PREDICTION OF CRIME TYPE AND OCCURRENCE USING MACHINE LEARNING ALGORITHM

Dr. A. Mummoorthy*1, Ms. A. Jaya Sree*2, Nallavalli Amulya*3, Karnatakapu Pavan Kalyan*4, Mahendrakar Anirudh*5

*1Associate Professor, Department Of Information Technology Malla Reddy College Of Malla Reddy College of Engineering & Technology Hyderabad, India.
*2Assistant Professor Department Of Information Technology Malla Reddy College Of Malla Reddy College of Engineering & Technology Hyderabad, India.
*3,4,5Final Year Student Department Of Information Technology Malla Reddy College Of Engineering & Technology Hyderabad, India.

DOI: https://www.doi.org/10.56726/IRJMETS51646

ABSTRACT

The rate of crime in India is increasing daily. Criminals have the resources to conduct crimes in the modern world thanks to social media, new technical advancements, and contemporary methods. A methodical approach that categorizes and examines crime trends in both analyses is called crime prediction. For pattern detection and criminal analysis, clustering techniques abound, but they don't always provide the full picture. The K means approach among them offers a more accurate means of predicting the outcome. Predicting the age groups more or less likely to commit crimes and the locations with higher crime rates is the major goal of our article. To reduce time complexity and increase result efficiency, we provide an enhanced K means technique.

Keywords: Crime, Clustering, And The Optimization Of The K-Means Technique.

I. INTRODUCTION

Every region of India is seeing an increase in crime. The main goals of crime analysis techniques are to recognize and forecast criminological patterns and trends. The police can identify different age groups that are more vulnerable to criminal activity by using our data mining technology, which we utilize to forecast the crime-prone locations. The identification of an effective clustering technique is necessary to complete the study of crimes. In data mining, we employ the K means clustering approach to classify a set of items based on how similar they are to each other compared to other categories. Numerous research publications deal with the prediction of criminal episodes and suggest different clustering approaches. In order to predict the result, we extract and analyze data in this study utilizing a successful clustering technique. K-means is among the better algorithms for resolving a well-known clustering issue. After a set number of iterations, a K-means algorithm designed for data clustering will minimize changes in cluster values. The cluster number is found by identifying the optimal k value. Value will aid in choosing a centroid value that is more advantageous. Right now, crime is increasing rapidly in India. Our study uses a larger dataset, and manually selecting the k value from this massive dataset is a laborious, complex, and complicated process that takes a long time.

We decide to find the k value using the elbow approach in order to solve this issue. We use the spider program to construct optimal K-means clustering in Python.

II. LITERATURE REVIEW

Shiju Sathyadevan et al. explain that in order to analyze the crime rates of each state, data classification is done through collection, classification, pattern recognition, prediction, and visualization[1]. They developed a model to categorize each offense using Naive Bayes classifiers. In this supervised learning task, you must suggest the class for every additional data point that is provided even though you already know the class for a set of training data points. With the use of decision trees, the apriori algorithm can be used to forecast outcomes and spot trends in criminal activity [2]. To determine each activity’s level, a heap map is used. Dark colors, for instance, are associated with low activity. By using improved feature selection techniques and the revised Iterative Dichotomiser3 algorithm, a decision tree approach can be used to identify suspicious emails [3]. Applying these...
fundamental elements makes decision trees faster and more aesthetically pleasing. The decision tree's parameters were optimized by the application of the genetic algorithm. Machine learning, supervised learning, and natural language processing (NLP) techniques are used to classify data and forecast crimes.

Furthermore, the data fusion approach's use of deep neural networks can forecast the likelihood of crimes [4]. Time and location records from the CLEAR system are utilized to create K-means clusters, which are then used to forecast crimes. The dataset includes parameters such as description, month, day, hour, etc., together with statistics on the prevalence of crime [5]. It primarily focuses on a method that may be used in conjunction with both numerical and categorical variables. Several unique data mining techniques are used in the development of crime analysis tools. It helps law enforcement officials in their crime-solving efforts [6]. By applying a clustering technique to crime records, analysis of crimes is made possible [7]. It makes it possible to identify and investigate a wide range of criminal activity patterns from start to finish. Finding local optima is facilitated by the cluster that is produced by the K-means algorithm, which randomly selects initial starting locations [8]. To resolve this, the beginning centroid for K-Means clustering was found using the partitioned data along with the data axis with the highest variance. Consequently, it is seen that the proposed technique needs less repetitions, resulting in a shorter clustering time. By applying merge sort, the Hidden Markov Model (HMM) can be more effectively grouped using the K-means algorithm [9].

