Gezira J. of Engineering and applied Sciences
ISSN Linking: 1858-5698
http://journals.uofg.edu.sd/index.php/gjeas/

Crime Analysis using K-Means Clustering Algorithm - A Case Study: Police Departments, Wad Madani, Sudan (2016-2018)

Gais Alhadi Babikir[1], Tarig K. Suliman*[2], Awadallah M. Ahmed*[3]

1. Department of Computer Science, University of Gezira, Wad Madani, Sudan. gais.alhadi@uofg.edu.sd
2. Department of Computer Science, University of Gezira, Wad Madani, Sudan.Tarig38@gmail.com
3. Department of Computer Science, University of Gezira, Wad Madani, Sudan. awadallah@uofg.edu.sd

INFORMATION:

Submission: 30/06/2022

Accepted: 30/03/2023

Publication: 11/12/2023

ABSTRACT

In today's world, recently global security is one aspect that has a high priority and importance by politicians and governments around the world, it aims to reduce the occurrence of crimes. Therefore, many studies aimed at knowing the types of crimes, the manner of their occurrence, the search for criminals, study their personal qualities and public behavior since each crime shares the criminals who commit it in certain character traits and patterns. This study aims to find the relationships between crimes and the characteristics of criminals who committed the crimes (such as age, gender, location, and job) in order to help Police Departments to make the right decisions through crime investigation. In this paper, the k-means clustering algorithm has been used as a data mining technique to discover the relationships between crime and criminal characteristics. Where statistical descriptive analysis was used for all processes, also we used WEKA as a tool to implement the k-means algorithm. It should be noted that the data of this study have been collected from three Police Departments in Wad Madani city, Sudan for criminal records for three years (2016, 2017, and 2018). The analysis of the results showed that there are some crimes related to the specific age of criminals. We also found that each region has its own crimes which are different from the other region. Moreover, the most common crimes committed by males: are fraud, scandalous acts, public employee objection, check, and public threat. And, the most common crimes committed by females: are dealing with alcohol, abuse, and slander. Furthermore, the analysis of the result showed that the most common crimes committed by students are sexual, theft, and violence.

KEYWORDS: **Crime, Clustering Algorithm, Data Mining, Machine Learning**

1. INTRODUCTION

In recent years, the crime rate is increasing very fast in Sudan because of the increase in poverty and unemployment. With the existing crime investigation techniques, officers have to spend a lot of time as well as manpower to identify suspects and criminals. However, the crime investigation process has to be faster and more efficient as a large amount of information is collected during crime investigation. The Police Department in Sudan is the largest unit for preventing crimes and maintaining the orders of law, rules and peace throughout the country. However, the problem in Sudanese police offices is that they still use the traditional manual process such as the First Information Report (FIR) to keep and analyze records of crimes and criminals.

The most incredible challenge for the Police Departments is to investigate the crime using traditional techniques to deal with a large amount of information and a huge volume of records. Hence, it's difficult to discover the relationship between data elements (such as age, gender, location, and job) of the crime record, which contains invisible information. As well as the difficulty of making the right decision in cases of searching for a criminal who committed a particular crime. Therefore, we need advanced technologies to deal with crimes. Indeed, data mining is an approach that can be useful in this perspective. It should be noted that data mining is a process that extracts useful information from a large amount of crime data so that potential crime suspects can be identified efficiently.

It is noteworthy that during the past ten years, several studies have found recent developments in crime-fighting applications aimed at adopting data mining techniques to assist in the crime investigation process. For example, a tool to change the offender's behavior has been presented in [1]. The authors used extracted factors including frequency, severity, duration, and nature to compare similarities between offenders. Tayal et al. [2] provided an approach to design and

