**Comparative Study of Approaches for Identification of Crime Prone Areas**

**Abstract:**

The identification of hotspots for crime and the deployment of resources to reduce it are both essential exploratory techniques. Numerous techniques have been developed to identify crime hotspots, but few studies 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. For comparison analysis, we also trained our model using the Support Vector Machine and Decision Tree algorithms. We tested this methodology on an authenticated dataset of Lucknow's crime incidence that we collected from 112 hotline numbers. The outcomes show that the suggested methodology is highly accurate in detecting crime-prone locations.

**Keywords:** Crime Prone, hotspots.

**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 that are more precise and effective. The concept of a crime hotspot has long attracted significant attention in the field of environmental criminology. 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. On the one hand, it aids in locating high-risk areas that demand additional care or perhaps intervention. On the other hand, crime hotspots can aid in illuminating the factors that influence the concentration of crime. This may inform the allocation of resources and the creation of policies. Crime hotspots are currently not well defined, and the word "hotspot" has many different connotations. In general, there are two ways to think about crime hotspots. Firstly, regions where crimes are concentrated, are referred to as crime hotspots. Methods like hotspot mapping (e.g., kernel density estimation) and clustering (e.g., K-means clustering, hierarchical clustering) can be used to find criminal hotspots. Secondly, crime hotspots require not only the clustering tendency of crimes but also significant concentration. In this way, hotspots refer to high-value clusters of crime cases, and the significance of clusters can be measured by some spatial statistics.

In this research, we conduct a comparative analysis of methods for locating crime hotspots, including Decision Tree Classification, Random Forest, and Support Vector Machines 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.

**Literature Review:**

Areas that are more likely to experience crimes are called crime-prone areas. These places are classified as having a high threshold for the crime.

In recent years, there has been a lot of interest in the topic of identifying crime-prone locations using machine learning. Numerous studies have been conducted on this subject, examining various methods for predicting crime trends and hotspot locations using machine learning algorithms. The following important studies offer a review of the relevant literature on the 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 crime prediction are all thoroughly reviewed in this research. The difficulties and restrictions of applying machine learning to crime prediction are also covered by the writers.

N. R. Charan and S. Jana's "Crime Hotspot Prediction Using Machine Learning: A Survey" In this paper, support vector machines, random forests, and artificial neural networks are covered in detail as machine learning techniques for crime hotspot prediction. The writers also go over the many traits and data sources that are used to forecast crime.

S. P. Patil, K. R. Chavan, and A. R. Kulkarni's "Predicting Crime Using Machine Learning Techniques: A Review" In this article, various machine learning methods for crime prediction, such as decision trees, k-nearest neighbours, and artificial neural networks, 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 review of the literature on the subject of locating crime-prone locations using machine learning techniques. They go over the many methods, strategies, and difficulties associated with crime prediction and offer insightful information for next studies.

**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 model is trained using the Lucknow District's 112 hotline crime data. Using the k-means algorithm, crime data is clustered, and a random forest classifier is trained on the clustered data to predict where crimes will occur. We also trained our model using the Support Vector Machine and Decision Tree techniques for comparative analysis. On a verified dataset of Lucknow's crime incidence that we compiled from 112 hotline numbers; we tested this methodology. The results show that the suggested methodology is highly accurate in detecting crime-prone locations.

**Methodology:**

The proposed methodology consists of the following steps:

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 process of gathering the data needed for model training is known as data collection. Prior to gathering data, we identify the type of issue we are trying to resolve, look into the sources of data that are available, then look into data that is readily 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.

1. **Data Preprocessing:-** Preprocessing is done to transform raw data into a format that can be used by machine learning. A data scientist can use an applied machine learning model to obtain more accurate findings by using structured and clean data. Data formatting, cleansing, and sampling are all part of the method.

**3.1 Data formatting-** When information is gathered from diverse sources by different persons, the significance of data formatting increases. Standardising record formats is a data scientist's first job. A specialist examines the consistency of the recording of variables for each attribute. Examples of variables include names of goods and services, costs, date formats, and addresses. Aspects represented by numeric ranges are similarly subject to the notion of data consistency.

**3.2 Data cleaning.-** This series of steps enables the elimination of noise and the correction of data discrepancies. Imputation approaches, such as replacing missing values with mean characteristics, can be used by a data scientist to fill in missing data. A specialist can also spot outliers or observations that dramatically depart from the distribution. If an outlier points to inaccurate data, a data scientist deletes or, if necessary, corrects it. The removal of erroneous and incomplete data objects is another step in this process.

This is the dataset overview which this paper is proposed on:



1. **Feature Selection:-** Choose the dataset's most important features to model. boosting accuracy and maybe cutting down on overfitting and training time (reduced total data and redundant data to train on).

