city are not well identified, and the term "hotspot" has many different connotations. To identify the crime hotspots in the particular city, two ways or approaches are used. Firstly, regions where crimes are concentrated are referred to as crime hotspots. Techniques of Information Technology like hotspot-mapping (e.g., kernel density estimation) & clustering or grouping (e.g., K-means clustering, hierarchical clustering) are used to find criminal hotspots. Second, crime hotspots require not just criminal grouping but also a significant number of criminals. In this sense, hotspots are the clusters or areas in the city where occurrence of criminal cases whose importance may be assessed using geographic data. In this research, we conduct a comparison between various methods for locating crime hotspots based on various algorithms like Decision Tree Classification, Random Forest (RF), and Support Vector Machines (SVM) that have been trained using K-means

clustered data. On a verified dataset of Lucknow's crime incidence that we compiled from 112 hotline numbers, we tested this methodology. The results demonstrate how effective the suggested methodology is at identifying crime-prone areas.

II. Literature Review

Areas or locations in the particular city where possibility of occurrence of crime is more or previously criminal activities noticed or recorded are called crime-prone areas. These places are classified as having a high threshold for the crime.

To identify the areas more prone to criminal activities in the particular city recently drew more attentions of the researchers with development of new algorithms or tools in machine learning. A lot of researches or studies performed on this topic or subject by various researchers in different part of the world using different tools and algorithms based on machine learning technologies. The following important studies or researches offers basic information and tools designed on this subject:

S. Indumathi and K. Priya's "Crime Prediction Using Data Mining". The decision trees, neural networks, and clustering methods utilised in data mining for criminal activities identification thoroughly reviewed in this research. The difficulties and restrictions of applying machine learning to criminal activities identification are also covered by the writers.

N. R. Charan and S. Jana's "Crime Hotspot Prediction Using Machine Learning: A Survey" In this research paper, Support Vector Machines (SVM's), Random Forests, and Artificial Neural Networks (ANN) are covered in detail as machine learning techniques for crime hotspot prediction. The writers also go over the many traits, characteristics, and data-sets from various sources that could be used to criminal activity identification.

S. P. Patil, K. R. Chavan, and A. R. Kulkarni's "Predicting Crime Using Machine Learning Techniques: A Review" In their paper, they identified different machine learning methods for crime prediction, such as Decision Trees, k-Nearest Neighbours, and Artificial Neural Networks (ANN), are reviewed. The authors also go into the numerous causes of crime, such as demographics, the environment, and the time of day.

M. P. Yadav and S. K. Singh's article "Crime Hotspot Prediction: A Review of Data Analytics Approaches" This study gives a thorough overview of several data analytics methods, such as spatial clustering, spatiotemporal analysis, and machine learning algorithms, utilised for crime hotspot prediction. The difficulties and restrictions of applying data analytics to crime prediction are also covered by the writers.

These studies offer a thorough study about the subject of locating crime prone locations in a specific city and implementation of different algorithms of machine learning. They go over the many methods, strategies, and difficulties associated with criminal activities finding or prediction in the specific part of the city and offer insightful information for next studies.

III. Proposed Idea

In this study, we provide a strategy for locating crime-prone locations by applying random forest classification and k-means clustering. Using the k-means algorithm to cluster crime data, the suggested methodology entails training a random forest classifier on the clustered data to forecast crime incidence in various locations. The proposed model is trained using the Lucknow District's 112 hotline crime data. Using the k-means algorithm, criminal activity data is clustered, and a random forest classifier is trained on the clustered data to predict where crimes will occur. We also trained proposed model using the SVM and Decision Tree (DT) techniques for analysis with other approaches. On a verified dataset of Lucknow's crime incidence that we compiled from 112 hotline numbers; we tested this methodology. On analysis with other approaches, the suggested methodology is highly accurate in detecting crime-prone locations.