implement crime detection and criminal identification of Indian cities using data mining techniques. The authors used a k-means algorithm for crime detection analysis based on similar crime attributes. They also showed that the results obtained achieved an accuracy of 93.62 and 93.99% in the formation of two crime clusters using selected crime attributes. A model has been proposed in [3] by Zubi and Mahmmud for analyzing crime and criminal data using the k-means algorithm. The authors collected used data manually from Police Departments in Libya, in order to help make a strategic decision on preventing the increase in the high crime rate. To analyze the collected data, they used WEKA and Microsoft Excel. Also, in [4], the data mining approach clustering algorithm was used to discover crime patterns and speed up the crime-solving process. In fact, the authors used a k-means clustering algorithm to help identify crime patterns, and a semi-supervised learning method to discover knowledge from crime records. They stated that modeling technology has been able to identify crime patterns from a large number of crimes, making the task of crime detectives easier. In [5], crime analysis was performed by performing k-means aggregation on the crime dataset using the rapid mining tool. They showed that the murder rate decreased between 1990 to 2011. Moreover, the trend of crime over the years had been easy to determine and can be used to design preventive methods for the future. The data mining approach has been used in [6] to analyze, investigate and verify crime patterns. In fact, the authors used a clustering technique to analyze crime data, in which the stored data was compiled using the K-mean algorithm. They showed that they could predict a crime based on its historical information. Where the system can identify areas with a potentially high crime rate and distinguish areas with a higher crime rate. In [13], Vignesh et al. proposed a framework for analyzing crime data using the K-means clustering technique. In [14], the k-means clustering algorithm was applied to analyze the crime pattern and predict the crime occurrence rate in the future to take appropriate measures against it. The authors in [15] focused on combining the features of the K-Means Clustering algorithm with Dynamic Time Wrapping Algorithm for efficient Crime prediction and analysis. Kumar et al. [16] used the K-means algorithm to analyze crime under different locations and time periods.

The reader is invited to consult the papers by Saeed et al. (2021) [7], Jabeen and Agarwal (2021) [8], Kaur et al. (2021) [9], Mahmud et al. (2021) [10], Jeyaboopathiraja and Maria Priscilla (2021) [11], and Hussain and Aljuboori (2021) [12] for more state-of-the-art.

From the presented studies we have seen that in many countries and even in different police areas, the K-means algorithm plays an important role in analyzing and predicting crimes in real datasets. Hence, in this study, we will use the k-means clustering algorithm as a data mining technique to discover the relationships between crime and criminal characteristics. Obviously, the major contribution of our work is to discover the relationships between crimes and the characteristics of criminals (such as criminal age, criminal gender, criminal location, and criminal job) to assist Wad Madani Police Departments in making correct decisions. More clearly, this study can contribute to reducing the crimes that can occur in Wad Madani city based on the characteristics of criminals.

The rest of this paper is organized as follows. Section 2 describes our research methodology. In Section 3, we present the obtained results and discussion. Finally, Section 4 concludes this paper.

2. METHODOLOGY

In this paper, we used the k-means algorithm in order to transform the processed data into useful information and knowledge. As we can in Figure. 1, the process begins with manually collecting data from Police Departments, and the data includes information on crimes and criminals in Wad Madani city that occurred in 2016, 2017, and 2018. Then, we create the dataset and apply the data preprocessing stage, where we clean the data from the wrong and inconsistent data. Next, we perform the k-mean clustering algorithm and plot view to get the cluster. Finally, we perform crime analysis on cluster form.

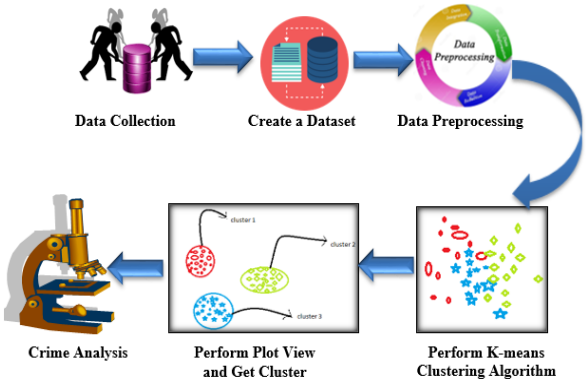

Figure.1 A Simplified methodology of crime analysis using simple k-means algorithm

A. Data collection

The data used in this study consists of 3000 distributed records, obtained from Wad Madani Police Department, with the following parameters

- Years: 2016, 2017, and 2018
- Location of Police Department: Middle, East, and South
- Gender: Male and Female
- Job: Worker, House Wife, Student and Driver
- Number of Crime Type: 134 Crime

Symbols and Notations that have been used in this paper.

