# Clustering method for criminal crime acts using K-means and principal component analysis

**Ratih Hafsarah Maharrani[1], Prih Diantono Abda'u[2], Muhammad Nur Faiz[1]**
[1]Cybersecurity Engineering, Department of Computer and Business, Cilacap State Polytechnic, Cilacap, Indonesia
[2]Informatics Engineering, Department of Computer and Business, Cilacap State Polytechnic, Cilacap, Indonesia

**Article Info**

***Article history:***

Received Nov 28, 2023
Revised Dec 17, 2023
Accepted Jan 6, 2024

***Keywords:***

Clustering
Criminal acts
Dimension reduction
Index davies bouldin
K-means algorithm
Principal component analysis
Z score

**ABSTRACT**

Criminality is an act of violating the values and norms of society that causes a lot of harm. Much of the criminal data is often just a collection of data that has no information. Analysis of crime data is key in efforts to reduce crime rates that provide an overview of the incidence of crime, patterns, levels of vulnerability, and the level of security of an area. This research proposes data analysis that provides an understanding of crime using data mining techniques, especially the K-means cluster method, both traditional and with principal component analysis (PCA) dimension reduction. Before the PCA process, the values are transformed first with Z score normalization. From the processing through the davies bouldin index (DBI) performance test with 3 clusters, it is concluded that traditional K-means produces a DBI Index value of 0.019 and K-means PCA of 0.299. Meanwhile, to see the optimal cluster, several iterations were performed and resulted in the most optimal DBI index of 4 clusters in K-means of 0.014 and K-means PCA of 0.172. From the performance test value, it means that in the context of clustering the traditional criminal K-means data is declared more optimal than K-means PCA.

***Corresponding Author:***

Ratih Hafsarah Maharrani
Cybersecurity Engineering, Department of Computer and Business, Cilacap State Polytechnic
Cilacap, Central Java, Indonesia
Email: ratih.hafsarah@pnc.ac.id

## 1. INTRODUCTION

Criminality is an act of crime that has materially harmed, endangered safety, is not approved by the community because of a feeling of insecurity and is still being sought to suppress its growth [1], [2]. This is influenced by several factors such as negative associations, economic problems, environment, welfare level and age [3], [4]. Criminality requires special attention, because all over the world there are people in conflict with the law, of which two-thirds are in prison and the rest are in the custody of social institutions [5]. In Indonesia, in the scope of the police, especially the Cilacap Regency Police, criminal reports have been grouped by type of crime in numbers, but the details of each crime are still separated in different forms/reports. The large amount of data is often just a collection of data that has no information because the dataset is separate, unstructured and requires complicated processing so that in data analysis it needs to be compared from one report to another.

Every data has a pattern, where this pattern can be used to describe the condition of order, classify the level of vulnerability or the level of security of an area. Analysis is needed to group criminal data through a high and low crime rate approach in an area. The analysis in question is data mining which is used in the search for hidden information in large databases, through observation, data identification and data processing into information that is interrelated with one another [6], [7]. Data mining with K-means cluster has helped in clustering homicide crime analysis data and successfully identified crime trends every year [8]. Other studies have also found geographical patterns in crime and thus proposed a spatiotemporal crime clustering technique with 2-Dimensional Hotspot analysis with indicators of time, weather, location, census parameters such as annual opinion and literacy rate. It was found that the hotspot analysis model has better performance [9]. Spatiotemporal data was also investigated for point crime detection and prediction of future events with fuzzy C-means through the sum of squared error (SSE) approach and the Dunn index in measuring the quality of the cluster. It was found that this method is effective in finding spatiotemporal crime clusters [10]. K-means cluster has also been developed in crime prediction for the discovery of similar features, patterns and values which are then categorized in Surigao del Norte municipalities. This research resulted in the prediction of the increase of crime in each year both indexed and non-indexed through clustering technique [11].

