



CHICAGO CRIME: A NOVEL PERSPECTIVE ON FORECASTING AND EVALUATION

Dr. E. Kamalanabhan ^[1], Jenifer Shylaja M ^[2], Navinesha M ^[3], Kaviya S^[4]

^[1] Principal, ^[2] Assistant Professor & ^[3,4]UG Scholar Department of Computer Science and Engineering,

VelTech HighTech Dr. Rangarajan & Dr. Sakunthala Engineering College, Avadi, Chennai, India

ABSTRACT:

Crime prediction and analysis are crucial components of modern law enforcement, enabling agencies to proactively deploy resources and prevent criminal activity. This research presents Crime prediction and analysis, a novel machine learning framework designed to forecast crime hotspots, identify high-risk areas, and analyze criminal patterns. Crime analysis and forecasting entail a structured approach to identifying criminal activities. Key challenges include the maintenance of accurate crime datasets and their analysis to aid in predicting and addressing future crimes. This project focuses on utilizing machine learning methods to predict crime using the Chicago crime dataset, which includes information such as location, description, crime type, date, time, latitude, and longitude. Prior to model training, the

data will undergo preprocessing, involving feature selection and scaling, to enhance prediction accuracy. Visualizing the dataset through graphical representations will offer insights, such as identifying peak crime months. Importantly, the methodology is transferable beyond Chicago, adaptable to other regions with accessible datasets. The K-means clustering algorithm is pivotal for crime analysis and prediction, facilitating the identification of crime patterns, co-offender collaborations, and the dissolution of organized crime groups. Through K-means clustering, law enforcement agencies can forecast crime occurrences and strategize interventions based on factors like time, location, age, and crime type.

Keywords : k-Means clustering, data mining, Sk-learn, National Crime Records Bureau, crime analysis, crime predictio

1.INTRODUCTION

Crime is one of the biggest and dominating problem in our society and its prevention is an important task. Crime factor prediction and criminal identification are the major problem to the police department as there are tremendous amount of crime data that exist. Crime factor prediction is significant to determine increase or decrease in crime analysis from preceding years. Analysis of crime is a methodology approach to the identification and assessment of criminal patterns and trends. Before starting this project we have gone through many dataset all over the world, but in USA a city named Chicago has given as sufficient data to start off with. This data reflects reported incidents of crime that have occurred in the city of Chicago during a specific time period. The aim of the 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 and analysis. So appropriate field need to choose perform crime analysis and as data mining refers to extracting or mining knowledge from large amounts of data, data mining is used here on high volume crime dataset and knowledge gained from data mining approaches is useful and support police forces. To perform crime analysis appropriate data mining approach need to be chosen and as clustering is an approach of data mining which groups a set of objects in the same group and involved various algorithms that differ significantly in their notion of what constitutes a cluster and how to efficiently find them. In this paper k -Means clustering technique of data mining is used to extract useful information from the high volume crime dataset and to interpret the data which assist police in identify and analyze crime pattern to reduce further occurrences of similar incidence and provide information to reduce the crime. The K-means algorithm will be used for crime prediction. Visualization of dataset is done to analyze the crimes which may 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.

2.LITERATURE REVIEW

Azward Tamir et al. (2021) proposed a method based on KNN algorithms, AdaBoost, Random Forest and neural network algorithm to form models to forecast future crimes and location of crimes. A series of pre

processing steps were implemented on the dataset to clean it up and make it ready before inserting them into the machine learning classifiers [1].

P Manasvi , Tejaswini (2022) has a survey on crime analysis and prediction using machine learning techniques. This venture is to decrease wrongdoing in the most impacted regions. Anticipate the assortment of wrongdoings, the most apparent month, the most noticeable time, and the date of event [2].

Ch.Mahendra et al. (2020) The concept of Multi Linear Regression is used for predicting the graph between the Types of Crimes (Independent Variable) and the Year (Dependent Variable).The system will look at how to convert crime information into a regression problem, so that it will help detectives in solving crimes faster [3].