Let $x \equiv \{G, J, L, A, CT, Y\}$ be the set of data points which:

- $G = \{g_1, g_2\} \equiv$ Gender,
- $J = \{J_1, J_2, J_3, J_4\} \equiv$ Job,
- $L = \{L_1, L_2, L_3\} \equiv$ Location,
- $A = \{a_1, a_2, a_3, \ldots, a_n\} \equiv$ Age,
- $CT = \{CT_1, CT_2, CT_3, \ldots, CT_n\} \equiv$ Crime Type and
- $Y = \{Y_1, Y_2, Y_3\} \equiv$ Year.

B. Data preprocessing

After the data was successfully collected from the police departments, an Excel sheet (dataset) was created to organize the data so that it would be easier to handle and analyze. In fact, these data need to be cleaned up and removed inconsistent evidence, so we need to preprocess the data. Obviously, this stage includes removing outliers in the data, predicting, and filling in missing values. It should be

noted that outliers are values that are unusual in the dataset, and can distort statistical analyses, potentially leading to misleading interpretations. So, removing this data and filling in missing values is an important aspect. It should be noted that since we already know the distribution of the data, we used common sense to find outliers that were incorrectly logged. Moreover, we plotted the data visually to find outliers. Hence, we have removed the outliers in the dataset, predicted the missing values, and filled them in to be fit and ready for analysis by WEKA.

C. Preform K-means clustering

After the preprocessing stage, the WEKA tool opened and the k-mean algorithm was performed. The steps of this K-means clustering algorithm can be described as follows.

Let X ≡ {G, J, L, A, CT, Y} be the set of data points and V ∈ X, {v ∈ X}.

- **Step1**: Calculate the distance between each data point and cluster center.
- **Step2**: Assign the data point to the center of the cluster, where the distance is closer to the center of the cluster, at least for all cluster centers.
- **Step3**: Recalculate the new cluster center using Eq. 1.

$$v_i = (1/c_i)\sum_{j=1}^{c_i} x_i \quad (1)$$

where c_i represents the number of data points in i^{th} cluster.

- **Step4**: Recalculate the distance between each data point and the new obtained centers of cluster.
- **Step5**: If there is no data point was reassigned then stop, otherwise repeat the previous steps starting from Step2.

Remark

In the next stages, we perform a plot view to get the cluster and analyze the given results.

3. RESULTS AND DISCUSSION

The following results have been obtained after analyzed the collected data using simple k-means algorithm and WEKA software.

To clarify the results accurately, in the following subsections, we will take the attribute crime type with the others attributes (i.e., Criminal Age, Criminal Location, Criminal Gender, and Criminal Job) separately to clarify their relationships with crime.

A. Crime type and criminal age

This subsection studies the relationship between the type of crime and the criminal age. See Figure.2 which illustrates the visualization of crime and age clusters. Obviously, we will consider two attributes (type of crime and criminal age) in the dataset and ignore the others. Then, the data will be divided into four clusters according to the criminal age attribute (See Table 1).

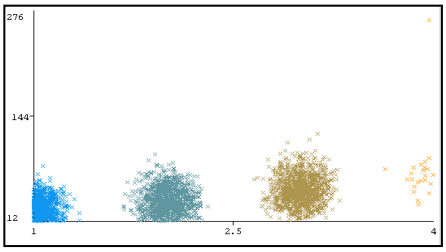

Figure. 2 Crime type and age clusters visualization

Table.1 Clusters of ages

Cluster No.	Number of Records	Range of Age
0	1260	18 to 25
1	1014	26 to 39
2	702	40 to 70
3	23	more than 70

Each cluster (age group) has its own crimes according to their psychological and functional needs and preferences. The analysis of the result showed that the crimes occurrence depending on the criminal age. For example, Figure. 3 shows the most common crimes with incidence rate for cluster 0 (i.e., range of age: 18 to 25).

