KNOWLEDGE DISCOVERY IN DATABASES FOR PREDICTION OF FUTURE CRIMES

Ishita Rastogi¹, Anjali Jha², K.C. Prabu Shankar³

¹Computer Science and Engineering, SRM Institute of Science and Technology, Chennai, India. E-mail: rastogishita30@gmail.com
²Computer Science and Engineering, SRM Institute of Science and Technology, Chennai, India. E-mail: anjalijha1999@gmail.com
³Assistant Professor, Computer Science and Engineering, SRM Institute of Science and Technology, Chennai, India. E-mail: prabushc@srmist.edu.in

ABSTRACT

Crime is a pernicious and pervasive social phenomenon that affects people all over the world. Every day, the ever-increasing crime rate has a negative impact on a country's economic development. As a result, identifying various causes, crime incidence relationships, and finding an optimal way to minimise crime rates is very important. Preventive steps must be taken by law enforcement in order to minimise crime rates. Precise real-time crime forecasts assist in minimising crime rates, yet they remain a difficult challenge for scientists to solve because crime events are influenced by a variety of complex factors. In this work, various data mining techniques are used for predicting the crime. Data mining is the process of analysing and reviewing vast preliminary databases for producing new knowledge that can be useful to the institution. The fresh and useful data is extracted with the aid of pre-existing datasets. Numerous perspectives to data mining research and prediction had been considered. The raw datasets were pre-processed according to the need, in the first step. Following that, data mining algorithms were used to extract information from these vast databases and extract hidden associations, which were then used to discover and report crime trends, which is useful for the crime analysts for analysing these networks for crime prediction and thus helps in crime prevention.

Keywords: Crime, Real-time crime Forecasts, Crime rates, Crime prediction, Datasets, Data Mining

I. INTRODUCTION

The crime rate is steadily rising day by day. Crime is one of those crucial factors that influence various important life decisions such as roaming at the appropriate time, moving to a new place, avoiding dangerous areas, and so on. Crimes have a negative impact on a community's reputation. Crimes affect a country's economy by putting a financial strain on the government because of the requirement of additional police forces, security, courts, and other services. Since crime is on the rise, we are at an alarming stage in terms of reducing it even more quickly. Previously, it has been the responsibility of criminal justice and law enforcement professionals to solve crimes. As the use of computerised systems to monitor crimes and trace offenders has grown, computer data analysts have begun to lend a hand in assisting law enforcement officers and investigators in speeding up the crime solving process. Since crime is neither spontaneous nor systematic, it is impossible to predict it. Criminology methods may be used to determine suspects and the likelihood of a crime occurring. The police department, detective departments and crime divisions use criminology to help them determine the true attributes of a suspect. Since the 19th century, the department of criminology has been included in the trials of criminal investigations. Crime is a social blight that costs our society severely in a variety of ways. The government of India has also taken measures to create applications and tools for use by federal law enforcement agencies and the state in conjunction with the National Crime Records Bureau (NCRB)[6].

In data mining, knowledge discovery refers to obtaining useful information or extracting patterns from vast quantities of data that lead to a significant prediction. When it comes to the data to be extracted, there are two types of data mining: Description mining and Classification with Prediction [2]. Mining of association rules,
frequent patterns, clusters or associations is commonly referred to as Description mining. Classification and prediction entails using probability equations to assign a class mark to data and then using numeric measures to forecast any features. While the accuracy of the expected results cannot be guaranteed, the results show that our application helps to reduce crime rates to some degree by providing protection in crime-prone areas. Thus, in order to create such an effective crime analytics tool, we must first collect and evaluate crime records [5]. And though we can’t predict who will be the crime’s victims, we can predict the location where it is most likely to occur.

