Introduction:

Crime is one of the concerning aspect of the society. Crimes affect our society in different ways. Crime investigation plays an important role in police system in the country. Criminal analysis and investigation is the process to explore and detect crime and criminals relationship .

•There are lots of data related to the crime in police station records, information related to the particular crime or the essential information which is directly or indirectly related to crime should be extracted. So there is need of such technology, which separate all these data from huge content. On the basis of previously known (historical) crime and criminals relationship record, the criminal investigation team can extract useful information so that they can identify the facts related to the committed crime and minimize the future crime possibilities .

•In the current era, number of crimes occurs in the society and this criminal rate increase day by day. There is tremendous growth of criminal data. Crime has negatively influenced the societies. Crime control is essential for the welfare, stability and development of society. Law enforcement agencies are seeking for the system to target crime structure efficiently. The intelligent crime data analysis provides the best understanding of the dynamics of unlawful activities, discovering patterns of criminal behavior that will be useful to understand where, when and why crimes can occur. There is a need for the advancements in the data storage collection, analysis and algorithm that can handle data and yield high accuracy.

Data Mining is the procedure which includes evaluating and examining large pre-existing databases in order to generate new information which may be essential to the organization. The extraction of new information is predicted using the existing datasets. Many approaches for analysis and prediction in data mining had been performed. But, many few efforts has made in the criminology field. Many few have taken efforts for comparing the information all these approaches produce. The police stations and other similar criminal justice agencies hold many large databases of information which can be used to predict or analyze the criminal movements and criminal activity involvement in the society. The criminals can also be predicted based on the crime data.

BIG DATA ANALYSIS:

1.Big data analytics is the process to examine the huge amount of data to find hidden information patterns and trends.
2.It helps in cost reduction, faster and better decision making. It provides a framework for storing and analyzing huge amounts of unstructured criminal data in real time.
3.An analytical system can cope up with predicting crimes. Investigation analytics system can deal with many information like text data, audio, video, DNA.
4.Combining with security intelligence sources, it provides the information about the latest vulnerabilities and identify outliers and anomalies in security data.
5. Big data security analytics can minimize flows of raw security events to a manageable number of alerts. It provides details to the investigator about the incident and its relationship with historical anomalies.
6.Big data analytics can be used for analyzing the financial transaction, log files to identify suspicious activities.

Crime Analysis procedure

The criminals when leaving the crime scene does leave some traces which can be used as a clue to identify the criminals.

The crime sequence and the patterns which several criminals follow when committing a crime make it easy for analyzing the crime.

This process includes several procedures to be followed in order to identify the criminals and getting more information based only on the clues or information given by the local people.

The criminal can be analyzed based on the information from the crime scene which is tested against the previous crime patterns and judging by the method which is implied to test and proceed with the information that can affect the prediction results.

The prediction can be further made useful for detecting the crimes in advance or by adding more cops to the sensitive areas which are identified by the system .

The police stations can put up special force when there are chances for crime ahead of time. This type of the system will ensure there are peace and prosperity among the citizens.

The crime analysis can be performed procedure which is similar to figure Fig.1 which specifies each module which is used for machine learning to predict the crime or form group of clusters of criminals according to crime records.

The criminals can hold certain properties and their crime characteristics and crime careers may vary from one criminal to another.

Such a type of information can be taken as the input dataset.

The input dataset is given to a preprocessor which performs the preprocessing based on the requirements.

Once the pre processing is completed the features or attributes from those information are extracted which may be in the form of text content from emails, the crime factors for a day, criminal characteristics, geo-location of the criminal, etc,.

The pre processed result is further given to the classification algorithm or the clustering algorithm based on the requirements. The requirements may be anything from selecting the crime prone areas to predicting the criminal based on the previous crime records.

The classification algorithm works in a supervised learning manner in which the training and testing phase is required in order to train the classifier to identify the new unknown crime record. This is known as prediction.

Whereas the clustering algorithm works in an un-supervised learning manner which automatically separates the crime records based on the number of groups to be created. The groups created in such a manner are known as clusters. Such a type of design can be a general template for applying crime prediction and crime analysis based on data mining algorithms.