IV. Methodology

The steps in the suggested technique are as follows:

1. **Data Collection: -** After analyzing the issue, we must collect information from the 112 Helpline and other sources that have the Lucknow district's crime statistics. The procedure of gathering the records of information or data required for training the proposed model is known as data collection. Prior to gathering data, we identify the type of issue we are trying to resolve, look into the sources from where data required to train the model are readily available and accessible to the public, and lastly look into the format of the data. Following all these presumptions, we next compile the Lucknow District's crime data in CSV format.
2. **Data Preprocessing:** Preprocessing is the procedure of translating unprocessed data into a format which could be used for learning of the machine. Using organized and clean data helps in achieving more precise results with models or techniques of machine learning. These techniques require data formatting, cleaning etc.

2.1 Data Presentation- When information is received from many sources by different people, it is required to follow a specific format so that it become more useful not only for the current application but also for future applications. An expert evaluates the consistency of variable recording for each characteristic. Variables include product and service names, prices, date formats, and addresses. Data consistency should apply to different aspects of problems that could be represented in terms of numerical values as well.

2.2 Cleaning of data: - This sequence of stages allows for the removal of noise and the correction of data inconsistencies. A data scientist can utilize imputation procedures to fill in missing data using mean of values present. An expert or professional could detect the findings that deviates the result drastically from the normal one. If an outlier in data set says that data is incorrect then simply delete it, if it is undesirable otherwise correct it. Another phase in this process is the elimination of erroneous and incomplete data items.

This is the dataset overview which used in the model that was proposed in the paper:

	District	Event	Circle	Police Station	Caller Source	Event Type	Event Sub-Type	Create Date/Time	Latitude	Longitude
0	LUCKNOW	P01042100004	C1	PS1	PHONE	Information Against Police	Misbehavior By Priv	01-04-2021 00:00	26.834	81.008
1	LUCKNOW	P01042104316	C1	PS1	PHONE	Threat In Person	Attack	01-04-2021 12:09	26.828	81.014
2	LUCKNOW	P01042104847	C1	PS1	PHONE	Dispute	Dispute In Hospital	01-04-2021 12:51	26.840	81.009
3	LUCKNOW	P01042105074	C1	PS1	PHONE	Gambling	Play Cards	01-04-2021 13:10	26.828	81.002
4	LUCKNOW	P01042105152	C1	PS1	PHONE	Threat In Person	Attack	01-04-2021 13:18	26.834	81.033

3. **Feature Selection:-** Choose the dataset's most important features that the proposed model should contains like improving the accuracy and reducing the time to learn from the data set selected.

3.1 Dimensionality reduction: Principal component analysis (PCA), a popular technique for reducing dimensions. It starts with all important dimensions (features) present in data set and using linear algebra technique to condense or reduce to a smaller number. For instance, we use's PCA

to reduce a set of 10 number of features to just 3 features.

3.2 Feature importance: This technique is commonly used in post modeling. For fitting a model to a data set, it should then inspect the various features present in it. The tool should eliminate the features which were not required or having the least importance.

3.3 Wrapper methods: Methods like genetic algorithms and recursive feature elimination must include the procedure to create large subsets of data that includes different feature options. These methods select and remove the features from the data-set that don't cause any change in the result.

4. **Dataset Splitting:** A dataset that should be used for learning of the machines must be divided into three subsets — training, test, and validation sets.

Training set- Part of the data-set which was used for training of the model is called training set which is normally 80 percent of the data-set. It should be chosen randomly to avoid over-fitting. An expert of the domain along with data scientist uses a training-set to train the proposed model with selected data set and define its most favourable features or parameters.

The test set- It was required to evaluate the generalizability of the trained model which describes the ability to identify the pattern by the model has learnt during training. The proposed model for the particular problem that discussed above is due to the lack of generalized approach that may cause over fitting. This must be avoided in any machine learning model by using different subsets for training and testing.

The validation set- The structural parameters in machine learning model's are commonly denoted as hyper parameters. For better configuration of the model these parameters were used. These settings, for example, might convey a model's complexity and capability.

5. **Training Model: -** K-Means Clustering, Decision Tree and ID3 Algorithm are used in proposed model Training to categorize the crime hotspots. Proposed models were trained using training data that was divided throughout the splitting process. We start the training of the proposed model as soon as we gathered and preprocessed data. The data collected above is divided into three parts for better training and learning. Training of the model is started by providing the required data. Based on a threshold or specific or generic value, our algorithm will evaluate the data and produce a model that can cluster and categorize the crime-prone locations. To create a proposed model, goals of the training the model should be known and clear. For this, we have employed an unsupervised clustering approach.