Only in some past decades, spatial data mining has become a realistic, inexpensive, and accessible method for a huge range of law enforcement officials due to the technology. We collect crime data from different sources such as blogs, web pages, news sites, social media and so on because the availability and accessibility of criminal records is restricted. This massive amount of information is used to create a crime database. The biggest obstacle we're facing is collecting crime reports from the law enforcement, which could impact the program's accuracy because it's based on the training set's accuracy. Identifying trends and patterns in crime is a difficult task. Crime investigators will spend a considerable amount of time sifting through data to see if a particular crime fits into an existing pattern in order to locate a trend. The data must be categorised as a new pattern if it doesn’t recognise with an existing pattern. It is possible to predict, foresee, and deter crime after identifying a trend. Clustering algorithms have previously been used for crime detection. In this case, 80 percent of the data will be trained using the provided algorithms, and the remaining 20% will be checked. The goal is to use an effective predictive model to predict the top most features that influence the high crime rate, which will ultimately assist police or law enforcement officials in taking the required measures.

The goal is to examine the most common types of crime over a given time span in order to predict potential future crime at a specific location using an effective predictive model that will ultimately assist police or law enforcement officials in taking appropriate action. A detailed implementation of Decision trees and Random Forest classification has been made for this reason. In Decision Trees [2], every internal node represents an attribute test, every branch represents a test result, and every leaf node represents a class label. The Random Forest [2] takes into account a number of different trees. Each tree classifies the data on its own, and the algorithm then selects the classification that the majority of individual trees agree on. As a result, the Random Forest Classifier algorithm was instrumental in obtaining an adequately predictable collection of features using decision trees.

In this paper, the following steps are involved for doing the crime analysis:

1. Collection of Data
2. Preprocessing
3. Data cleaning
4. Training and Testing dataset
5. Best model accuracy
6. Prediction

![Crime Analysis Diagram](image-url)

**Fig. 1**

II. LITERATURE SURVEY

A. Literature Review

[1] Hitesh Kumar Reddy. et al. developed a platform using Google Maps and a number of R packages that gives a configuration to visualise crime systems and analyse them using a number of machine learning algorithms. To extract information and discover unknown associations in data sets linked to crimes, machine learning algorithms are used to. They haven't used any other classification algorithms that could increase prediction accuracy. [4]
Devan M.S. et al. discuss the system that predicts crime prone areas. They used various test sets to assess the accuracy of classification and prediction. The Bayes theorem is used to classify the data, which has a high accuracy rate of over 90%. They developed a model using this algorithm after training it on a large number of news papers. Their system inputs attributes/factors of a location and uses the Apriori algorithm to generate recurrent figures for that location. The pattern/figure is used to build a decision tree model. Their software can predict crime hotspots in India on a specific day, but not the time, which is also a key factor in crime. [3] H. Benjamin Fredrick David et al. used data mining as the main approach to predict and analyse the crime. This paper was a comparative analysis of various articles and methods. Text/NLP, spatial, and geolocation-based approaches were the main methods used. [7] Rohit Patil et al. propose their crime prediction model in statistical implementation. They used K-means to divide the data into clusters based on high and low values. The K-means result is used as an input to the Apriori algorithm, which finds associations between a variety of other attributes. They were unable to compare this method because they were unaware of such a criminal prediction system implementation model in the region. [8] C. Vijayalakshmi et al. did major work on murders happening in a specific area but have considered various types of crimes also. The best outcome was determined and taken into consideration using Bayesian, Levenberg, and Scaled algorithms on train and test results. The best result was obtained with the scaled algorithm, meaning a precision of .78. They didn't apply it to large amounts of data. [2] Prajakta Yerpude et al. conclude that in terms of accuracy, precision, recall, and F1 ranking, the Random Forest Classifier produces the most balanced results. Though Linear Regression produced the low accuracy results in these output stages, the data did not fit well into the straight line when the target and remaining features were taken into account. [5] Malathi A. et al. concentrated on Apriori and MV algorithms, with a few improvements, for helping in the procedure of putting in remaining values and identifying crime trends. They used these approaches on real crime data, but for information discovery from the crime reports, they used a semi-supervised learning method in this paper. [6] Manish Gupta et al. proposed an interface in order to carry out police activities effectively, as a crime analysis method for CCIS focused on existing decision support and data mining techniques. Their system too, doesn’t give an exact time prediction for analysis. [9] Neetu Singh et al. suggest that the Multiple Linear Regression model assists in building a prognosticative model that anticipates the sorted out instances with 85.5 percent of accuracy. The city with the highest crime risk zone among the others has been shown by the K-Means Clustering. Although, they worked with the dataset limited to a specific state. [10] Tushar Sonawanev et al. conclude their work in the form of a connection between different crimes and their location, i.e. state/city. The crime prediction is illustrated using a variety of diagrams, including pie charts, heat maps, spikes, and graphs. It's difficult to make a precise forecast solely based on graphs.

