An Experimental Approach for Evaluating the Performance of Clustering Techniques for Crime Detection

Mrs. Sukhvinder Kaur Walia[1] , Dr. Shraddha Masih[2], Dr. Ugrasen Suman[3]

[1]Research Scholar, DAVV, Indore, waliasukhvinder2008@gmail.com
[2] Professor, SCSIT, DAVV, Indore, smasih.scsit@gmail.com
[3]Professor & Head, SCSIT, DAVV, Indore, ugrasen123@yahoo.com

Abstract: In the 21st century, when technology connects the entire world, our society suffers from modern age crime and criminals. A big challenge with the law enforcement agencies is to find and detect crimes. In machine learning, clustering algorithm is used to analyze crime related textual data focusing on identifying crime patterns from crime text data. In this research three clustering methods such as K-means, Agglomerative and DBSCAN are applied to crime reports to group similar description into clusters. These clusters were then analyzed to reveal the underlying patterns such as common crime type, crime time and crime location. The performance of these clustering techniques was evaluated using clustering assessing metrics Silhouette Score, ARI, homogeneity and completeness. It was observed that Silhouette score of K-Means, Agglomerative and DBSCAN were recorded 97%, 2.89% and 0.02%. The Adjust Rand Index of the three was 100%, 0% and 32.65%.The Homogeneity and completeness were 100%, 53.88%, 0.0% and 100%, 50.73%, 100%. The scores shows that K-means gives the better results in all the four metrics as compare to agglomerative and DBSCAN algorithm.

Keywords: K-means, DBSCAN, Agglomerative, clustering, machine learning.

1. Introduction:

Community safety is a primary concern for developed nations. Consequently, governments implement the necessary measures to mitigate crime rates. This, in turn, facilitates economic growth and enhances the quality of life. Crime analysis is a crucial component of criminology that focuses on examining behavioral patterns and aims to identify indicators of such phenomena [1]. However, numerous complexities have emerged in the implementation of crime prevention strategies. This is due to the diversity of criminal activities, motivations, consequences, management approaches, and preventive measures. Owing to these complexities and multifaceted characteristics, crime prediction has evolved into a potent and extensively utilized methodology. Consequently, law enforcement agencies allocate substantial time and resources for the detection and forecasting of criminal trends.

In light of increasing technological advancements and progress in artificial intelligence (AI), Machine Learning (ML) methodologies can mitigate this effort by efficiently analyzing substantial volumes of data to discern crime patterns[2]. Various artificial intelligence techniques have been extensively investigated to prevent criminal activity and enhance public safety across different nations. In the future, these machine learning models may be utilized for predicting potential criminal incidents and their associated characteristics.

Machine learning algorithms will enable law enforcement agencies to optimize resource allocation by identifying high-risk areas based on temporal, typological, or other relevant factors[3]. In addition, the investigation of past criminal behavior may provide more insights into the social composition of localities. One way to solve crime puzzles is through police reporting systems. The police reporting system contains a first investigation report

(FIR), witness report, legal documents, crime scene records, and weapon details. analyzing unstructured crime text data presents various clues owing to the complexity and variability in this type of data. In this regard, ML and NLP play a vital role in solving crime puzzles and predicting important hints for detecting crime patterns in text data.

2. Problem Statement:

The rapid increase in crime-related textual data, such as police reports, news articles, and incident descriptions, presents a significant challenge for law enforcement and policy makers. These large volumes of unstructured data contain valuable information that can help identify crime patterns, trends, and areas requiring intervention. However, manually analyzing this data is both time-consuming and resource-intensive. Traditional methods of crime analysis often rely on predefined categories and human judgment, which may overlook subtle or complex patterns in the data. Furthermore, the diversity and inconsistency in crime descriptions make it difficult to categorize and extract actionable insights.