6. **Unsupervised clustering approach K-means clustering**

The k-means technique is then used to cluster the preprocessed crime data. A well-liked clustering algorithm called K-means which divides data into k clusters on the basis of similar crime patterns. Based on the similarity of their crime patterns, the crime data in this instance is grouped into k crime-

prone zones. The elbow approach, a method for figuring out the ideal number of clusters for a certain dataset, is used to figure out how many clusters there should be.

Random forest classification

Using the clustered crime data, a random forest classifier is trained. A prominent categorization strategy known as random forest is used to mix data from many decision trees to create predictions. In this scenario, the random forest classifier is trained to predict the likelihood of crime in specific places based on previous crime trends in those regions. Using clustered data to train the random forest classifier improves its accuracy.

Decision Tree Classification

The Model is also trained on Decision Tree Classification for comparative analyses.

7. **Model Validation:** The goal of this stage or phase is to develop the generalized model that can accurately and rapidly formulate a target value in a reliable manner. A data expert can achieve this goal by adjusting model parameters. These parameters should be optimized in most reliable way to get optimal algorithm performance. Cross validation during the process of model evaluation is the most reliable and successful approach.

7.1 Confusion Matrix- A confusion matrix is a N x N matrix used to evaluate the efficacy of a classification model, where N is the total number of target classes. We compare the actual values with the values predicted by the model with the help of the matrix.

Model in machine learning is considered to be good model has high TP (True Positive) and TN (True Negative) rates while having low FP (False Positive) and FN (False Negative) rates. In our model, we used a confusion matrix since it is always preferred to utilize this as the machine learning model's assessment criterion.

Accuracy and Reliability are the major criteria used to evaluate the performance for our model. These could be computed by using the following: accuracy, precision, recall f1 score, and unweighted average recall (UAR).

7.2 Cross-validation- The most popular tuning technique is cross-validation. It includes creating ten equal folds out of a training dataset. A specific model is trained on only nine folds before being evaluated on the tenth (the one that was previously ignored). Up until every fold is set aside and put to use in testing, training continues. A specialist determines a cross-validated score for each set of hyperparameters using a model performance measure. A data expert uses various sets of hyperparameters in various models, in order to find the accuracy of each model and model with highest accuracy is selected.

To determine which model has the highest prediction accuracy, a data scientist trains models using various sets of hyperparameters. Indicated by the cross-validated score is the model's average performance across ten hold-out folds.

7.3 Retrain the Model: - Utilise a variety of model evaluation strategies after running the model on test data. If an error is greater than expected, the model will be retrained until the error is reduced.

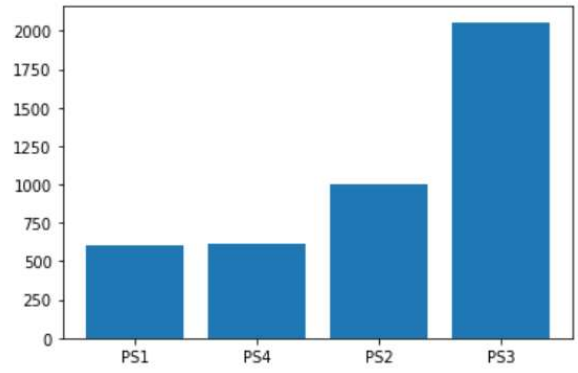
8. Deployment: - After the model has been successfully trained, tested on test data, and the calculation of error, we retrain the model if necessary, and then deploy it using the flask library to produce API so that we can use it to develop our website.

V. Results

The proposed methodology was applied to a real-world dataset of crime incidence in a city. The dataset contained information about different types of crime incidents and their spatial distribution. The dataset was preprocessed to remove any noise and inconsistencies. The preprocessed dataset was then clustered using the k-means algorithm.

The elbow method was used to determine the optimal number of clusters. The elbow method identified 5 clusters as the optimal number of crime-prone areas. The crime data was then used to train a random forest classifier. The random forest classifier was trained on the clustered data to predict crime incidence in different areas.