**4.1 Dimensionality reduction:** Principal component analysis (PCA), a popular technique for reducing the number of dimensions, starts with a high number of dimensions (features) and using linear algebra to condense them to a smaller number. For instance, you could use PCA to reduce a set of 10 number features to just 3.

**4.2 Feature importance (post modeling):** Fit a model to a set of data, then inspect which features were most important to the results, and remove the least important ones.

**4.3 Wrapper methods-** such as genetic algorithms and recursive feature elimination involve creating large subsets of feature options and then removing the ones which don’t matter.

1. **Dataset Splitting:-** A dataset used for machine learning should be partitioned into three subsets — training, test, and validation sets.

**Training set.**- A data scientist uses a training set to train a model and define its optimal parameters — parameters it has to learn from data.

**The test set-** A test set is necessary to evaluate the generalizability of the trained model. The latter describes a model's ability, after training on training data, to recognize patterns in new, unseen data. The above-discussed lack of generalizability is caused by model overfitting, which must be avoided by using varied subsets for training and testing.

**The validation set-** A model's hyperparameters—higher-level structural parameters—cannot be directly learned from data, thus a validation set is used to fine-tune them. These settings, for example, might convey a model's complexity and how quickly it identifies patterns in data.

1. **Training Model:-** K-Means Clustering and the Decision Tree ID3 Algorithm are used in Model Training to categorise the crime hotspots. Models are trained using training data that has been divided throughout the splitting process. We start model training after preprocessing the gathered data and dividing it into three parts. This procedure comprises "feeding" training data to the algorithm. Based on a threshold value, our algorithm will evaluate the data and produce a model that can cluster and categorise the crime-prone locations. To create a model is the goal of model training. For this, we have employed an unsupervised clustering approach.
2. **K-means clustering**

The k-means technique is then used to cluster the preprocessed crime data. A well-liked clustering algorithm called K-means divides data into k clusters based on how similar they are. 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**

A random forest classifier is trained using the clustered crime data. Data from several decision trees are combined to produce predictions via a popular classification approach called random forest. The random forest classifier in this case is trained to predict the likelihood of crime in specific locations based on those locations' past crime trends. The accuracy of the random forest classifier is increased by using the clustered data to train it.

**Decision Tree Classification**

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

1. **Testing The Model:-** This step's objective is to create the most basic model that can accurately and quickly formulate a target value. Through model adjustment, a data scientist can accomplish this objective. To attain an algorithm's optimal performance, model parameters are optimised in this manner.

Cross-validation is one of the more effective techniques for model evaluation and adjustment.

**7.1 Confusion Matrix-** An N x N matrix called a confusion matrix is used to assess the effectiveness of a classification model, where N is the total number of target classes. In the matrix, the actual goal values are contrasted with those that the machine learning model anticipated.

High TP and TN rates and low FP and FN rates are characteristics of a good model. We employed a confusion matrix in our model because it is always preferable to use this as the machine learning model's evaluation criterion.

For our model, the measures utilised to assess model performance were 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. 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.

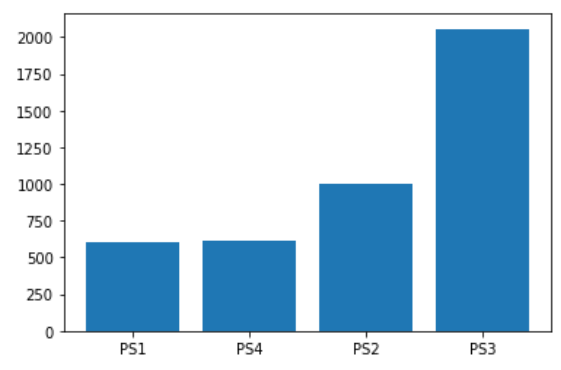
1. **Retrain the Model:-** Utilise a variety of model evaluation strategies after running the model on test data. If an error is greater than expected, the model will be retrained until the error is reduced.
2. **Deployment:-** After the model has been successfully trained, tested on test data, and the 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.