This research seeks to address these challenges by exploring the use of **unsupervised clustering algorithms** to automatically identify patterns in crime text data. By grouping similar crime reports together, these algorithms can reveal underlying trends, such as crime types, geographical hotspots, and temporal patterns, without the need for predefined labels. The problem lies in selecting the most effective clustering method that can accurately capture these patterns while handling the inherent noise and scarcity in text data. Therefore, this study aims to evaluate and compare various clustering techniques—such as **K-Means**, **DBSCAN**, and **Agglomerative Clustering**—to determine which method best identifies meaningful crime patterns and trends in large-scale crime text datasets.

The study focused in finding crime patterns from crime text data. After pre-processing, NER and clustering techniques were used to detect crime patterns. Feature selection and extraction were used to extract prominent features that help to improve crime-detection rates. Clustering techniques such as K-Means, DBSCAN and BIRCH were used to identify the patterns. The main objective of this study is to evaluate the performance of three clustering techniques on crime textual data using Python.

The remainder of this paper is organized as follows: Section II describes the related work on the various ML and NLP techniques used for crime text analysis. Section III demonstrates the methodology used for the performance evaluation of the clustering algorithms, and Section IV presents the experiments and results showing and comparing the accuracy of each clustering method, thus representing the best clustering technique for extracting crime patterns from crime text data. Lastly section V gives conclusion and future work

3. Related Work:

This section examines the most recent developments and applications of text mining, natural language processing (NLP), and machine-learning (ML) approaches in crime text. These methods use various types of crime data, including witness narrative reports, police reports, and online content. Numerous studies have been conducted on crime detection and prediction. proposed a new clustering method to predict crime patterns[3]. The dataset used was newspaper crime news. In this study, they proposed a technique to extract the relations between crime news articles. In [4] proposed a method to predict crime patterns based on crime-committed locations. In this research, robbery data from the UK were used for crime analysis. Three fundamental criminal patterns in four graph theory combinations were used in [5] to detect crime based on crime location and crime time. In [6] used the K-means clustering technique to detect crime patterns. In this study a system was proposed to extract crime patterns based

on partitioning clustering. Chen and Kurland proposed a method to identify the crime pattern on basis of Time, Place and Modus Operandi [7]. In this study, the Apriori algorithm was used to perform the crime analysis. In [8] author used a decision tree algorithm to detect suspicious criminal activities. A tool called Z-CRIME, which shows high accuracy in terms of precision and recall, was proposed in this study. Wang et al. (2013) proposed a pattern detection algorithm called Series Finder to build crime patterns. The datasets used in this study were collected from the Crime Analysis Unit of the Cambridge Police Department. The algorithm combines the features of individual crimes and yields encouraging results.

Pre-processing and named entity identification are the two primary phases of the suggested rule-based Arabic named entity recognition (NER) system for Arabian crime texts[9]. While named entity identification uses grammatical rules and patterns, preprocessing entails sentence splitting, tokenization, and part-of-speech tagging. The efficacy of the system in recognizing and classifying named items in Arabic crime-related texts was demonstrated by its 91% precision and 89% recall when tested on a corpus of Arabic criminal newspaper documents.

Based on a literature review, it is evident that text clustering is a technique employed in natural language processing and machine learning to categorize similar documents or textual elements according to their content. This process entails the analysis of textual data to identify patterns and similarities, followed by the organization of texts into clusters, wherein items within each cluster exhibit greater similarity to one another than to those in other clusters. Unsupervised learning techniques are capable of efficiently organizing and analyzing massive document collections without predetermined categories, thereby allowing the detection of hidden patterns and enhancing information retrieval.

In this study, we utilized three clustering techniques for crime text: K-means, agglomerative clustering, and density-based spatial clustering of applications with noise (DBSCAN). The research process typically encompasses steps such as text preprocessing (tokenization, stop word removal, and stemming), feature extraction (frequently employing techniques such as TF-IDF), and pattern generation using a clustering algorithm. The main objective of this study was to evaluate the performance of the three clustering techniques based on crime text data. The evaluative measures used were cluster quality, silhouette score, ARI, homogeneity and completeness.