In this research, a comparison of 2 data mining methods is proposed, namely K-means cluster (traditional) and K-means cluster using PCA dimension reduction. PCA as a dimension reduction method is used in the process of reducing the number of features or variables in a dataset while maintaining some of the information in the dataset. This method is used because K-means cluster works well for large amounts of data in crime analysis. Crime data can be large and very diverse, while K-means cluster can cope well with such data through data analysis based on existing patterns and this method is centroid-based unsupervised learning by partitioning a set of data into several parts [12], [13]. In the initial normalization of the PCA method, the use of Z score is necessary to convert the variables in the dataset so that they have a uniform scale because the crime dataset has different scales such as the range of the number of crimes in each region.

This research contributes to the clustering of criminal offenses by looking at changes in data from month to month by comparing two methods, namely traditional K-means and K-means using PCA dimension reduction. The criminal data is grouped based on the type of loss caused, time, type of crime and other patterns that appear more clearly. The clustering results were then tested using davies bouldin index (DBI) to see the most optimal number of clusters. From the research results, it is expected to know the most optimal algorithm from the number of clusters where the selected algorithm can display the types of crimes that often occur each month so as to help law enforcement officials to identify crime-prone areas and identify seasonal patterns in preventing or overcoming crime in an area. In addition, this research can be used as a foundation for the development of other clustering techniques or using other approaches in the analysis of criminal offenses.

## 2. METHOD

Based on Figure 1, the first step is the collection of research datasets, where the scope of this research is the total amount of criminal data per month in each police station in Cilacap Regency. Not all variables are used in this study, including the physical loss caused per type of crime, age, time of the incident, education level and motive for committing a criminal act which is shown in Table 1. The dataset used is criminal data for all Cilacap District Police in 2021-2022, which comes from 2 different types of data, the type of incident (derived from crime report data) and case anatomy recapitulation data, all in excel format consisting of 63 columns and 520 rows.

Table 1. Variable representation

| Variable | Sub variable | | |
|---|---|---|---|
| Time of Incident | X1 : 06.00-08.59;<br>X4 : 15.00-17.59; | X2 : 09.00-11.59;<br>X5 : 18.00-20.59; | X3 : 12.00-14.59<br>X6 : 21.00-23.59 |
| Offender Education | X7 : Unknown Education Level<br>X10 : Junior High School | X8 : Elementary school | X9 : High School; |
| Age of Perpetrator | X11 : 16-21 Years;<br>X14 : 41-50 Years; | X12 : 22-30 Years | X13 : 31-40 Years; |
| Victim's condition | X15 : Severe Injury; | X16 : Minor Injuries; | X17 : Passed Away; |
| Type of Crime | X18 : Conventional Crime; | X19 : Transnational Crime; | X20 : Interference with People |
| Crime Motive | X21 : Economy; | X22 : Due to Negligence/Culva; | X23 : Intentionally/Dolus; |

Furthermore, the existing dataset comes from several separate tables, so it is necessary to do data integration so that all data can be made into one. This is done to make it easier in the next analysis stage. Then next is data cleaning or cleaning of missing or duplicated data, thus causing the analysis results to be more accurate to be processed to the next stage. In data cleaning, deletion is carried out on variables that are not used in the calculation process and on data where no crime has occurred at all so that the total amount of data used after the data cleaning process is only 312 records. The stages of the research method are shown in Figure 1.