**Drawbacks:**
- High computational requirements
- Needs large amount of labeled data
- Training takes more time

### III. PROPOSED METHOD

Crime is the biggest threat to humanity. There are a lot of crimes that happen frequently. Perhaps it is spreading rapidly and broadly. Small towns, villages, and big cities all experience crime. Crimes come in a variety of forms, such as rape, robbery, false imprisonment, assault, murder, and homicide. The increased rate of crime necessitates much faster case resolution. The police force is responsible for keeping an eye on and curbing the criminal activity, which is growing more quickly. The police department faces two primary challenges: criminal identification and crime prediction, given the abundance of available crime data. Technology is required in order to speed up the case solution process.

"With this new regression-based feature section algorithm, we can achieve the best accuracy by avoiding the noise information in feature engineering."

The above problem made me to go for a research about how can solve a crime case made easier. Through many documentation and cases, it came out that machine learning and data science can make the work easier and faster.

The aim of this project is to make crime prediction using the features present in the dataset. The dataset is extracted from the official sites. With the help of machine learning algorithm, using python as core we can predict the type of crime which will Occur in a particular area.

The objective would be to train a model for prediction. The training would be done using the training data set which will be validated using the test dataset. Building the model will be done using better algorithm depending upon the accuracy.

(XGBoost, AdaBoost, Random Forest, KNN) classification algorithms will be used for crime prediction. Visualization of dataset is done to analyze the crimes which may have occurred in the country. This work helps the law enforcement agencies to predict and detect crimes in Chicago with improved accuracy and thus reduces the crime rate.

### IV. METHODOLOGY

This algorithm splits the whole process into two stages. Phase I: In this phase, the initial centroids are computed. The Elbow method algorithm is used to find the ideal k value in this process.

First algorithm: The algorithm for k-means clustering Enter:
T is equal to \{T_1, T_2,... T_{10}\}. Product: Eight clusters in all. \( K = 8 \). Steps:
1. Select eight data items at random from set T to use as beginning centroids;
2. Repeat.
3. Determine the separation between each data point \( T_i \) and its centroids, then assign \( T_i \) to the cluster with the closest centroid.

Phase II: In this phase, the distances between each data item and the starting centroids are computed. The data points are then grouped with the clusters that have the closest centroids.

We'll record the distance between each data point and the cluster with which it is related. After that, the cluster centroid will be found again. The above procedure is repeated until the newly obtained cluster value is equal to the prior cluster value.

We're working on implementing Spyder. Here, Spyder 3.7 is being used. Spyder is an integrated development environment designed for systematic Python programming. Here, a number of packages were implemented, such as matplotlib, numpy, sklearn, pandas, etc. This makes charting easier. We're working on implementing Spyder. Here, Spyder 3.7 is being used. An integrated development environment for methodical Python programming is called Spyder. Several packages, including matplotlib, numpy, sklearn, pandas, etc., were implemented here. Which aids in the plotting of a data frame table and elbow graph using the \( K \)-means clustering algorithm?

As seen in Fig. 1, datasets are gathered from Kaggle datasets and imported into Spyder in CSV format.

Using the elbow approach, we do normalization in order to determine the precise number of clusters (\( k \)). The elbow technique computes the SSE by applying \( k \)-means clustering for a range of \( k \) values (2–15) on the acquired dataset. For every value of \( k \), the SSE is represented as a line chart. The elbow graph (Fig. 2) exhibits a little fluctuation at cluster value 8, which led to its recognition as the optimal \( k \) value. Our objective is to determine a modest number of \( k \) that maintains the surviving clusters and a low sum of squared errors. The elbow at \( k=8 \) in Fig. 4 is a clear indication that 8 is the ideal number of clusters. Thus, we are clustering the dataset into eight distinct groups according to the value of \( k \), which is eight.