B. Inference from the survey

Crime prediction depends on a lot of factors and the prediction can never be 100% accurate as the crimes take place randomly [4]. Although, studying the authentic datasets very carefully, the locations most prone to crimes can be known. To make the prediction more accurate, time prediction is also an important factor [4]. Even though, the victims and culprits can’t be predicted, the crime can be prevented if the time, location and type of crime that is going to occur is known. All the approaches used until now give the results up to a certain accuracy. We mainly need to focus on finding ways to increase the accuracy of the prediction which will make real-time use of the system easier. Moreover, using huge data sets along with the combined algorithms will also contribute towards improving the accuracy of the result.

C. Analytical Insights

By predicting the crime rate of various incidents that occur in different places, it is possible to decide which crimes should be taken into account and given greater attention, as well as to make people aware of all possible threats in their area. Models use basic assumptions, a variety of methods, and a huge number of data sets in order to achieve the highest precision in the shortest amount of time. These types of models will give us a better chance of coping with problems and determining which events should be overlooked and which should be taken into account.

D. Demerits

No reliable prediction can be made due to the limited datasets. The law enforcement department's ability to provide crime data reports is limited. The consistency of the training set decides the program's accuracy. One of
the most common roadblocks faced in all of the studies was the failure to relate to vast volumes of data in order to produce results based on comparative analysis.

III. PROPOSED WORK

A. Description

Our proposed system will focus on increasing the accuracy rate of the prediction and taking the least time possible in providing the results. The ability to pre-process datasets and select the most accurate algorithm is a crucial attribute of our proposed system. The system would also be able to work with huge datasets. Our system would adapt to the newly fed datasets and predict accordingly. This system will make predictions based on a larger and more diverse dataset. It will examine the attributes and make predictions according to which, the preventive measures can be taken. We hope to build a system that will predict what type of crime is most likely to occur at a particular location during a specific time.

We plan to include features like longitude and latitude of a region to define the location, date and time of occurrence. The model will employ current advanced methods and practices to provide reliable and accurate prediction rates and recommend appropriate preventive measures. In conclusion, our model should be able to accurately predict what type of crime is most likely to occur at a particular location during a specific time and date.

B. Merits

Our framework would accurately predict crime, allowing law enforcements to take appropriate preventative measures. It is beneficial as our system would be able to provide:

1) Immediate steps to deal with it.

2) Specific steps to avoid recurrences of previous offences

3) Comprehensive strategies that recommend formal action against a particular type of crime.
IV. SYSTEM ARCHITECTURE

A. Dataset details

This dataset includes recorded crime events from 2001 to the present, except the most recent seven days. The information is taken from the National Crime Records Bureau (NCRB). Addresses are only displayed at the block level to secure the crime victims’ privacy, and individual locations are not known. This data set will be used in our research.

The attributes of the original dataset are as follows:

<table>
<thead>
<tr>
<th>Attribute ID</th>
<th>Attribute Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ID</td>
<td>Case ID</td>
</tr>
<tr>
<td>CaseNumber</td>
<td>Case Number</td>
</tr>
<tr>
<td>Date</td>
<td>Date of crime</td>
</tr>
<tr>
<td>Block</td>
<td>Block ID</td>
</tr>
<tr>
<td>PrimaryType</td>
<td>Crime type</td>
</tr>
<tr>
<td>Description</td>
<td>Crime Description</td>
</tr>
<tr>
<td>Location description</td>
<td>Location description</td>
</tr>
<tr>
<td>Latitude</td>
<td>Latitude of location</td>
</tr>
<tr>
<td>Longitude</td>
<td>Longitude of location</td>
</tr>
</tbody>
</table>

B. Data pre processing

Data visualisation and cleaning are examples of data pre-processing. Converting time stamps to break year, months, day, hour, minute, and seconds is part of data cleaning.