Criminals analysis methods:

1. **TEXT, CONTENT AND NLP-BASED METHODS**
2. **CRIME PATTERNS AND EVIDENCE-BASED METHODS**
3. **SPATIAL AND GEO-LOCATION BASED METHODS**
4. **PRISONER BASED METHODS**
5. **COMMUNICATION BASED METHODS**

1)TEXT, CONTENT AND NLP-BASED METHODS

Sharma [1] proposed a concept which depicts zero crime in the society. For detecting the suspicious criminal activities, he has concentrated on the importance of data mining technology and designed a proactive application for that purpose.

In his paper, he proposed a tool which applies an enhanced Decision Tree Algorithm to detect the suspicious e-mails about the criminal activities.

An improved ID3 Algorithm with an enhanced feature selection method and attribute-importance factor is applied to produce a better and faster Decision Tree based on the information entropy which is explicitly derived from a series of training data sets from several classes.

He proposed a new algorithm which is a combination of Advanced ID3 classification algorithm and enhanced feature selection method for the better efficiency of the algorithm.

Hamdy et al. [8] described an approach based on the people's interaction with social networks and mobile usage such as location markers and call logs.

Their work also introduced a model for detecting suspicious behavior based on social network feeds and it not only describes a new method using the social interaction of people but, their work proposes a new system to help crime analysis create faster and precise decisions.

The suspicious movement of the entity can be determined using the sequence of inference rules. Their constructed model is able to predict and characterize human behavior from reality data sources.

Method	Input	Dataset Used	Pre-Processing	Feature Extraction	Classification Clustering	Strength	Weakness	Outcome

2)CRIME PATTERNS AND EVIDENCE-BASED METHODS

Bogahawatte and Adikari [2] proposed an approach in which they highlighted the usage of data mining techniques, clustering and classification for effective investigation of crimes and criminal identification by developing a system named Intelligent Crime Investigation System (ICSIS) that could identify a criminal based up on the evidence collected from the crime location. They used clustering to identify the crime patterns which are used to commit crimes knowing the fact that each crime has certain patterns. The database is trained with a supervised

learning algorithm, Naïve Bayes to predict possible suspects from the criminal records. His approach includes developing a multi-agent for crime pattern identification. There are agents for the place, time, role trademark and substance of criminals which separates the role of the criminals in components. The system is a multi-agent system and made with managed Java Beans. It makes it easy to encapsulate the requested entities in the work into objects and returns it to the bean for exposing properties. Classifying the criminals/ suspects is based on the Naïve Bayes classifier for identifying most possible suspects from crime data. Clustering the criminals is based on the model to help to identify patterns of committing crimes.

Agarwal et al. [3] used the rapid miner tool for analyzing the crime rates and anticipation of crime rate using different data mining techniques. Their work done is for crime analysis using the K-Means Clustering algorithm. The main objective of their crime analysis work is to extract the crime patterns, predict the crime based on the spatial distribution of existing data and detection of crime. Their analysis includes the tracking homicide crime rates from one year to the next

Kiani et al. [4] performed a crime analysis work based on the clustering and classification techniques. Their work includes the extraction of crime patterns by crime analysis based on available criminal information, prediction of crimes based on the spatial distribution of existing data and crime recognition. They proposed a model in which the analysis and prediction of crimes are done through the optimization of outlier detection operator parameters which is performed through the Genetic Algorithm. The features are weighted in this model and the low-value features were deleted through selecting a suitable threshold. After which the clusters are clustered by the k-means clustering algorithm for classification of crime dataset.

Satyadevan et al. [5] has done a work which will display high probability for crime occurrence and can visualize crime prone areas. Instead of just focusing on the crime occurrences, they are focusing mainly on the crime factors of each day. They used the Naïve Bayes, Logistic Regression and SVM classifiers for classification of crime patterns and crime factors of each day. Their method consists of a pattern identification phase which can identify the trends and patterns in crime using the Apriori Algorithm. The prediction of crime spots is done with the help of Decision Tree algorithm which will detect the crime possible areas and their patterns.