DATA ANALYSIS:
The algorithm's flow
1. Data Collection: Compile past crime statistics with details about the place, time, kind of crime, and any other pertinent characteristics. Public sources such as government databases and APIs are where you can find this information.
2. Data Preprocessing: Use methods like one-hot encoding to transform category variables into numerical representations, handle missing values, and eliminate outliers from the data.
3. Engineering features: Determine important characteristics from the unprocessed data that can be used to forecast the kinds and frequency of crimes. This could include elements like the day of the week, the time of day, the location, the socioeconomic status of the region, etc.
4. Model Selection: Select suitable machine learning models for regression (to forecast occurrence counts) and classification (to forecast crime categories).
5. Neural Networks, Forests, and Support Vector Machines are examples of models that can be used for occurrence prediction; models such as Poisson or Linear Regression.
6. Model Training: Use the training data to train your selected models, then divide the data into testing and training sets.
7. Model Evaluation: Assess your models' performance using relevant metrics, such as mean squared error or mean absolute error for regression, F1 score for classification, accuracy, precision, and recall.
8. Deployment: After you're happy with the model's functionality, use it in a practical application.
9. Monitoring and Maintenance: To guarantee your model's accuracy and applicability over time, keep a close eye on how it performs in the real-world setting and occasionally retrain it using new data. It's crucial to keep ethical considerations in mind throughout this process, such as data bias and the potential effects of the
model’s predictions on people and communities. Furthermore, make sure that all laws pertaining to data security and privacy are followed.

V. RESULT AND ANALYSIS

The dataset of Indian crimes is displayed in Fig 1. This dataset includes extensive information on a wide range of crimes committed in India between 2001 and 2010. The dataset is rescaled using normalization, with a total of 1053 values. The pre-processed dataset that results is then subjected to the k-means clustering technique. Finding the ideal number of clusters to divide the data into should be the first step in developing an unsupervised algorithm. Plot the total within curve in accordance with the number of clusters, k. We get the optimal and most precise k value by selecting numbers at random between 2 and 15. We examine the changes in the eighth value from this graph, and as a result, we determine that k=8.

The variables for storing total inside sum of squares for each k-means are labeled total_within_ss on the y-axis and no_of_clusters on the X-axis of the graph above.

There is a bend (knee) in the plot, which often indicates the number of clusters. That is, k=8. Here, we select k for the iteration from 2 to 8 based on the graph in Figure 2. The dataset is created using k-means clustering after the ideal k value has been determined.
The results of the clustering are displayed below. There are more male offenders overall in cluster 2, or value 13746. In cluster 6, there are more female offenders. There are more males under the age of 18 in cluster 4. In cluster 4, there are more females under the age of 18. Following clustering, fig. 4 displays the additional clusters as the dataset’s second attribute, as seen below.

This implementation outcome serves as the year 2001–2010 case study. It displays the illicit activity of the entire state. The case study makes it abundantly evident that the majority of male and female criminals are from Maharashtra and Madhya Pradesh, respectively. There are more male and female offenders under the age of eighteen in cluster 4, or those from Chhattisgarh.

Fig 5 clearly shows the total number of criminals is more in UP i.e. 13983 criminals. This result will help the police departments and other authorities to be more aware of the criminal behaviors and also to plan different strategies according to this result to prevent future crimes.

VI. CONCLUSION

Python has been used for the implementation. Based on the results of each cluster, we may determine which state has a higher or lower number of criminals. Increased awareness among the authorities is definitely beneficial. The technique of grouping similar data points into distinct groups is called clustering. There are no similarities between the points in various clusters, and the points within a cluster are as similar as feasible. The optimized K-means algorithm produces an efficient result with fewer iterations and better accuracy for the final cluster.

Predicting crimes is aided by this model. A machine learning method can forecast the offender’s age, sex, and relationship. In this case, the classifier and regression that were used yield an accuracy of almost 98%. The dataset can be enhanced and used in other countries if the circumstances are nearly the same. An overall prediction of crime is provided by the model.

This model can be enhanced by using deep learning techniques.

VII. REFERENCES