4. Research Methodology

The methodology used for the performance analysis of clustering techniques on crime text data involves four basic stages: data collection, data cleaning, data reduction and feature extraction, implementation of clustering techniques on crime text datasets, and measurement of the accuracy of each technique using the Silhouette Score, cluster purity, and completeness. Crime-related text data, such as police reports, witness narrations, and legal court documents, were used as a dataset for this research. The crime-text preprocessing algorithm proposed in [10] was used for preprocessing because crime data are highly complex and noisy.

4.1 Crime text preprocessing

Legal court documents that are highly unstructured and contain complex noisy raw data require cleaning for optimal criminal analysis. Crime text preprocessing converts unstructured data into an organized, analyzable, structured format, thereby enabling efficient information extraction and interpretation. This process involves techniques such as tokenization, parsing, and normalization to break down the raw text into manageable units and standardize their representation. These methodologies create a structured framework for crime analysis, including pattern recognition, to extract meaningful information and entity extraction.

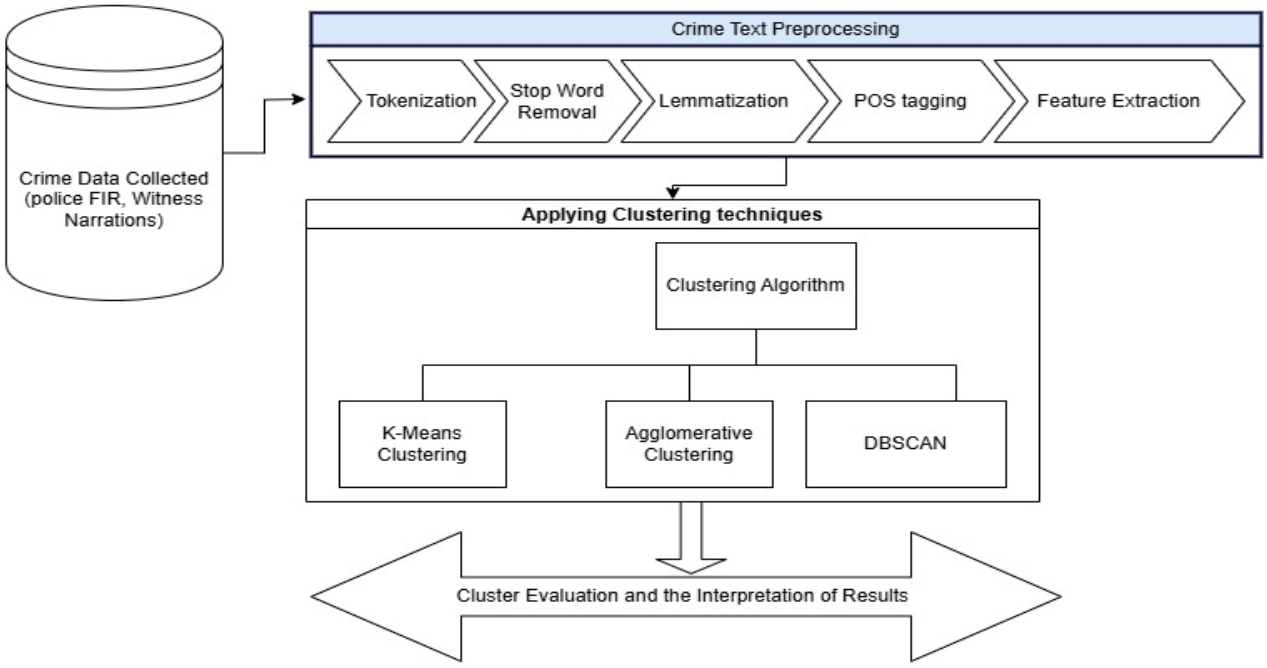


Fig 1: Methodology used for Crime Detection

5. Clustering algorithm

The implementation of each clustering algorithm was carried out with attention paid to parameter tuning and optimization for text data clustering. The algorithm is then applied to the preprocessed text data, iterating until convergence by minimizing the distance between the data points and their respective cluster centroids. Hierarchical Clustering is performed using agglomerative methods, where a dendrogram helps to visualize how clusters form at various levels of similarity, allowing for flexibility in deciding the number of clusters. In the context of Density-Based Spatial Clustering of Applications with Noise (DBSCAN), epsilon (radius) and minimum sample parameters are used to successfully identify clusters with diverse shapes and densities. This approach is particularly valuable for detecting outliers in the textual data.

5.1. K-Means clustering

K-means clustering is an unsupervised machine-learning technique employed to divide a dataset into a set of distinct groups called clusters. It measures the mean-squared distance between the data points and the center of the clusters [11].