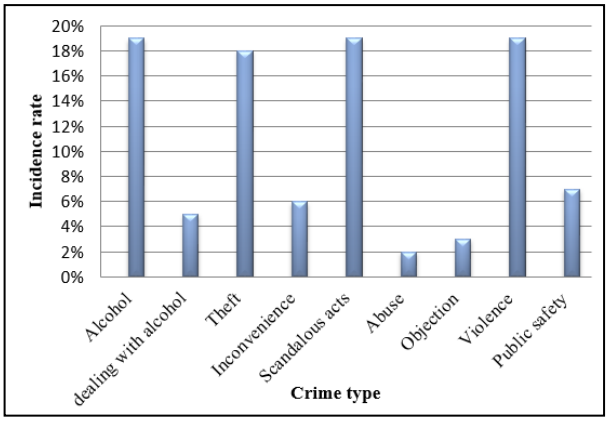

Figure. 3 The most common age crimes with incidence rate for cluster 0

As we can see, the most common crimes are: drinking alcohol, theft, scandalous acts, and violence with a ratio of %19, 18%, 19%, and 19% respectively. It should be noted that the reason for these crimes is the age at which the criminal goes through the stage of youth that makes him try to drink alcohol, theft, scandalous acts, and violence.

Moreover, as see we can see in Figure.4, the crimes of alcohol, theft, and violence are the most common crimes for cluster 1 (i.e., age between 26 and 39). Also, as we can see in Figure.5, the most common crimes for cluster 2 (i.e., age between 40 and 70) are: alcohol, theft, objection, violence, and check. In Figure.6, it clear that the most common crimes for cluster 3 are alcohol, theft, and check (i.e., age more than 70).

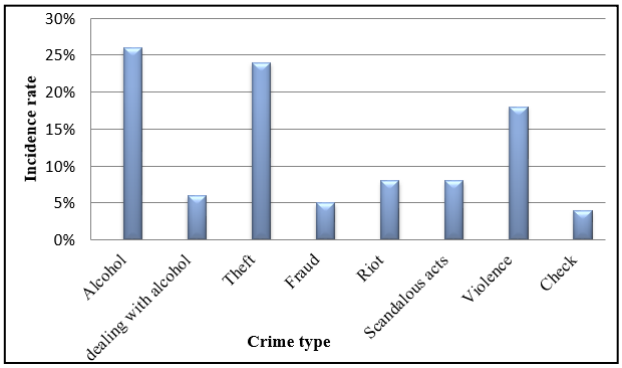

Figure. 4 The most common age crimes with incidence rate for cluster 1

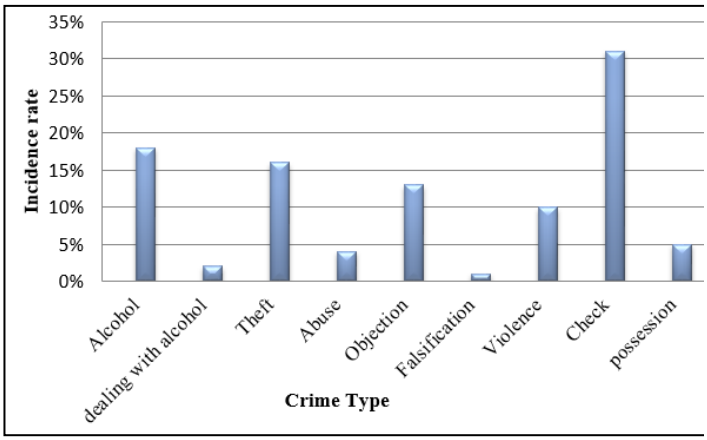

Figure. 5 The most common age crimes with incidence rate for cluster 2

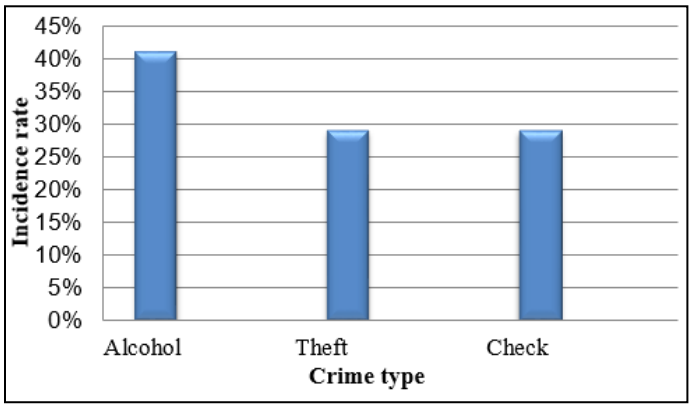

Figure. 6 The most common age crimes with incidence rate for cluster 3

B. Crime type and criminal location

This subsection studies the relationship between the type of crime and the location of the criminal. In fact, we will consider two attributes (crime type and criminal location) in the dataset and ignored the other ones. Therefore, the data has been divided into three clusters according to the criminal location attribute (See Table 2). In Figure.7, we present a visualization of crime and location clusters.