Bruin et al. [7] proposed a technique which is used to determine the clustering of criminals based on the criminal careers. The criminal profile per offense per year is extracted from the database and a profile distance is calculated. After that, the distance matrix in profile per year is created. The distance matrix including the frequency value is made to form clusters by using naïve clustering algorithm. They made a criminal profile which is established in a way of representing the crime profile of an offender for a single year. With this information, the large group of criminals is easily analyzed and they predicted the future behavior of individual suspects. It will be useful for establishing the clear picture on different existing types of criminal careers They tested the tool on actual Dutch National Criminal Record Database for extracting the factors for identifying the criminal careers of a person.

3)SPATIAL AND GEO-LOCATION BASED METHODS

Huang et al. [6] focused on a different approach for criminal activity prediction based on mining location based Social Network interactions. By using these interactions, they can collect information using the geographical interactions and data collections from the people. They devised a working procedure in which a series of features are categorized from the Foursquare and Gowalla used in the San Francisco Bay area. The crime patterns and the crime occurrences are tracked with the geographical features which are extracted from the map and they are analyzed to detect the urban areas with high crime activities. Their work aims at exploiting the location-based social network data to investigate the criminal activities in urban areas. By using the Haversine formula the distance between the two points i.e. the crime location and venue location is calculated and shown in the Google Maps API and OpenStreetMap.

4)PRISONER BASED METHODS

Sheehy et al. [10] came up with a research idea which was geared towards the treatment of the mentally ill people inside the prison. According to their work, the mentally ill criminals are identified using their Social Security Number (SSN) with all the criminal personal records and their crime career records attached. As the outcome, the Criminals are classified into "high", "medium" and "low" levels of recidivism risk potential according to their mental health. Their objective was to describe and classify the criminals into a misdemeanor and a felony which can be referred and not referred based on the mental health of the criminals. Their ill activities are monitored and data collection is continuous. By these, the criminals can be separated from other criminals who are hazardous and those who can cause damage to other inmates along with them. Further, their study also involves the classification of the mental health of the criminals into two categories i.e. "referred" and "not-referred". This helps the guards to identify the prisoners who are referred for the mental health check-up. The research work they had undergone will provide a summary of the inmates who are seriously mentally ill and those who are to be separated from the other inmates.

5)COMMUNICATION BASED METHODS

Taha et al. [9] has developed a forensic investigation tool for identifying the influential members who create an impact in a criminal organization. The immediate leaders can also be identified in a criminal organization. Removing these influential members can weaken the strength of the criminal organization. Their work is based on this methodology. They proposed a new work which is known as SIIMCO which first constructs the graph representing the criminal group or organization as a network from either mobile communication data of the criminal organization or based on the crime records. The system works on the basis of the created networks. These networks represent the criminal organization or crime incident reports. The vertex represents the individual criminals and the link represents the relationships or communication link between those two criminals. They employed certain formulas that quantify the degree of influence/ importance of each vertex in the network relative to all other vertices i.e. criminals in the graph. Based on this their system identifies the immediate leaders with the weighted graph which connects the criminals and identify them for further processing

Crime Analysis

Recently computer has a significant role in every field, also in tracing criminals and tracking their crimes. Since the criminals can have specific properties and may have different crime characteristics and crime careers. Law enforcement agencies take advantage of data analysis techniques to improve Criminology. It refers to the process used for identifying crimes and criminals' characteristics, and to detects the patterns of previous committed crimes in order to predict.

the main aims of crime analysis are:

- Applying crime analysis to find crime patterns based on real criminal information.
- Prediction of crimes and their occurrence using various data mining techniques.
- Crime recognition.

Classification

Classification is a supervised machine learning technique that categorizes data into groups (classes) and is used for making the future predictions. In other words, classification is used to predict the classes of a given data based on certain attributes known as predictors

They have used the following classification methods:

- Naive Bayes classification method to create a model by training crime data related to vandalism, murder, robbery, burglary, sex abuse, gang rape, arson, armed robbery, highway robbery, snatching, etc.
- Apriori algorithm to identify crime patterns that occur frequently.
- Decision Tree to predict the crimes areas. The techniques used are easy to apply and interpret, but they can only predict the crime spot without predicting the crime occurrence time.

Clustering

Clustering is an unsupervised data mining technique used to split a group of items and data into clusters based on certain characteristics, each cluster contains a group of similar data. Clustering is used in undefined and unfixed classes and without supervision when it comes to grouping the objects.