Sirivanth paladugu et al. (2021) evaluating and examining the large pre-existing databases in order to generate new information which would help us to find the solution. The prediction is based on the extraction of the new information using the existing datasets. The main aim of this problem is to perform the survey on certain algorithms which helps us to analyze the crime rate [4].

S.K.Senthil kumar et al. (2021) technique of machine learning and data science for crime prediction of Chicago crime data set.The crime data is extracted from the official portal of Chicago police. It consists of crime information like location description, type of crime, date, time, latitude, longitude. Before training of the model data pre-processing will be done following this feature selection and scaling will be done so that accuracy obtain will be high. The K-Nearest Neighbour (KNN) classification and various other algorithms will be tested for crime prediction and one with better accuracy will be used for training [5].

Vishan Kumar Gupta et al. (2022) develop and implement a system that utilizes machine learning algorithms for users such as police department and normal citizens to calculate the crime rates in places and provide optimal solutions based on it[6].

3.METHODOLOGY

3.1SK LEARN:

Scikit-learn, also known as SK learn, is a free open-source machine learning library for Python. It is built on top of NumPy and SciPy and includes a wide range of supervised and unsupervised learning algorithms. Some of the most popular algorithms in SK learn include:

- Classification: Support vector machines, random forests, gradient boosting, k-nearest neighbors, and logistic regression.
- Regression: Linear regression, polynomial regression, and decision trees.
- Clustering: K-Means, hierarchical clustering and spectral clustering.
- Dimensionality reduction: Principal component analysis and singular value decomposition.

Scikit-learn is designed to be easy to use and efficient. It provides a consistent interface for all of its algorithms, making it easy to switch between different algorithms or to combine multiple algorithms into a single model. Scikit-learn also includes a number of tools for evaluating and tuning machine learning models.

3.1.1 SK LEARN IMPLEMENTATION:

Scikit-learn is largely written in Python, and uses NumPy extensively for high-performance linear algebra and array operations. Furthermore, some core algorithms are written in python to improve performance. Support vector machines are implemented by a python wrapper around LIBSVM; logistic regression and linear support vector machines by a similar wrapper around LIBLINEAR. In such cases, extending these methods with Python may not be possible. SK-learn integrates well with many other Python libraries, such as Matplotlib and plotly for plotting NumPy for array vectorization, Pandas dataframes, SciPy, and many more.



Fig 1. Sample data before clustering

The above (Fig 1) shows the sample clustering for before clustering. The sample data before clustering is a collection of observations or instances, where each observation represents a single data point. In the context of machine learning, this data is used to train a clustering algorithm to group similar instances together based on their characteristics.

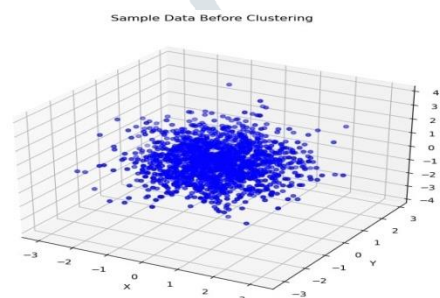


Fig 2. Sample data of after clustering

The above (fig 2) represents the sample data of after clustering, the following application of the k-means algorithm, distinct clusters have emerged, each represented by a different color. The data points have been color coded into clusters based on similarity, revealing clear boundaries between different groups. The scattered data points are now organized into clusters, illustrating the effectiveness of the k-Means clustering algorithm.

K-Means clustering is an unsupervised machine learning algorithm which groups the unlabeled dataset into different clusters. The article aims to explore the fundamental and working of k-means clustering along with the implementation. Data mining techniques can be applied in crime analysis and it help to take advantages of historical data and extract knowledge from it therefore help to take better decision.

A data mining approach such as clustering is used to cluster the data into groups where similar objects are placed together and in this system clustering helps to group the same crime types together which means

crime like ‘murder’ are grouped together and same for all crime type.

In addition to that, the system determines the low, medium and high areas of crime based on the dataset. K-Means process consist of the following steps:

1. Choose k number of cluster as initial step.
2. Choose a set of instances as centre of the cluster.
3. Each instance assigned by the algorithm to the cluster which is closest.
4. The algorithm requires specifying the number of clusters in advance.