Steps in K-Means clustering

Step 1: Initialize the values of the cluster centroids and set the number of clusters.
Step 2: Each data point was allocated to the closest cluster centroid.
Step 3: Update the centroids by calculating the mean of all the assigned data points inside a cluster.
Step 4: Repeat Step 3 until the calculated centroid matches the previous centroid.

5.2 Agglomerative Clustering

Agglomerative clustering is a type of hierarchical clustering that groups data points based on their similarities. Every data point begins as an individual cluster and follows a bottom–up approach. This clustering technique

represents the clusters in the form of a dendrogram, where each data element starts as its own cluster and gradually combines the closest and nearest data points until all the data points combine with each other and form a target cluster[11].

Steps for Agglomerative Clustering

Step-1: *Each data point was treated as an individual cluster.*

Step-2: *Calculate the cluster distance, find the two most similar clusters, and merge them into one cluster.*

Step-3: *Continue finding and merging the next closest pair of clusters.*

Step-4: *This process continues until all data points are merged into a single cluster or a specific number of clusters is reached.*

5.3 DBSCAN

Density-Based Spatial Clustering of Applications with Noise is an algorithm in unsupervised machine learning that detects clusters of various forms in datasets by analyzing density patterns. This method does not require prior knowledge of the number of clusters and can identify groups of different shape. The algorithm is based on two main parameters such as **Epsilon (ε),** the maximum distance within which points are considered neighbors, and **MinPts,** the minimum number of points required to form a dense region[13].

Steps for DBSCAN

Step-1:Define two parameters , Epsilon (ε) and MinPts.

Step-2: Classify Each Point for each point in the dataset, count how many other points lie within the ε distance, Core point, Border Point and noise

Step-3: Starting with an unvisited core point, assign it to a new cluster. All the directly reachable points within the ε distance of this core point were added to the cluster.

Step-4: Repeat step -3 until no more points can be added to the cluster.

Step-5: Any points that are not assigned to any cluster at the end are marked as noise.

Step-6: Repeat until all points are visited and continue the process for each unvisited core point in the dataset, forming clusters around them.