Table.2 Clusters of criminal location

Cluster No.	Number of Records	Location
0	1329	Middle
1	882	South
2	788	East

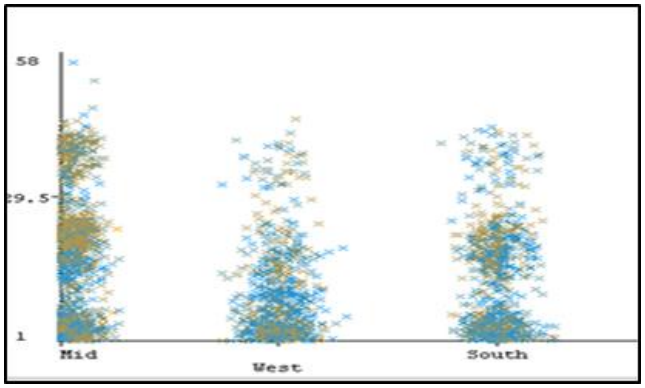

Figure. 7 Crime type and location clusters visualization

The analysis of the result showed that the crimes of alcohol, theft, objection, falsifying, and check are the most common crimes in the Middle Police Department (i.e., cluster 0) as shown in Figure.8. Where in the South Police Department (i.e., cluster 1) the most common crimes are shown in Figure.9, which are: alcohol, theft, violence, objection, and check. In the East Police Department (i.e., cluster 2) the most common crimes are shown in Figure.10, which are: alcohol, theft, scandalous acts, and prostitution.

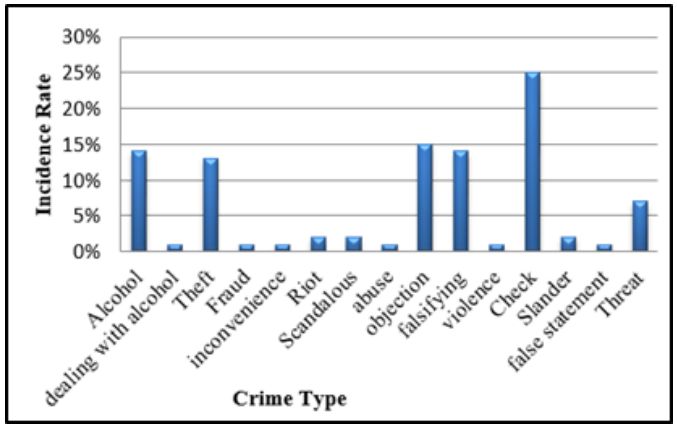

Figure. 8 The most common crimes with incidence rate in the Middle Police Department (i.e., cluster 0)

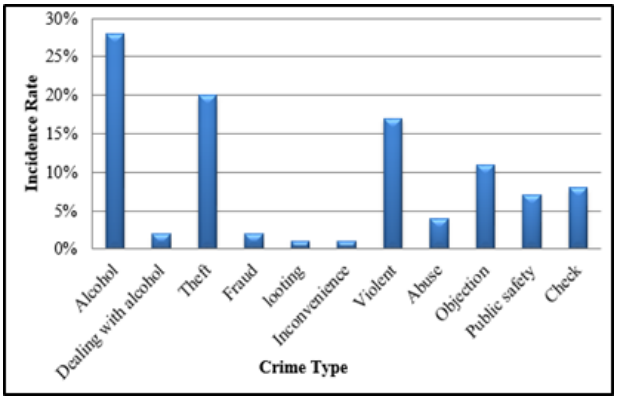

Figure. 9 The most common crimes with incidence rate in the South Police Department (i.e., cluster 1)