R. et al. [1] have used diverse clustering methods to analyze crimes in the dataset State Crime Records Bureau (SCRB) of Tamilnadu, India. The data has been obtained which has approximately 38 several cities and districts crime data.

The clustering method that has been used for the analysis was DBSCAN and K-means. To resolve the efficiency of the two clustering approaches, they used Silhouette coefficient. The

study concluded DBSCAN clustering with a high accuracy and resulted in more accurate clusters.

P. Vrushali et al. [16] focused on the frequency rate during diverse years to classify clustered crimes and the study was conducted on real data which offers an innovative framework for clustering and predicting crimes.

To analyze the crime data, in addition, to categorize crimes by grouping the related patterns, several clustering methods were applied.

In conclusion, several crime categorization and prediction methods were discussed that were applied to different crimes datasets. All the discussed papers have applied more than one machine learning technique to achieve the expected results.

Therefore, we will need to apply many techniques to our dataset to analyse the data behavior, in order to create an efficient module to predict the crime category with the optimal accuracy

. **METHODOLOGY**

Description of the Proposed Techniques

This section presents the description of the machine learning techniques that is used in our study to predict the crime category based on behavioral data. In our study we used supervised learning approach and chose Naïve Bayes and Decision Tree algorithms, which fall under the classification technique.

Naïve Bayes Technique

Naïve Bayes is a simple supervised classification algorithm. It uses the concept of Bayes Theorem, which classifies tuples to a class label related to a dataset based on the calculations of the conditional probability for each class label and attribute.

The Naïve Bayes algorithm works by the given probability rule.

$$P(H \mid X) = \frac{P(X \mid H)\, P(H)}{P(X)}$$

Where X is a data sample of dimensions $x = < x1, x2, x3 \ldots . xn >$ with an unknown class label, and X belongs to H hypothesis class label. Where $P(H)$ is the prior probability (initial probability) associated with hypothesis H [17].

Decision Tree Technique (J48)

Decision Tree is a supervised learning technique, which recursively partitions the instance space. Its main use is to predict the class labels.

The Decision Tree is composed of internal nodes, which represents set of predictors (attributes), edges, which represent a specific value or range of values of the input predictors (attributes), and leaf nodes, which represent the class labels.

The internal nodes along with their edges split the instance space into two or more partitions and each terminal node (leaf node) of the tree is a class label [18].

J48 classifier is a simple C4.5 Decision Tree for classification. In this technique, a binary tree is constructed to model the classification process. Once the tree is built, it is applied to each tuple in the database and results in a classification for that tuple.

J48 splits the data into range based on the attribute (predictors) values. J48 allows classification by Decision Tree or the rules generated from them [17].

The Basic Steps in J48 classifier:

- For the instances of the same class, the tree represents a leaf labeled with the class.
- By applying a test on each attribute, the potential information is calculated for all of them.
- Then, a test is applying on the attribute to calculate the information gain.
- Then, based on the selection criterion, the best attribute is selected and used for branching. The "Entropy" is used in this process, which is a measure of the data disorder. The Entropy of $\vec{y}$ is calculated by

$$Entropy\ (\vec{y}) = -\sum_{j=1}^{n} \frac{|y_i|}{|\vec{y}|}\ log\ (\frac{|y_i|}{|\vec{y}|})$$

$$Entropy\ (j|\vec{y}) = \frac{|y_i|}{|\vec{y}|}\ log\ (\frac{|y_j|}{|\vec{y}|})$$

And Gain is

$$Gain\ (\vec{y}, j) = Entropy\ (\vec{y} - Entropy\ (j|\vec{y}))$$

And Gain is $Gain\ (\vec{y}, j) = Entropy\ (\vec{y} - Entropy\ (j|\vec{y}))$

The objective is to maximize the Gain, dividing by overall entropy due to the split argument by value j [19].

5. EMPIRICAL STUDIES

This section describes the dataset, statistical analysis of the dataset, and experimental setup.

5.1 Description of Dataset

The selected dataset reflects reported incidents that occurred in the City of Chicago from 2013 to 2017. Data is extracted from the Chicago Police Department's Citizen Law Enforcement Analysis and Reporting (CLEAR) system [20]. The dataset contained 12109 records and 18 attributes/features, including the target. The attributes contain nominal and numeric data. The techniques used in our study are Naïve Bayes and Decision Tree.