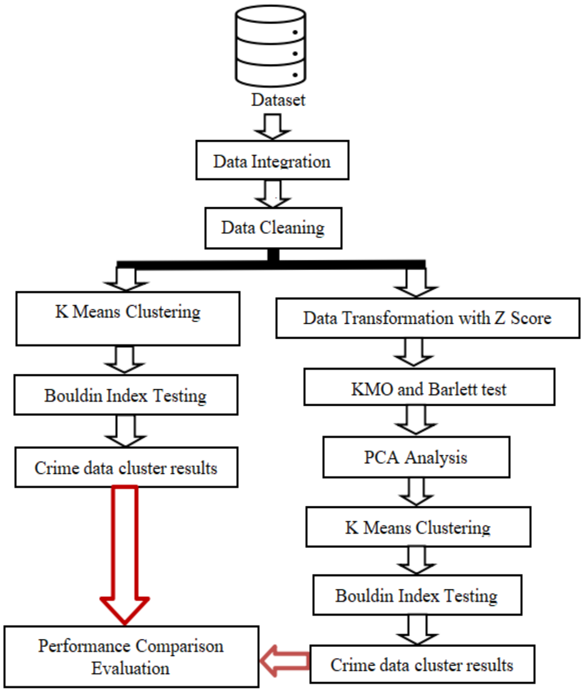

Figure 1. Research method

PCA is a statistical method as a data dimensionality reduction technique through the identification of patterns as well as the underlying structure of the data [14], [15]. The goal of PCA is to reduce the size of the data as much as possible without losing important information by embedding the data into a linear subspace with lower dimensions, where observation values located close to the reduced dimensional space are shown to have similar properties [16], [17]. The calculation of PCA values is based on the calculation of eigenvalues and eigenvectors that represent the data distribution of a dataset [18]. By using PCA, the previous n variables are selected into k new variables called principal components (PC) where the number of k is less than n. The stages in PCA analysis consist of data normalization such as Z score normalization, Kaiser-Meyer Olkin (KMO) and Barlett test [19].

A. Z score normalization

Z-score is a method to convert data into a score relative to the population mean and standard deviation. In PCA, Z score refers to the process of normalizing input variables before applying PCA analysis, so that any attribute with a large domain will not be more dominant with smaller attributes [20]. This normalization is important to ensure the variables have an equal scale before PCA is processed, besides that this is due to the difference in the size of the range in each variable. The Z score formula is written in (1):

$$Z_i = \frac{(xi - \mu)}{\sigma} \tag{1}$$

where Z : Z score (standardized score), I : 1st, 2nd, 3rd, ....n data, x: Observed value (raw score), μ: population mean, σ: population standard deviation.

B. KMO

KMO is a statistical measure to determine the suitability of results for software prediction through analysis of test data in measuring the feasibility or suitability of data for PCA analysis [21]. The sample adequacy test ranges between 0 and 1 where $0.9 < \text{KMO} < 1.00$ (very good data adequacy); $\text{KMO} > 0.8$ (good data adequacy); $\text{KMO} > 0.7$ (median adequacy); $\text{KMO} > 0.6$ (low adequacy) and $\text{KMO} < 0.5$ (very low adequacy/inappropriate data) [22].

C. Barlett's test

Barlett's test is used to determine whether the P value is smaller than the significant level ($\propto = 0.05$), where the null hypothesis can be rejected, which means that there is a significant correlation in the data and not all factors have the same variance [23]. In testing the freedom between variables the hypothesis in the Barlett's test can be explained if $H0 : P = 1$ (there is no correlation between variables) and $H1 : P \neq 1$ (there is a correlation between variables).

Clustering is a process in chain mining that identifies groupings of objects based on information obtained from data, based on the principle of maximizing similarity between class members and minimizing similarity between classes or clusters [24], [25]. The higher the intra-cluster similarity and the lower the inter-cluster similarity, the better the clustering results [26]. The clustering methods proposed in this study are traditional K-means and K-means PCA. In the traditional K-means process, the data does not undergo a previous conversion process. With the first step of determining the number of initial clusters (3 clusters), then determining the centroid value (cluster center) which is done randomly. Meanwhile, if determining the centroid value which is the stage of iteration, the formula (2) is used:

$$\overline{v_{ij}} = \frac{1}{N_i} \sum_{k=0}^{N_i} x_{kj} \tag{2}$$

after that, calculate the distance between the centroid point and the point of each object in (3).

$$D_e = \sqrt{(x_i - s_i)^2 + (y_i - t_i)^2} \tag{3}$$

The last step is grouping objects to determine cluster members by taking into account the minimum distance of the object and repeating until the resulting centroid value is fixed and cluster members do not move to other clusters.

Meanwhile, the k-means PCA method is a variation of the k-means method that uses the process of reducing the dimensions/features of the dataset. Data in PCA is stored into 2 columns with the same number of rows in each column, namely 312 rows. The steps performed in PCA are the same as conventional K-means, but at the beginning the data will be transformed into a new space with the main components obtained from PCA so as to reduce the number of features to minimize calculation complexity.