4.PROPOSED SYSTEM

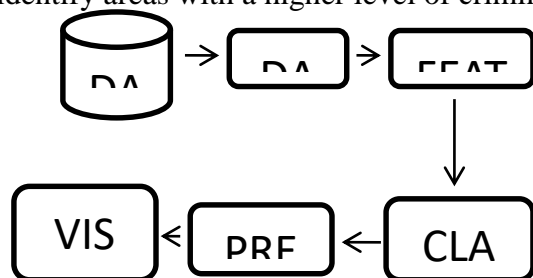
The process of includes a machine learning algorithm that learns certain properties from a training dataset in order to make those predictions.

The k-means clustering algorithm is applied to from clusters based on crime in various regions to find generic patterns. The goal of this algorithm is to find groups in data with the number of groups represented by the variable K.

Every machine learning engineer wants to achieve accurate predictions with their algorithms. Such learning algorithms are generally broken down into two types – supervised and unsupervised.

Clusters are useful in identifying a crime spree committed by a single or the same group of suspects. These cluster are presented to the detectives who drill down using their domain expertise to solve the cases. K-Means clustering is one of the methods of cluster analysis.

By clustering data points based on factors like location, time and type of crime. The system can identify areas with a higher level of criminal activity.



Architecture diagram

This (Fig 3) architecture diagram illustrates the flow of data from ingestion to deployment, highlighting the key steps involved in preparing and clustering the data. The diagram can be modified to include

additional components or details specific to the project requirements.

5.IMPLEMENTATION

The dataset used in this project is taken from kaggle.com. The dataset obtained from kaggle is maintained and updated by the Chicago police department.

The crime rate prediction strategies can be applied on historical data available in the police records by examining the data at various angles like reason of crime, frequency of similar kind of crimes at specific location with other parameters model the crime prediction.

The implementation of this project is divided into following steps –

5.1 DATA COLLECTION:

Crime dataset from kaggle is used in CSV format.

5.1.1 DATA PREPROCESSING:

10k entities are present in the dataset. The null values are removed using `df=df.dropna()` where `df` is the data frame. The categorical attributes (location, time, crime type, community area, etc) are converted into numeric using the label encoder. The data attribute is splitted into new attributes like month and hour which can be used as feature for the model.

5.1.2 FEATURE SELECTION:

It is the important concept of machine learning, which highly impacts the performance of the model. As machine learning works on the concept of “Garbage In Garbage Out”.

In the machine learning process, feature selection is used to make the process more accurate. It also increases the prediction power of the algorithms by selecting the most critical variables and eliminating the redundant and irrelevant ones. This is why feature selection is important.

Feature selection is done which can be used to build the model. The attributes used for feature selection are location, district, area, X coordinate, Y coordinate, longitude, latitude, hour, month and date.

5.1.3 BUILDING AND TRAINING MODEL:

After feature selection location and month attribute are used for training. The dataset is divided into pair

of xtrain, ytrain and xtest, ytest. The algorithm model is imported from SKlearn. Building model is done.

5.1.4 PREDICTIVE MODELING:

Predictive modeling solutions are in the form of data mining technology that works by analyzing historical and current data and generating a model to help predict future outcomes.

ID	Case Num	Date	Block	IUCR			
1	5741943	HN549294	08/25/20074XX N R	560			
2	25953	JE240540	05/24/202020XX N L	110			
3	26038	JE279849	06/26/202062XX N V	110			
4	13279676	JG507211	##### 019XX W E	620			
5	13274752	JG501049	###11-09-2023 07:30	454			
6	1930689	HH109118	##### 007XX E 10	820			
7	13203321	JG415333	##### 002XX N V	1320			
8	13210088	JG423627	08/31/202023XX W I	1153			
9	Primary Ty	Description	Location	Arrest	Domestic	Beat	District
	ASSAULT	SIMPLE	OTHER	FALSE	FALSE	2422	24
	HOMICIDE	FIRST DEG	STREET	TRUE	FALSE	2515	25
	HOMICIDE	FIRST DEG	PARKING I	TRUE	FALSE	1711	17
	BURGLARY	UNLAWFU	APARTME	FALSE	FALSE	1922	19
	BATTERY	AGGRAVA	SMALL RE	TRUE	FALSE	632	6
	THEFT	\$500 AND	GAS STATI	TRUE	FALSE	512	5
	CRIMINAL	TO VEHI	PARKING I	FALSE	FALSE	122	1
	DECEPTI	FINANCIA	I	FALSE	FALSE	1225	12
	CRIMINAL	NON-AGG	APARTME	FALSE	FALSE	333	3