**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.

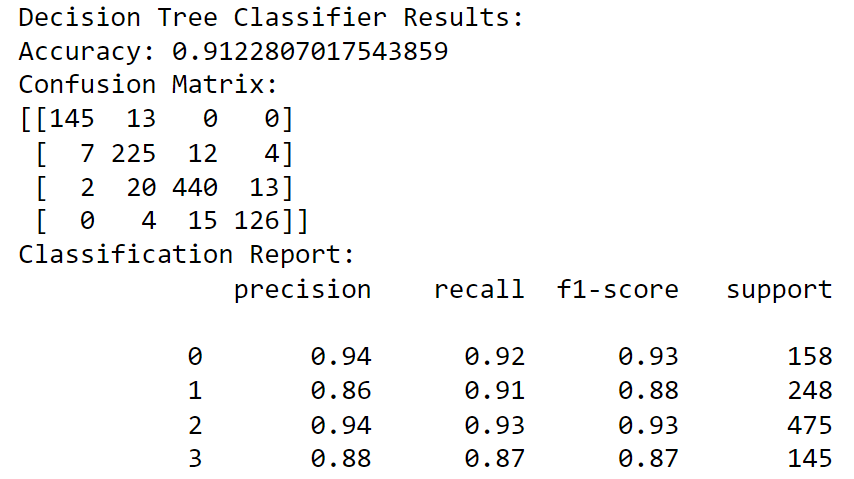
The elbow method was used to determine the optimal number of clusters. The elbow method identified 5 clusters as the optimal number of crime-prone areas. The crime data was then used to train a random forest classifier. The random forest classifier was trained on the clustered data to predict crime incidence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

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

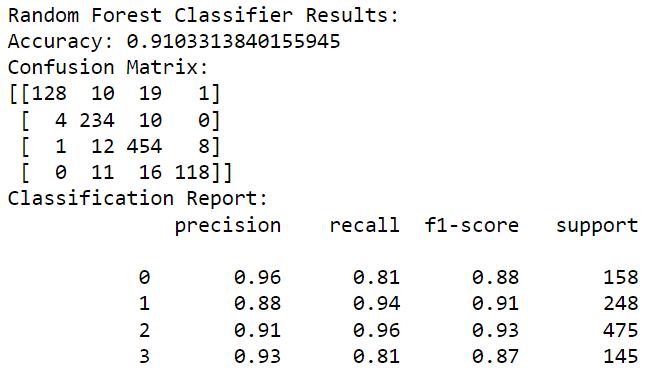


The performance of the proposed methodology was evaluated using a testing set. The proposed methodology is trained using different algorithms for comparative analyses.

**Performance with Decision Tree Classifier :**



**Performance with Random Forest Classifier:**



**Conclusion:**

Prevention of criminal activity is among the top most

priorities of the world in the building of a well-connected,

a halcyon community where people are protected from all kinds

of mental and physical harm. This paper provides a means

of preventing crime by providing a possible solution to the

a common problem of “lack of police force” in regions with

regular criminal activity. This paper provides a new outlook

in the way how crime related data has been explored in the

research world by proposing a way to judiciously use existing

police force to alleviate the impact of crime. The methodology

delineated in the paper gives an example of how the quantity of

police force needed in a particular district is directly dependent

on the instance of time as criminal unrest varies from district to

district concerning a particular time point.

Prevention of criminal activity is among the top most

priorities of the world in the building of a well-connected,

a halcyon community where people are protected from all kinds

of mental and physical harm. This paper provides a means

of preventing crime by providing a possible solution to the

common problem of “lack of police force” in regions with

regular criminal activity. This paper provides a new outlook

in the way how crime related data has been explored in the

research world by proposing a way to judiciously use existing

police force to alleviate the impact of crime. The methodology

delineated in the paper gives an example of how the quantity of

police force needed in a particular district is directly dependent

on the instance of time as criminal unrest varies from district to

district with respect to a particular time point.

Prevention of criminal activity is among the top most

priorities of the world in the building of a well connected,

halcyon community where people are protected from all kinds

of mental and physical harm. This paper provides a means

of preventing crime by providing a possible solution to the

common problem of “lack of police force” in regions with

regular criminal activity. This paper provides a new outlook

in the way how crime related data has been explored in the

research world by proposing a way to judiciously use existing

police force to alleviate the impact of crime. The methodology

delineated in the paper gives an example of how the quantity of

police force needed at a particular district is directly dependent

on the instance of time as criminal unrest varies from district to

district with respect to a particular time point.

The world places a high priority on preventing crime in order to create a linked, utopian society where individuals are safe from all types of emotional and physical harm. The prevalent issue of "lack of police force" in areas with consistent criminal activity is addressed in this paper as a strategy of deterring crime. By suggesting a strategy to wisely use the current police force to lessen the impact of crime, this study offers a unique perspective on how crime-related data has been investigated in the research community. Because criminal disturbance differs from district to district with respect to a specific time point, the approach described in the article provides an example of how the amount of police force needed in a particular district is directly dependent on the instance of time. Our study's dataset is a list of crimes that have happened in the Lucknow District.

Our solution uses location-specific components and attributes to provide the area's prevalent patterns. The pattern is used to build a decision tree model. We create a model for each location by training 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%20 policing%20sample %20data.xlsx

**7.** https://www.researchgate.net/figure/Map-showing-crime-prone-areas\_fig7\_280722606