From the two algorithms, the next stage is accuration testing where the accuracy calculation is done by calculating the DBI so that the most optimal one is obtained in determining the criminal area. DBI is a measure in evaluating the work of a clustering, which has a positive correlation with "within-class" cases and a negative correlation with "between-class" cases [27]. The way DBI works is through validating the amount of data and property in a dataset to see how good the existing clusters are, so that 3 values are obtained, namely the distance between clusters, the centroid data center and the ratio owned by the cluster.

## 3. RESULTS AND DISCUSSION

In the research conducted, there are 2 processes carried out where the final results will be compared with each other. In the initial discussion, it will be discussed about the K-means stage which previously conducted a dimension reduction analysis using PCA. To facilitate the calculation of values in the PCA process (KMO and bartlett test), the analysis is processed using the python library. All values are taken in each month of the incident, where the processing results are expected to show the type of vulnerability of each with detailed variables in Table 2. An unbalanced range of Table 2 values in this crime data (for example, some areas have a high number of crimes while others have a low number) can have a significant influence on the clustering results. This may lead to the dominance of certain clusters by areas with a high number of crimes.

Based on the data in Table 2, it can be seen that the data has a range due to differences in the number of events in each month in each region. For this reason, it is converted into a different scale and easier to interpret with a Z score. The process of transforming the dataset using Z score standardization is done by transforming the dataset to allow for a fair comparison within a region so that it can be seen how criminal data values deviate from the average and compare them to other values in the dataset. Table 3 shows some of the data from the Z score data transformation using in (3) where previously at the beginning of the processing, the dataset has been integrated and data cleaned.

Classification results can be influenced by one of them is from the dimension and complexity of high data, to reduce errors during the classification process, it is necessary to reduce the dimensions using PCA. The stages of the PCA method in Table 4 include testing the KMO test value to see the overall sample adequacy and Bartlett's test which is used to determine whether there is a relationship between the variables used. These two tests are used in the context of factor analysis to check the fit of the data and the significance of the correlation matrix before proceeding to further stages of analysis. In the results of Table 4, the KMO value is 0.73, which indicates that the data has a fairly good fit for factor analysis. The Barlett's test results show a fairly large chi-square value (16310.2068) with a df (degrees of freedom) of 22 and a very small p-value (0.0000000), indicating a significant correlation between the variables in the dataset.

Then from the scree plot graph in Figure 2, it can be explained that there are 2 factors that produce a total eigenvalue> 1 (PC1-PC2) which summarizes most of the information from the data and allows simpler mapping. Factors that have an eigenvalue (1) cannot be used or are excluded from the calculation. The point at which the curve begins to flatten or slope may be an indication that the principal components after that point contribute less to the variability.

Table 1. Research dataset

| Attribute No | Month | X1 | X2 | X3 | X4 | X5 | X6 | X7 | X8 | X9 | X10 | … | … | X30 | X31 | X32 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | January | 0 | 0 | 2 | 1 | 0 | 4 | 7 | 1 | 1 | 0 | … | … | 3 | 3 | 2 |
| 2 | January | 0 | 0 | 1 | 0 | 0 | 2 | 0 | 1 | 1 | 0 | … | … | 1 | 1 | 1 |
| 3 | January | 1 | 1 | 1 | 4 | 1 | 3 | 3 | 4 | 3 | 3 | …. | …. | 6 | 1 | 0 |
| 4 | January | 0 | 1 | 2 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | … | … | 1 | 2 | 0 |
| 5 | January | 0 | 2 | 2 | 1 | 1 | 2 | 6 | 0 | 0 | 0 | … | … | 1 | 2 | 3 |
| 6 | January | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | … | … | 0 | 1 | 0 |
| 7 | January | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 2 | 0 | 0 | … | … | 1 | 0 | 0 |
| 8 | January | 1 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | … | … | 0 | 0 | 1 |
| 9 | January | 1 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | … | … | 0 | 0 | 0 |
| 10 | January | 1 | 1 | 0 | 0 | 1 | 2 | 2 | 0 | 0 | 0 | … | … | 4 | 0 | 2 |
| 11 | January | 0 | 14 | 35 | 57 | 11 | 18 | 151 | 2 | 7 | 4 | … | … | 0 | 17 | 134 |
| …. | …. | …. | …. | …. | …. | …. | …. | …. | …. | …. | …. | …. | …. | …. | …. | …. |
| 188 | August | 1 | 1 | 4 | 1 | 1 | 3 | 4 | 2 | 2 | 0 | …. | …. | 4 | 1 | 0 |
| 189 | August | 0 | 0 | 2 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | …. | …. | 1 | 1 | 0 |
| …. | …. | …. | …. | …. | …. | …. | …. | …. | …. | …. | …. | …. | …. | …. | …. | …. |
| 311 | December | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | … | … | 0 | 0 | 0 |
| 312 | December | 0 | 0 | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | … | … | 0 | 1 | 0 |