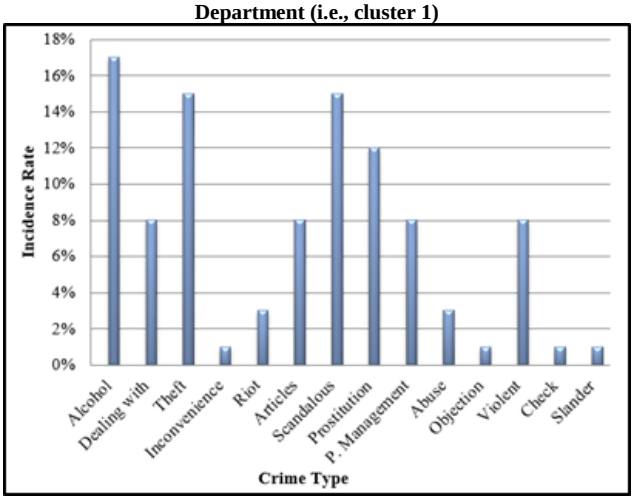

Figure. 10 The most common crimes with incidence rate in the East Police Department (i.e., cluster 2)

C. Crime type and criminal gender

In this subsection, we examined the relationship between crime type and the criminal gender (male, female). See Figure.11 which illustrates the visualization of crime and gender clusters. In fact, we will consider two attributes (type of crime and gender of the criminal) in the dataset and ignore the others. Then, the data will be divided into two clusters according to the criminal gender attribute (See Table 3).

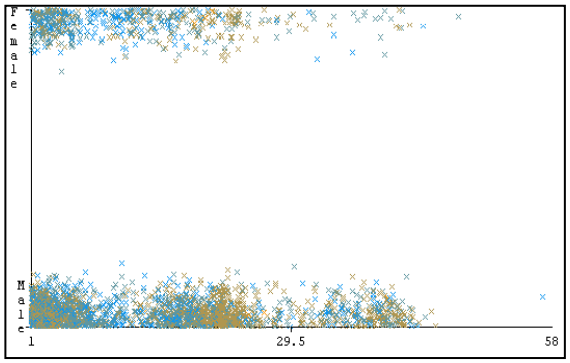

Figure. 11 Crime type and gender clusters visualization

Table.3 Clusters of criminal gender

Cluster No.	Number of Records	Gender
0	1214	Male
1	285	Female

As we know men differ from women in physiological, functional, and behavioral. Therefore, according to this difference we can see in Figure.12, there are some crimes committed by men and have the highest rate compared to females, while in Figure.13 we see that there are some crimes in which women exceed men. It is clear from Figure.12 that the crime of fraud was committed by 94% for males and 6% for females, scandalous acts 61% for males and 39% for females, and the crime of public employee objection is committed by men of 100%. The crime of check was committed by 92% for males and 8% for females, while the crime of public threat was committed by 82% for males and 18% for females.

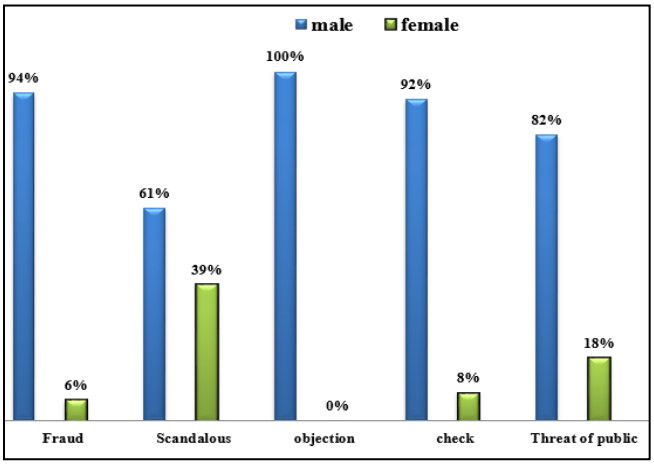

Figure. 12 The most common male crime compared to female crime

Furthermore, in Figure.13 we see that there are some crimes in which women exceed men. The crime of dealing with alcohol was committed by 58% for females and 42% for males, while the crime of abuse was committed by 75% for females and 25% for males. Finally, the crime of slander was committed by 72% for females and 28% for males.