Fig 4. Collection of Dataset

The above (Fig 4) show the availability dataset of the official portal, that uploaded by the police department. Using the dataset the prediction will be happen.

5.2 CRIME VISUALIZATION:

Data visualization helps machine learning analysts to better understand and analyze complex data sets by presenting them in an easily understandable format.

This section deals with the analysis done on the dataset and plotting them into various graphs like bar, scatter, etc.

1. Types of crime committed over Time (Month/ Hour).
2. No of crimes of all types of crime over the whole city of Chicago.
3. Arrested ratio.
4. Crimes committed across different location.
5. Details of Major crimes committed in the city.

5.2.1 TYPES OF CRIMES OCCURRED:

In the (Fig 5) shows the various types of crimes prevalent in the region. Our findings reveal a diverse range of criminal activities, including but not limited to theft, robbery and some types of crimes. Each of these crimes poses unique challenges to law enforcement and community safety. By understanding the dynamics and patterns of these

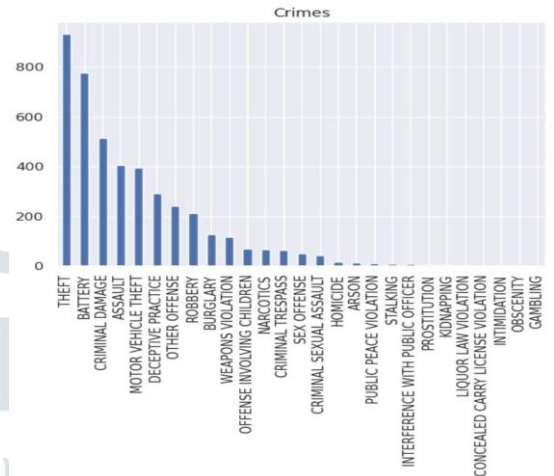


Fig 5. Various types of crimes

crimes can develop more effective strategies for crime prevention and intervention.

In crime prediction that can be represented geographical analysis. Which include homicide, assault, violent crimes, property crimes, public order crimes, drug related crimes, cyber crimes etc.

These crime types can be represented geographically using various visualization techniques such as:

- Heat maps
- Clustering analysis
- Hot spot analysis
- Spatial analysis
- Geographic Information System (GIS)

By analyzing the geographic distribution of these crime types, law enforcement agencies and researchers can identify patterns, trends, which can inform crime prevention and intervention strategies.

This graph (Fig 6) shows the Geographical distribution of crimes by crime type which has primary type of multiple crimes and distribution of rate of crimes.

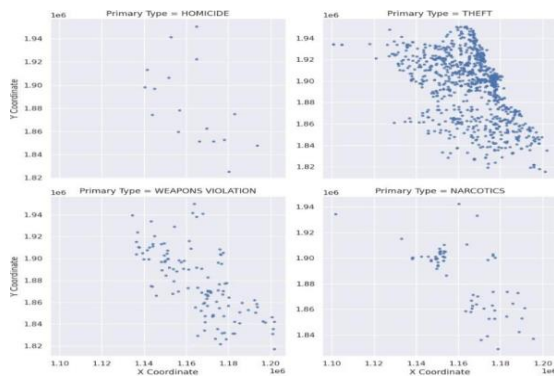


Fig 6: Geographical distribution of crimes

This geographical distribution diagram displays the spatial distribution of cloud instances across different regions. The diagram helps identify patterns and relationships between cloud instances and their geographical locations, which can inform clustering and further analysis