Table 2. Z score normalization transformation

| Attributes No | X1 | X2 | X3 | X4 | X5 | X6 | X7 | X8 | X9 | … | X22 | X23 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | -0,259 | -0,212 | -0,028 | -0,137 | -0,275 | 0,658 | 0,140 | 0,734 | 0,315 | … | -0,104 | 0,748 |
| 2 | -0,259 | -0,212 | -0,122 | -0,241 | -0,275 | 0,174 | -0,187 | 0,734 | 0,315 | … | -0,166 | 0,251 |
| 3 | 0,462 | -0,072 | -0,122 | 0,173 | -0,067 | 0,416 | -0,047 | 3,916 | 1,549 | … | -0,166 | -0,245 |
| 4 | -0,259 | -0,072 | -0,028 | -0,137 | -0,275 | -0,309 | -0,140 | -0,326 | -0,303 | … | -0,135 | -0,245 |
| 5 | -0,259 | 0,069 | -0,028 | -0,137 | -0,067 | 0,174 | 0,093 | -0,326 | -0,303 | … | -0,135 | 1,244 |
| 6 | -0,259 | -0,212 | -0,215 | -0,241 | -0,275 | -0,067 | -0,187 | -0,326 | -0,303 | … | -0,166 | -0,245 |
| 7 | -0,259 | -0,212 | -0,215 | -0,241 | -0,067 | -0,309 | -0,187 | 1,795 | -0,303 | … | -0,197 | -0,245 |
| 8 | 0,462 | -0,072 | -0,215 | -0,241 | -0,067 | 0,174 | -0,094 | -0,326 | -0,303 | | -0,197 | 0,251 |
| … | … | … | … | … | … | … | … | … | … | … | … | … |
| …. | …. | …. | …. | …. | …. | …. | …. | …. | …. | …. | …. | …. |
| 188 | -0,259 | -0,212 | -0,028 | -0,137 | -0,275 | -0,309 | -0,140 | -0,326 | -0,303 | …. | -0,166 | -0,245 |
| 189 | -0,259 | -0,212 | -0,215 | -0,137 | -0,067 | 0,174 | -0,094 | -0,326 | -0,303 | …. | -0,135 | -0,245 |
| … | … | … | … | … | … | … | … | … | … | … | … | … |
| 311 | -0,259 | -0,212 | -0,122 | -0,241 | -0,275 | -0,309 | -0,187 | -0,326 | -0,303 | … | -0,166 | -0,245 |
| 312 | -0,259 | -0,212 | -0,215 | -0,241 | -0,275 | -0,309 | -0,187 | -0,326 | -0,303 | … | -0,197 | -0,245 |

Table 3. KMO and Barlett test results

| Test type | Result | |
|---|---|---|
| KMO test | | 0.7300477 |
| Barlett test | Approx chi square | 16310.2068 |
| | Df | 22 |
| | P | 0.0000000 |