The below Graph show the Crime occurrence which are categorized by Police Station under a city.



The performance of the proposed methodology was evaluated using a testing set. The proposed methodology is trained using different algorithms for comparative analyses.

Performance with Decision Tree Classifier:

```
Decision Tree Classifier Results:
Accuracy: 0.9122807017543859
Confusion Matrix:
[[145  13   0   0]
 [  7 225  12   4]
 [  2  20 440  13]
 [  0   4  15 126]]
Classification Report:
              precision    recall  f1-score   support

           0       0.94      0.92      0.93       158
           1       0.86      0.91      0.88       248
           2       0.94      0.93      0.93       475
           3       0.88      0.87      0.87       145
```

Performance with Random Forest Classifier:

```
Random Forest Classifier Results:
Accuracy: 0.9103313840155945
Confusion Matrix:
[[128  10  19   1]
 [  4 234  10   0]
 [  1  12 454   8]
 [  0  11  16 118]]
Classification Report:
              precision    recall  f1-score   support

           0       0.96      0.81      0.88       158
           1       0.88      0.94      0.91       248
           2       0.91      0.96      0.93       475
           3       0.93      0.81      0.87       145
```

VI. Conclusions

The world prioritizes crime prevention to build a connected, ideal society in which individuals are secure from all sorts of mental and physical damage. This study addresses the widespread issue of "lack of police force" in regions with regular criminal activity as a crime-deterring tactic. This study provides a new viewpoint on how crime-related data has been explored in the research community by offering a plan to strategically utilise the present police force to decrease the impact of crime. Because criminal disturbance varies by district in relation to a certain time point, the technique outlined in the article demonstrates how the quantity of police force required in a given district is directly dependent on

the instance of time. The dataset for our investigation is a list of crimes that have occurred in the Lucknow District.

Our solution provides the area's frequent patterns by utilizing location-specific components and features. The pattern is employed in the construction of a decision tree model. By training, we develop a model for each area.

on these common patterns. Crime trends are not static since they evolve throughout time. Training entails imparting knowledge to the system depending on specific inputs.

Therefore, by looking at the crime patterns, the algorithm automatically learns the conversion patterns in crime. The components of crime change over time as well. We can discover new causes of crime by looking through the crime data. Full precision cannot be attained because we are just taking a small number of elements into account. Instead of fixing specific traits, we need to look for other crime-related characteristics of locations to improve prediction outcomes. Up until this point, we used specific attributes to train our algorithm, but we intend to add more variables to increase accuracy. Our model forecasts Lucknow's high-crime areas for a specific day. If we take into account a certain state or region, it will be more accurate. The fact that we cannot anticipate when a crime will occur is another issue. We must forecast not only the areas where crime is likely to occur but also the appropriate time because time plays a significant role in the crime.

Comparative analysis of various algorithms that we used to train our proposed model shows the accuracy and performance of different approaches for the Identification of Crime Prone Areas.

REFERENCES

[1] *De Bruin, J.S., Cocx,T.K,Kosters, W.A.,Laros,J. and Kok, J. N (2006) Data mining approaches to criminal carrer analysis ,"in Proceedings of the Sixth International Conference on Data Mining (ICDM"06) ,Pp. 171-177.*

[2] *Tong Wang, Cynthia Rudin, Daniel Wagner, and Rich Sevieri. Detecting patterns of crime with series finder. In Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECMLPKDD 2013), 2013.*

[3] *https://www.altexsoft.com/blog/datascience/machine-learning-project-structure-stages-roles-and-tools/*

[4] *https://www.analyticsvidhya.com/blog/2021/04/steps-to-complete-a-machine-learning-project/*

[5] *https://towardsdatascience.com/5-unique-python-modules-for-creating-machine-learning-and-data-science-projects-that-stand-out-a890519de3ae*

[6] *https://manthan.mic.gov.in/sampledata/PS7%20Predictive%20Policing/PS 7%20predictive%20policing%20sample %20data.xlsx*

[7] *https://www.researchgate.net/figure/Map-showing-crime-prone-areas_fig7_280722606*