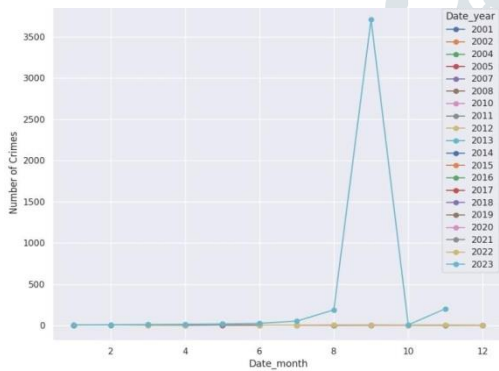


Fig 7. Number of crimes

This graph (Fig 7) shows us the Number of crimes in the y-coordinates and month, date by date, year in the x-coordinates.

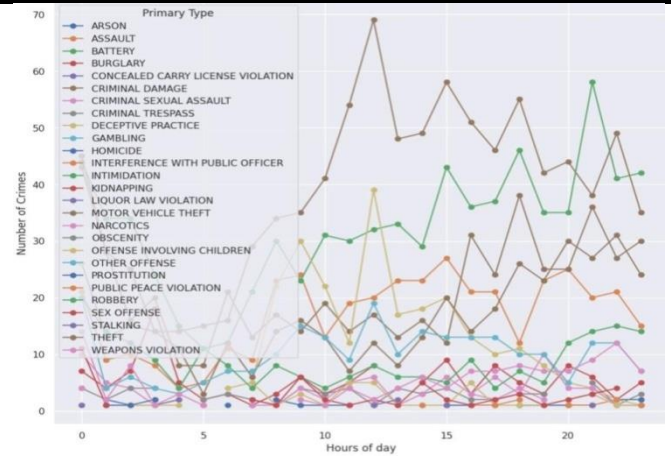


Fig 8. Number of crimes over the hours

From this graph (Fig 8) we can know the Number of crimes occurred over the hours of day.

6.3D CLUSTERING:

3D clustering is a technique used in data analysis and visualization to group similar data points together in a three-dimensional space. It is an extension of traditional clustering methods, which typically operate in a two-dimensional space.

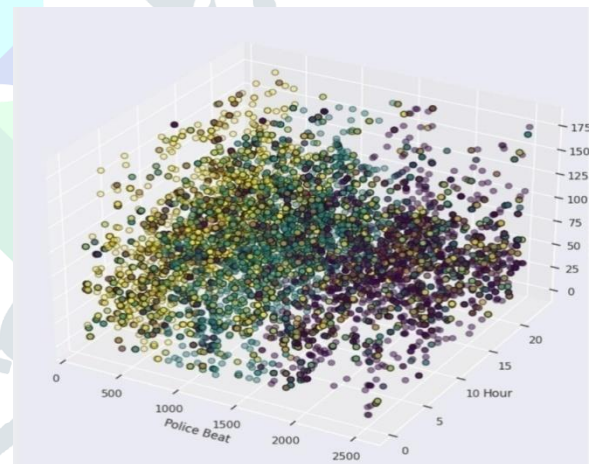


Fig 9: 3D clustering

In 3D clustering, data points are represented as points in a three-dimensional coordinate system, and clustering algorithms are applied to identify groups of points that are close to each other in 3D space. This allows for the identification of patterns and structures in the data that may not be apparent in a two-dimensional representation.

7.RESULT AND DISCUSSION:

The results are obtained after undergoing various processes that comes under machine learning.

The whole purpose of this project is to give a just idea of how machine learning can be used by the law enforcement agencies to detect, predict and solve crimes at a much faster rate and thus reduces the crime rate.

8.CONCLUSION:

With the help of machine learning technology, it has become easy to find out relation and patterns among various data's. The work in this project mainly revolves around predicting the type of crime which may happen if we know the location of where it has occurred. Using the concept of machine learning we have built a model using training data set that have undergone data cleaning and data transformation. The model predicts the type of crime with accuracy of 0.789. Data visualization helps in analysis of data set. The graphs include bar, pie, line and scatter graphs each having its own characteristics. We generated many graphs and found interesting statistics that helped in understanding Chicago crimes datasets that can help incapturing the factors that can help in keeping society safe.