6. Experimental and Results

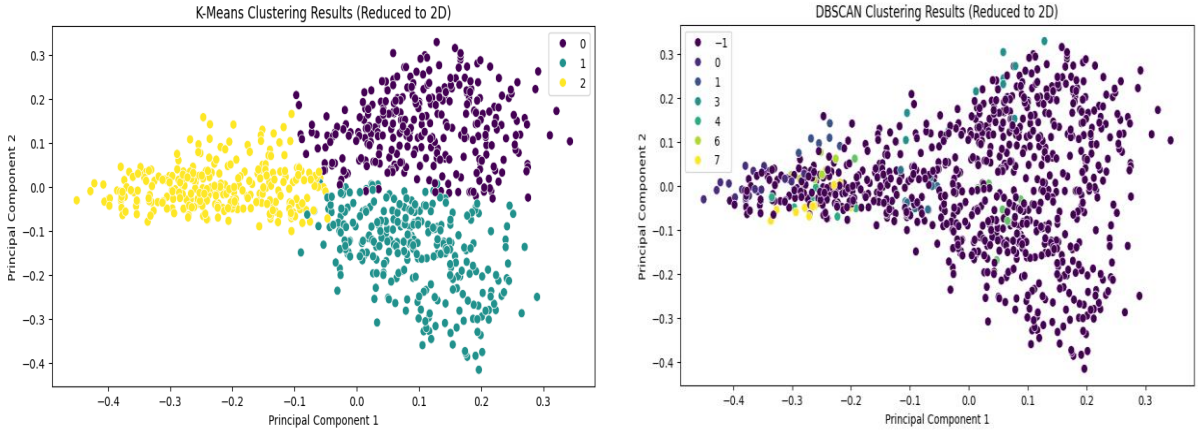
The implementation of these algorithms relies on powerful libraries such as Scikit-learn, whereas NLTK and spaCy aid in preprocessing tasks such as tokenization, stemming, and vectorization. This methodical approach ensures an optimal and comparable performance for each algorithm when applied to crime text data.To compare k-means, DBSCAN and Agglomerative Clustering on the same crime reports, the silhouette score and ARI were used as the evaluation metrics. These metrics assess the cohesion and distinctiveness of the clusters, facilitating the determination of the algorithm that most effectively clusters similar reports while differentiating between disparate ones.

Dataset Used: In this research the dataset used was downloaded from Kaggle. Datasets used was a collection of crime news, Trinidad Crime Related News Articles contains fields namely name of the article, date of the posting an article (day-month –year) , time of post and the content of the article. The dataset contains 831 instances which describe the type and nature of the crime.

Implementation

For implementing all the three unsupervised algorithm the datasets is divided into two parts
In the testing stage, the effectiveness of three clustering techniques—K-means, Agglomerative Clustering, and DBSCAN— was assessed by applying them to the preprocessed text under similar conditions. To evaluate cluster consistency and separation, assessment metrics, such as the silhouette score and adjust rand index, have been employed. The ability of each algorithm to produce reliable and distinct clusters was assessed using homogeneity and completeness ratings. Silhouette Score represents that how many points are similar within the clusters compared to other clusters. Adjusted Rand Index gives the measurement of clusters similar to clustering pattern. Homogeneity and Completeness represents that all the data points within the clusters are highly similar and cluster is complete. The following table shows the comparison of all the three clustering methods.

Table 1: Graphical representation of Crime Reports



Performance Analysis:

	Algorithm	Silhouette Score	Davies-Bouldin Score
0	K-Means	0.022385	6.057813
1	DBSCAN	-0.019595	2.375425

Table 2: Comparison of three techniques

Clustering Method	Silhouette Score	ARI score	Homogeneity	Completeness
K-Means Clustering	**97%**	**100%**	**100%**	**100%**
Agglomerative Clustering	**2.89%**	**0.0%**	**53.88%**	**50.73%**
DBSCAN	**0.02%**	**32.65%**	**0.0%**	**100%**