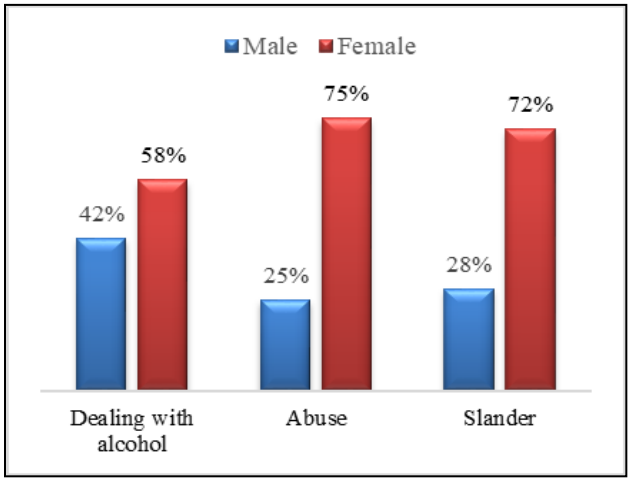

Figure. 13. The most common female crime compared to male crime

D. Crime type and criminal job

In this part, we will examine the relationship between crime type and the criminal job. Clearly, we will consider two attributes (crime type and criminal job) in our dataset and ignore the others. Hence, the data has been divided into four clusters according to the criminal job attribute (See Table 4). Figure.14 illustrates the visualization of crime and job clusters.

Table.4 Clusters of criminal job

Cluster No.	Number of Records	Job
0	1212	Worker
1	714	House wife
2	523	Student
3	550	Driver

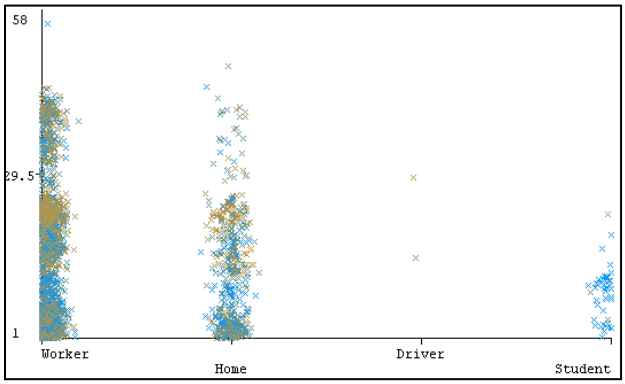

Figure. 14. Crime type and job clusters visualization

The analysis of the result showed that the most common crimes committed by students are sexual, theft, and violence with a ratio of %27, 68%, and 5% respectively (See Figure. 15). It should be noted that theft crime is the most common crime in the student community as they are trying to collect a lot of money in an easy way.

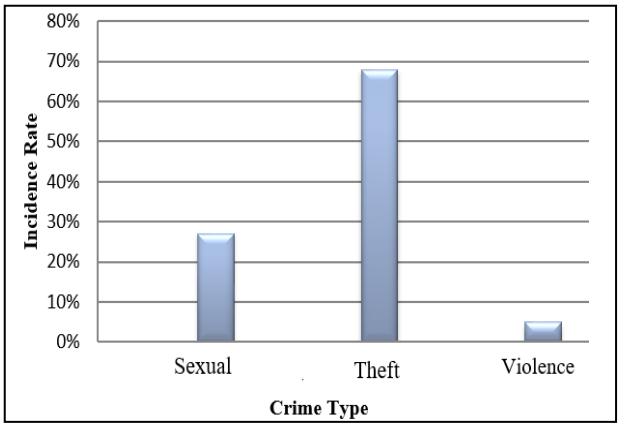

Figure. 15. The most common crimes with incidence rate committed by students

4. CONCLUSION

In this study, the k-means algorithm has been used as a data mining technique to discover the relationships between crimes and the characteristics of criminals. The data of this study have been collected from three Police Departments in Wad Madani city for three years (2016, 2017, and 2018). The results of the experiments showed the most common crimes for the criminals according to their ages, gender, locations, and jobs.

As the data size and the covering geographical area are increased. The solution provided by the crime analysis using the K-Means clustering technique needs more computation power. So, it needs various data mining techniques. In future work, intelligent investigation techniques can be developed for crime patterns such as association rule mining and data warehouse development for tracking crime trends and designing policing strategies.