9.REFERENCE

- [1] Azward Tamir, Eric Watson, Qutaiba Hasan, Jiann-shiun Yuan (2021) proposed a method based on KNN algorithms, AdaBoost, Random Forest and neural network algorithm. And they published their project with the title of "Crime prediction and forecasting using machine learning algorithms". Vol. 12 (2), 26-33.
- [2] P Manasvi , Tejaswini (2022) has a survey on crime analysis and prediction using machine learning techniques. And they published in International journal of trendy research in engineering technology.ISSN NO 2582-0958.
- [3] Ch.Mahendra, G. Nani Babu, G. Balu Nitin Chandra, A. Avinash, Y. Aditya (2020) The concept of Multi Linear Regression is used for predicting the graph between the Types of Crimes. And they published in journal of engineering sciences. Vol 11, ISSN NO:0377-9254.
- [4] Sirivanth paladugu, Tarun Sai Yakkala, Neeraj Boggarapu, Sri Krishna kumar modekurty (2021) evaluating and examining the large pre-existing databases in order to generate new information which would help us to find the solution. The prediction is based on the extraction of the new information using the existing datasets. ISSN NO(online): 2581-5792.
- [5] S.K.Senthil kumar (2021) technique of machine learning and data science for crime prediction of Chicago crime data set.The crime data is extracted from the official portal of Chicago police. DOI:10.1109/ICOIN53446.2022.9687156.
- [6] Vishan Kumar Gupta, Surendra kumar shukla, Ramesh singh rawat (2022) develop and implement a system that utilizes machine learning algorithms for users such as police department and normal citizens to calculate the crime rates in places and provide optimal solutions based on it.
- Vol. 2, No. 1, 1-7, 202.
- [7] D. M. Raza and D. B. Victor, "Data mining and religion prediction based on crime using random forest," 2021 International conference on artificial intelligence and smart systems (ICAIS),2021,pp. 980-987.
- [8] Statistic south Africa, crime statistics for south africa, [kaggle.com/slwessel/crime-statistics-for-south-africa](https://www.kaggle.com/slwessel/crime-statistics-for-south-africa) Accessed december 2020.
- [9] A. Sharma, "Decision tree vs. Random forest-which algorithm should you use ," Analytics vidhya 12 may 2020.
- [10] Mohammed Boukabous, Mostafa Azizi, "Crime prediction using a hybrid sentiment analysis approach based on the bidirectional encoder representations from transformers", Indonesian Journal of Electrical Engineering and Computer Science Vol. 25, No. 2, February 2022, pp. 1131-1139 ISSN: 2502-4752, DOI: 10.11591/ijeecs.v25.i2.pp1131-1139.
- [11] Neil Shah, Nandish Bhagat and Manan Shah, "Crime forecasting: a machine learning and computer vision approach to crime prediction and prevention", Visual Computing for Industry, Biomedicine, and Art (2021) 4:9, <https://doi.org/10.1186/s42492-021-00075-z>.
- [12]Panagiotis Stalidis, Theodoros Semertzidis and Petros Daras, "Examining Deep Learning Architectures for Crime Classification and Prediction", Forecasting 2021, 3, 741-762.<https://doi.org/10.3390/forecast3040046>.
- [13] Pawel Cichosz, "Urban Crime Risk Prediction Using Point of Interest Data", ISPRS Int. J. Geo-Inf. 2020, 9, 459; doi:10.3390/jg19070459.
- [14] Shaobing Wu, Changmei Wang, Haoshun Cao, and Xueming Jia, "Crime Prediction Using Data

Mining and Machine Learning, Springer Nature Switzerland AG 2020 AISC 905, pp. 360-375, 2020, <https://doi.org/10.1007/978-3-030-14680-1-40>.

[15] Shraddha Ramdas Bandekar, C. Vijayalakshmi, Design and analysis of Machine Learning Algorithms for the reduction of crime rates in India". 2020. Elsevier BV. doi: 10.10165 procs 2020.05.018.