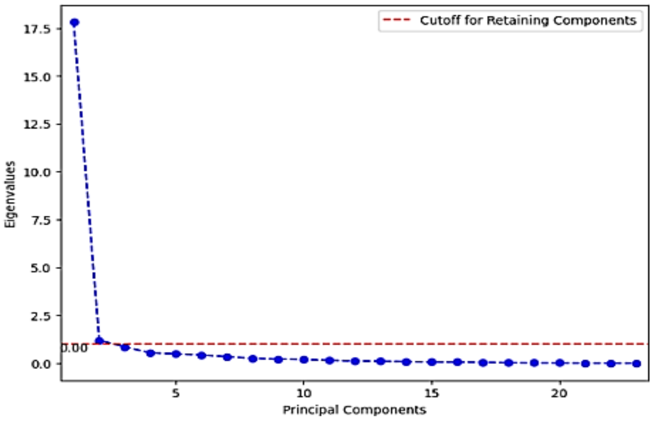

Figure 2. Scree plot of eigenvalues of PCA results

### 3.1. K-means cluster grouping

As a comparison conducted in this study, the criminal data clustering process is applied to different types of data, namely the original dataset and the dataset derived from processing using PCA. PCA resulted in data with 2 PCs as shown in Table 5. From the PCA results, there is a reduction in data dimensions by balancing between information loss and the optimal number of dimensions that must be maintained [28]. This data is then continued into the K-means clustering method.

The number of clusters determined at the beginning of the study was 3 clusters, which would represent criminal areas with very prone, prone, and moderately prone potential at each location per month of occurrence. The initial centroid used was chosen randomly and was updated each iteration based on the average location of all points in the corresponding cluster. This process will continue until convergence where there is no longer a significant change in the placement of data points into certain clusters or in the position of the centroid. At the beginning of the iteration (the number of clusters is 3), the results of K-means processing using PCA and traditional K-means are as written in Table 6.

From the results of Table 6, it can be seen that the number of members of each cluster is similar, but to produce good cluster quality in clustering algorithms, validation testing with the DBI is necessary. The DBI will understand the quality of the cluster by measuring how well the optimal cluster of a dataset is produced and how far the clusters are separated from each other. The lower the DBI value, the better the separation between clusters. The experiment is tested on the model built, where the number of clusters is determined from 2 to 6. From the segmentation results, the lowest DBI index value of the two methods is evaluated.

Table 4. Dataset slice from PCA processing

| Attributes No | pc_1 | pc_2 |
|---|---|---|
| 1 | 1,5 | 0,5 |
| 2 | -0,5 | 0,2 |
| 3 | 3,6 | 2,5 |
| 4 | -0,9 | 0,3 |
| 5 | -0,5 | 0,2 |
| 6 | -1,2 | -0,8 |
| 7 | -0,3 | 1,1 |
| 8 | -0,9 | -0,6 |
| … | …. | …. |
| …. | …. | …. |
| 188 | -1,0 | 0,2 |
| 189 | -0,9 | 0,1 |
| … | …. | …. |
| 311 | -1,2 | -0,8 |
| 312 | -1,3 | -0,8 |

Table 5. The result of grouping the criminal data into 3 clusters

| Police station (Polsek) | Month | K-means with PCA | Traditional K-means |
|---|---|---|---|
| Polsek Cilacap Tengah | January | cluster_0 | cluster_0 |
| Polsek Adipala | January | cluster_0 | cluster_0 |
| Polsek Kroya | January | cluster_0 | cluster_0 |
| Polsek Binangun | January | cluster_0 | cluster_0 |
| Polsek Nusawungu | January | cluster_0 | cluster_0 |
| Polsek Jeruklegi | January | cluster_0 | cluster_0 |
| Polsek Kawunganten | January | cluster_0 | cluster_0 |
| Polsek Bantarsari | January | cluster_0 | cluster_0 |
| …. | …. | …. | cluster_0 |
| Polsek Nusawungu | August | cluster_0 | cluster_0 |
| Polsek Jeruklegi | August | cluster_0 | cluster_0 |
| …. | …. | …. | …. |
| Polsek Maos | December | cluster_0 | cluster_0 |
| Polsek Sampang | December | cluster_0 | cluster_0 |
| Total cluster | | Cluster 0: 300 items<br>Cluster 1: 6 items<br>Cluster 2: 6 items | Cluster 0: 300 items<br>Cluster 1: 6 items<br>Cluster 2: 6 items |

The calculation of the DBI value on the clustering dataset, whether reduced (using PCA) or not, can produce several variations in value. From the data in Table 7, it can be concluded that the most optimal DBI value, which is the lowest DBI value (the smallest inter-cluster or intracluster distance) is obtained by the two methods with the number of k = 4. It can be concluded that the number of clusters in this study with the number of k = 3 is not optimal in clustering criminal offenses. The results of the DBI index in the traditional K-means method in this study produced a lower DBI value (0.014) than K-means using PCA variable reduction (0.172), indicating that the clusters generated by traditional k-means have a better degree of separation between clusters, or it could also indicate that PCA may sacrifice some important information when reducing the dimensionality of the data.