REFERENCES

[1] J. S., De Bruin, T. K. Cocx, W. A. Kosters, J. F. Laros, and J. N. Kok, 2006. Data mining approaches to criminal career analysis. In Sixth International Conference on Data Mining (ICDM'06), pp. 171-177. IEEE.

[2] D. K. Tayal, A. Jain, S. Arora, S. Agarwal, T. Gupta, and N. Tyagi, 2015. Crime detection and criminal identification in India using data mining techniques. AI & society, 30(1), 117-127.

[3] Z.S. Zubi, and A. A. Mahmmud, 2013. Using data mining techniques to analyze crime patterns in the libyan national crime data. In Proceedings of the 1st WSEAS Interna-tional Conference on Image Processing and Pattern Recognition, pp. 79-85.

[4] S.V. Nath, 2006. Crime pattern detection using data mining. In 2006 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology Workshops, pp. 41-44. IEEE.

[5] A. Joshi, A.S. Sabitha, and T. Choudhury, 2017. Crime analysis using K-means clustering. In 2017 3rd International conference on computational intelligence and networks (CINE), pp. 33-39. IEEE.

[6] S. R. Deshmukh, A. S. Dalvi, T. J. Bhalerao, A. A. Dahale, R. S. Bharati, and C. R. Kadam, 2015. Crime investigation using data mining. International Journal of Advanced Research in Computer and Communication Engineering 4, no. 3: 22-24.

[7] S. Saeed, M. M. M. Bagram, and M. M. Iqbal, 2021. An Intelligent Analysis of Crime Data using Data Mining Algorithms. Technical Journal, 26(01), 102-115.

[8] N. Jabeen, and P. Agarwal, 2021. Data Mining in Crime Analysis. In Proceedings of Second International Conference on Smart Energy and Communication (pp. 97-103). Springer, Singapore.

[9] P. Kaur, G. Rani, T. Sharma, and, A. Sharma, 2021. A Comparative Study to analyze crime threats using data mining and machine learning approach. In 2021 International Conference on System, Computation, Automation and Networking (ICSCAN) (pp. 1-4). IEEE.

[10] S. Mahmud, M. Nuha, and A. Sattar, 2021. Crime Rate Prediction Using Machine Learning and Data Mining. In Soft Computing Techniques and Applications (pp. 59-69). Springer, Singapore.

[11] J. Jeyaboopathiraja, and J. Maria Priscilla, 2021. A Thorough Analysis of Machine Learning and Deep Learning Methods for Crime Data Analysis. In Inventive Computation and Information Technologies (pp. 795-812). Springer, Singapore.

[12] F. S. Hussain, and A. F. Aljuboori, 2021. Survey on Crime Analysis Using Data Mining Based on Mobile Platforms. Journal of Al-Qadisiyah for computer science and mathematics, 13(1), Page-36.

[13] K. Vignesh, P. Nagaraj, V. Muneeswaran, S. Selva Birunda, S. Ishwarya Lakshmi, and R. Aishwarya, 2022. A Framework for Analyzing Crime Dataset in R Using Unsupervised Optimized K-means Clustering Technique. In Congress on Intelligent Systems: Proceedings of CIS 2021, Volume 1 (pp. 593-607). Singapore: Springer Nature Singapore.

[14] S. Umasare, S. Phirke, S. Thakur, and V. Kulkarni,2022. Crime Rate Analysis and Prediction Using K-Means. Journal homepage: www. ijrpr. com ISSN, 2582, 7421.

[15] V. Gendre, N. K. Chandrakar, L. K. P. Bhaiya, and V. K. Swarnkar, 2022. Efficient Crime Analysis Based on Hybrid Approach by Combining Dynamic Time Wrapping Algorithm with K-Means Clustering Approach. International Journal for Research in Applied Science & Engineering Technology (IJRASET**).** ISSN: 2321-9653.

[16] J. Kumar, M. Sravani, M. Akhil, P. Sureshkumar, and V. Yasaswi, 2022. Crime Rate Prediction Based on K-means Clustering and Decision Tree Algorithm. In Computer Networks and Inventive Communication Technologies: Proceedings of Fourth ICCNCT 2021 (pp. 451-462). Springer Singapore.