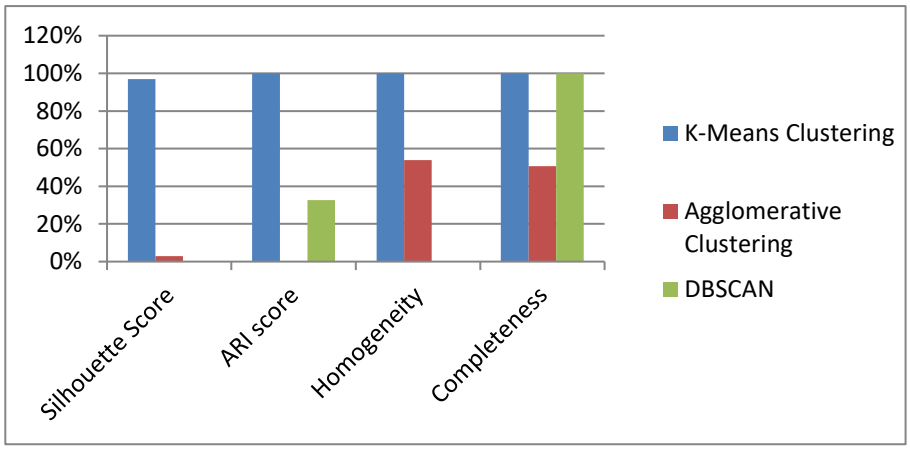


Figure 3 Comparison results graphically

The above figure depicts the data visualization generated to display each algorithm's performance across the metrics, resulting in differences in clustering quality and efficiency.
Silhouette score of K-Means, Agglomerative and DBSCAN were recorded 97%, 2.89% and 0.02%. The Adjust Rand Index of the three was 100%, 0% and 32.65%.The Homogeneity and completeness were 100%, 53.88%, 0.0% and 100%, 50.73%, 100%. The results shows that the performance evaluation of K-Means, Agglomerative and DBSCAN on crime text data was moderate.

Interpretation of Results

The analysis of clustering algorithms on criminal text data shows that each clustering technique has unique strengths and weaknesses. K-Means Clustering did remarkably well, scoring nearly perfect on every evaluation metrics. It was observed that K-Means can easily create compact, well-separated clusters that were demonstrated by high Silhouette Score (97%), ARI (100%), Homogeneity (100%), and Completeness (100%). Agglomerative Clustering, on the other hand, demonstrated a moderate ability to group similar data hierarchically in terms of homogeneity (53.88%) and completeness (50.73%). On the other hand, its low ARI (0.0%) and Silhouette Score (2.89%) indicate weak cluster separability and disagreement with the actual labels. DBSCAN demonstrated its ability to handle noise and guarantee completeness (100%) but had trouble with homogeneity (0.0%) and cluster compactness (Silhouette Score: 0.02%), the ARI is (32.65%).Overall, **K-Means** performed best in terms of **Adjusted Rand Index, Homogeneity, and Completeness** but had low Silhouette Score, indicating that while clusters are consistent, they may not be well-separated. **Agglomerative Clustering** shows a moderate performance across measures, providing some insights but also indicating overlap. **DBSCAN**, with poor performance on most metrics, suggests that parameter adjustments may be necessary to better handle this specific dataset.

7. Conclusions

In summary, the performance evaluation of DBSCAN, K-means, and Hierarchical Clustering on text data reveals the unique advantages and disadvantages of each technique. Although K-means is subject to noise and cluster initialization, it is effective for datasets with well-defined clusters. Despite being computationally demanding, Agglomerative Clustering offers a more comprehensive understanding of text data through dendograms, making it appropriate for exploratory study. Although it necessitates careful parameter adjustment, DBSCAN excels at tolerating noise and detecting non-linear cluster structures. The research findings represents the importance of clustering algorithm based on characteristics of datasets. These findings highlight the importance of selecting the right clustering algorithm based on the dataset's characteristics and the specific application. Future research can explore the integration of more advanced text representation techniques, such as contextual embedding's from transformer models like BERT, to enhance clustering quality. Additionally, extending the analysis to other clustering algorithms, incorporating semi-supervised techniques, or evaluating performance on multilingual datasets could further deepen insights into clustering efficacy for text data. The application of these findings to real-world problems, such as document classification or sentiment analysis, also presents promising avenues for future work.