Table 6. Cluster evaluation results

| Total K | K-means | K-means using PCA |
|---|---|---|
| K = 2 | 0.016 | 0.174 |
| K = 3 | 0.019 | 0.229 |
| K = 4 | 0.014 | 0.172 |
| K = 5 | 0.021 | 0.244 |
| K = 6 | 0.025 | 0.320 |

From the cluster data generated in Figure 3, it can be seen that the 4 clusters formed can be defined as the level of vulnerability, namely cluster 0 = low, cluster 1 = medium, cluster 2 = high, and cluster 3 = very high. In cluster 3, only one area in January is considered highly vulnerable, which, when viewed from the data in Table 2 row 11 before processing which states the results, includes areas with many crimes. While other regions have a small number of crimes, there are even some regions that in certain months have no crimes at all, so they are included in cluster 0 where crimes rarely occur. This unbalanced range in the crime data (for example, some areas have a high number of crimes while others have a low number) can have a significant influence on the clustering results. This may lead to the dominance of certain clusters by areas with a high number of crimes.

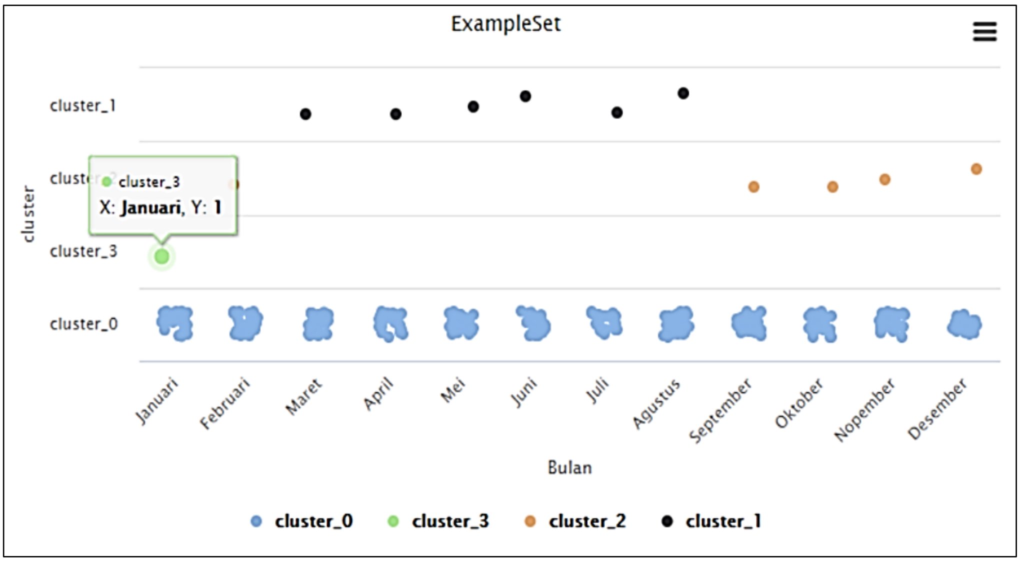

Figure 3. Criminal offense clustering results

## 4. CONCLUSION

Analyzing crime data monthly in each region can provide valuable insights into seasonal patterns and help identify potential factors contributing to crime fluctuations. Understanding these seasonal variations allows law enforcement to allocate resources more effectively and focus on specific areas and times when crime rates are high. Additionally, policymakers can implement targeted interventions in specific months to address the root causes of crime, resulting in more efficient crime prevention. Based on the results of the study, the number of clusters formed based on the determination of the initial number of clusters of 3 obtained the results of the DBI index K-means of 0.019 and K-means PCA of 0.299. This cluster is still considered less than optimal because the results of the DBI index with a different number of k (k = 4) can produce a smaller index of 0.014 in K-means and 0.172 in K-means cluster. So it can be concluded based on the results of DBI, K-means clustering is better than K-means PCA for the case of criminal data.