5.2 Experimental Setup

The experiment was conducted using Weka. Weka was used for data preprocessing and applying machine learning techniques. Moreover, the dataset was split into 70% for the training data, and 30% for the testing data to investigate the performance of the proposed technique. For data preprocessing, multiple preprocessing steps have been applied to the dataset, which include:

5.2.1. Data Cleaning

The dataset was pruned from the missing values. Since some the data are missing completely at random, listwise deletion (deleting rows) method was applied. Also, we manually reduced the dataset records to achieve a balanced distribution of the class labels.

5.2.2. Feature Selection

We have reduced the features from 22 to 10, by dropping the features, which describe the crime itself, such as case number, arrest, and description of the crime, since they only known after the occurrence of the crime and keeping the spatial features since the classification model relies on them. The selected features were: Block, Location Description, Domestic, Beat, District Ward, Community Area, X Coordinates, Y Coordinates Primary Type is the target variable. Since the number of the classification labels in the original dataset was too high, and several crime categories can be labeled under four main categories, so we reduced them as follow for better prediction.

- **Forbidden practices:** Narcotics, prostitution, gambling, obscenity, and other narcotic violation.

- **Theft:** Burglary, deceptive practice, motor vehicle theft, and robbery.

- **Assault**: Crime sexual assault, offense involving children, sex offense, homicide, and human trafficking.

- **Public peace violation:** Weapons violation, criminal defacement, criminal trespass, arson, kidnapping, stalking, intimidation, public indecency

RESULTS AND DISCUSSION

This section demonstrates the applied feature selection and shows the results of the proposed techniques.

6.1. Results of Investigating the Effect of Feature Selection on the Dataset

In feature selection, the features were selected by applying backward feature selection. Initially, the number of features was ten. They were ranked based on Information Gain Attribute Evaluation algorithm, then the attribute with the least rank was dropped until left with 9 attributes since the accuracy began to be reduced. Table 1 shows the accuracies of features subset for both Naïve Bayes and Decision Tree.

Using features	Na ïve Bayes	Decision Tree
All	83.40 %	91.68 %
Top 9 features	83.34 %	91.58 %
Top 8 features	82.82 %	90.71 %
Top 7 features	82.60 %	88.66%

As shown in the table above, Decision tree classifier gives better accuracy than Naive Bayes with 9 features and same accuracy was achieved with all the features. Therefore 9 features were selected since it consumes fewer resources. For Naïve Bayes highest accuracy was achieved with all the features.

Table 2 summarizes the results for Correctly Classified Instances (CC), Accuracy (AC), Receiver Operating Characteristic Curve (ROC), Precision, and Recall for Naïve Bayes and Decision Tree.

	Na ïve Bayes	Decision Tree
CC	7063	7763
AC	83.33 %	91.59 %
ROC	0.969	0.976
Precision	0.851	0.926
Recall	0.833	0.916

In Decision Tree:

o 2033 instances from class “Theft” were classified correctly and 70 instances were classified incorrectly.

o All the instances belong to class “Public Peace Violation” were classified correctly.

o 1936 instances from class “Forbidden Practices” were classified correctly and 160 instances were classified incorrectly.

o 1584 instances from class “Assault” were classified correctly and 557 instances were classified incorrectly.

- In Naïve Bayes:

o 1949 instances from class “Theft” were classified correctly and 145 instances were classified incorrectly.

o 1688 instances from class “Public Peace Violation” were classified correctly and 448 instances were classified incorrectly.

o 2024 instances from class “Forbidden Practices” were classified correctly and 72 instances were classified incorrectly.

o 1402 instances from class “Assault” were classified correctly and 739 instances were classified incorrectly.

Conclusion:

This study aimed to predict the category of crimes using “City of Chicago from 2013 to 2017” dataset. The techniques that were used in our study were Naïve Bayes and Decision Tree. In the preprocessing stage, missing values were handled using listwise deletion. Feature Selection was performed using backward feature selection technique. Decision Tree classifier performance was better than Naïve Bayes, its prediction accuracy reached 91.59 %. However, Naive Bayes accuracy is 83.40 %. Therefore, comparing the two results, Decision Tree outperformed Naïve Bayes