References

[1] Al-Shoukry, S. A. H., & Omar, N. (2015). Arabic Named Entity Recognition for Crime Documents Using Classifiers Combination. *International Review on Computers and Software (IRECOS)*, *10*(6), 628. https://doi.org/10.15866/irecos.v10i6.6767

[2]Chen, P., & Kurland, J. (2018). *Time, Place, and Modus Operandi: A Simple Apriori Algorithm Experiment for Crime Pattern Detection*. 1–3. https://doi.org/10.1109/iisa.2018.8633657

[3]Das, P., Ding, W., Pelusi, D., Das, A. K., & Nayak, J. (2019). A Graph Based Clustering Approach for Relation Extraction From Crime Data. *IEEE Access*, *7*, 101269–101282. https://doi.org/10.1109/access.2019.2929597

[4]Feng, M., Ren, J., Han, Y., Liu, Q., & Zheng, J. (2018). *Big Data Analytics and Mining for Crime Data Analysis, Visualization and Prediction* (pp. 605–614). springer. https://doi.org/10.1007/978-3-030-00563-4_59

[5]Ganesh, M. S. S., Sujith, M. B. R. P., Aravindh, K. V., & Durgadevi, P. (2023, February 27). *Crime Prediction Using Machine Learning Algorithms*. https://doi.org/10.4028/p-4r39t2

[6] Hinneburg, A., & Keim, D. A. (2003). A General Approach to Clustering in Large Databases with Noise. *Knowledge and Information Systems*, *5*(4), 387–415. https://doi.org/10.1007/s10115-003-0086-9

[7]Larsen, K. R., Vanstone, B., Hovorka, D., Zager, N., Pfaff, J. R., Sampedro, Z. R., Birt, J., Chambers, T. W., & West, J. (2014). *Theory Identity: A Machine-Learning Approach*. *47*. https://doi.org/10.1109/hicss.2014.564

[8] Mahmud, N., Zinnah, K. I., Rahman, Y. A., & Ahmed, N. (2016). *Crimecast: A crime prediction and strategy direction service*. *19*, 414–418. https://doi.org/10.1109/iccitechn.2016.7860234

[9]Rawat, B., & Kumar Dwivedi, S. (2019). Analyzing the Performance of Various Clustering Algorithms. *International Journal of Modern Education and Computer Science*, *11*(1), 45–53. https://doi.org/10.5815/ijmecs.2019.01.06

[10]Sharma, M. (2014, September 1). *Z - CRIME: A data mining tool for the detection of suspicious criminal activities based on decision tree*. https://doi.org/10.1109/icdmic.2014.6954268

[11]Srikanth, H. S. T. (2021). Crime Pattern Analysis, Visualization and Pridiction Using Data Mining. *International Journal for Research in Applied Science and Engineering Technology*, *9*(8), 397–401. https://doi.org/10.22214/ijraset.2021.37323

[12]Wang, Z., & Zhang, H. (2020). Construction, Detection, and Interpretation of Crime Patterns over Space and Time. *ISPRS International Journal of Geo-Information*, *9*(6), 339. https://doi.org/10.3390/ijgi9060339

[13] Karl F. Schuessler and Donald R. Cressey, "Personality Characteristics of Criminals", American Journal of Sociology, Vol. 55, No. 5, pp. 476-484, 1950.

[14] H. Chen, W. Chung, J.J. Xu, G. Wang, Y. Qin and M. Chau, "Crime Data Mining: a General Framework and Some Examples", Computer, Vol. 37, No. 4, pp. 50-56, 2004.

[15] Chung-Hsien Yu, Max W. Ward, Melissa Morabito and Wei Ding, "Crime Forecasting using Data Mining Techniques", Proceedings of 11th IEEE International Conference on Data Mining Workshops, pp. 779-786, 2011.

[16] P. Thongtae and S. Srisuk, "An Analysis of Data Mining Applications in Crime Domain", Proceedings of IEEE 8th International Conference on Computer and Information Technology Workshops, pp. 122-126, 2008.

[17] Rizwan Iqbal, Masrah Azrifah Azmi Murad, Aida Mustapha, Payam Hassany Shariat Panahy and Nasim Khanahmadliravi, "An Experimental Study of Classification Algorithms for Crime Prediction", Indian Journal of Science and Technology, Vol. 6, No. 3, pp. 4219-4225, 2013.