- \( \text{train\_data}'[\text{yr}'] = \text{pd.TimedeltaIndex(train\_data}'[\text{timestamp'}]).yr \)
• train_data['month'] = pd.Timestamp(train_data['timestamp']).month
• train_data['date'] = pd.Timestamp(train_data['timestamp']).date
• train_data['hr'] = pd.Timestamp(train_data['timestamp']).hr
• train_data['min'] = pd.Timestamp(train_data['timestamp']).minute
• train_data['sec'] = pd.Timestamp(train_data['timestamp']).second

C. Training using Random Forest Algorithm and Decision Trees

The dataset has been split into train and test set, and the models are trained using the Random Forest and Decision Tree machine learning algorithms. The dataset is divided as follows:

X_Train represents the following input

train_data[['Latitude', 'Longitude', 'yr', 'month', 'date', 'hr', 'min']]

Y_Train represents the following input

train_data[['primarytype']]}

As we are predicting the crime type, we set the Y column/ target column as Crime Type.

D. Crime prediction

The crime type is predicted by obtaining test feedback from the user through a Graphical User Interface (GUI). To predict the type of crime in a given area, the latitude, longitude, and timestamp should be provided as data. Data is pre-processed for the appropriate format after the user input is received.

X_test = np.array(input).reshape(1,-1)

For Random Forest model, the trained model is used for prediction as below:

rf.predict(X_test)

For Decision Tree model, the trained model is used for prediction as below:

df.predict(X_test)
Multiple interpretations of what crime entails, as well as non-systematic classification of the related offences and a lack of recommended measures, are hindering the management and orchestration of effective directives, regulations, and legislative initiatives at the state, national, and international levels, resulting in inadequate crime
incident handling. This paper contributes towards the efficiency of crime prevention measures by proposing a schema-based crime incident description that: 1) describes the characteristics of a crime incident and their possible elements, and 2) offers a two-level offence classification scheme based on relevant criteria. The proposed schema may be extended with a list of suggested acts, related steps, and appropriate policies to combat the form of offence and, as a result, the specific incident. This matching would allow for more effective tracking, handling, and moderating of various crimes and their manifestations as individual incidents.

![Fig. 5 Pictorial Representation of the Occurrences of Types of Crime](image_url)

We analysed top 5 crimes that occurred daily, monthly and weekly.
Fig. 6(a) 

Top 5 daily crimes 2020

Fig. 6(b) 

Top 5 Weekly crimes 2020

Fig. 6(c) 

Top 5 monthly crimes 2020
We also run an analysis on the arrests as a result.

Fig. 7(a)

Fig. 7(b)

Fig. 7(c)
Finally, our model predicts the type of crime depending upon the dataset when the required attributes are being input.

![Fig. 8(a)](image)

**IX. CONCLUSION AND FUTURE WORK**

We conclude that the system identifies and presents the characteristics of criminal cases, as well as a classification system for similar offences and a schema that ties the different elements together and discusses their interrelationships in order to recommend appropriate acts, interventions, and policies. The recognition of crime characteristics allows for a more detailed explanation of particular crimes, resulting in a clearer understanding, handling, and management of their occurrences.

Additional functionality may be added in the future because the approach towards incident description is flexible and feature-based. In addition, depending on particular viewpoints, they may expand or consolidate their respective elements.
This paper also introduced a two-level classification scheme and further research that took into account the current state of crime and new types of crime. The scheme covers the most popular types of offences which can be beneficial to law enforcement. The suggested scheme is also included as part of the feature-based definition for classifying a crime occurrence under a related criminal offence.

REFERENCES