## ACKNOWLEDGEMENTS

Our gratitude goes to the Research and Community Service Center of Cilacap State Polytechnic in the 2023 internal research financing scheme (DIPA) for assisting in the research process.

## REFERENCES

[1] L. Sun *et al.*, "Explore the correlation between environmental factors and the spatial distribution of property crime," *ISPRS International Journal of Geo-Information*, vol. 11, no. 8, 2022, doi: 10.3390/ijgi11080428.

[2] E. I. Pratiwi, "Law enforcement efforts against the crime of body shaming through mediation," *Pancasila and Law Review*, vol. 1, no. 2, 2021, doi: 10.25041/plr.v1i2.2127.

[3] Y. N. Septiawan, "Analysis of the causes of narcotics recidivities in class IIa prisons in Bogor," *Walisongo Law Review (Walrev)*, vol. 2, no. 1, 2020, doi: 10.21580/walrev.2020.2.1.5321.

[4] A. Chariri, "Criminal settlement of criminal acts of motorcycle theft by child through restorative justice," *International Journal of Educational Research and Social Sciences*, vol. 3, no. 4, 2022, doi: 10.51601/ijersc.v3i4.442.

[5] N. R. Anandia, "Crime of murder by a child as the perpetrator a criminal psychology perspective," *Locus Journal of Academic Literature Review*, 2022, doi: 10.56128/ljoalr.v1i7.93.

[6] S. B. Patel, S. M. Shah, and M. N. Patel, "An efficient search space exploration technique for high utility itemset mining," in *Procedia Computer Science*, 2022, doi: 10.1016/j.procs.2023.01.074.

[7] S. Khademizadeh, Z. Nematollahi, and F. Danesh, "Analysis of book circulation data and a book recommendation system in academic libraries using data mining techniques," *Library and Information Science Research*, vol. 44, no. 4, 2022, doi: 10.1016/j.lisr.2022.101191.

[8] J. Agarwal, R. Nagpal, and R. Sehgal, "Crime analysis using K-Means clustering," *International Journal of Computers and Applications*, vol. 83, no. 4, 2013, doi: 10.5120/14433-2579.

[9] G. Hajela, M. Chawla, and A. Rasool, "A clustering based hotspot identification approach for crime prediction," in *Procedia Computer Science*, 2020, doi: 10.1016/j.procs.2020.03.357.

[10] M. Y. Ansari, A. Prakash, and Mainuddin, "Application of spatiotemporal fuzzy C-means clustering for crime spot detection," *Defence Science Journal*, vol. 68, no. 4, 2018, doi: 10.14429/dsj.68.12518.

[11] A. J. P. Delima, "Applying data mining techniques in predicting index and non-index crimes," *International Journal of Machine Learning and Computing*, vol. 9, no. 4, pp. 533–538, Aug. 2019, doi: 10.18178/ijmlc.2019.9.4.837.

[12] K. E. Setiawan, A. Kurniawan, A. Chowanda, and D. Suhartono, "Clustering models for hospitals in Jakarta using fuzzy c-means and k-means," in *Procedia Computer Science*, 2022, doi: 10.1016/j.procs.2022.12.146.

[13] M. Sinan, J. Leng, K. Shah, and T. Abdeljawad, "Advances in numerical simulation with a clustering method based on K–means algorithm and Adams Bashforth scheme for fractional order laser chaotic system," *Alexandria Engineering Journal*, vol. 75, 2023, doi: 10.1016/j.aej.2023.05.080.

[14] J. A. Bawa, P. Ayuba, and O. K. Akande, "Factors influencing the performance of indoor environmental quality of pharmaceutical factory buildings in Southwest Nigeria," in *IOP Conference Series: Earth and Environmental Science*, 2022, doi: 10.1088/1755-1315/1054/1/012023.

[15] S. Marukatat, "Tutorial on PCA and approximate PCA and approximate kernel PCA," *Artificial Intelligence Review*, vol. 56, no. 6, 2023, doi: 10.1007/s10462-022-10297-z.

[16] N. M. N. Mathivanan, N. A. MdGhani, and R. M. Janor, "A comparative study on dimensionality reduction between principal component analysis and K-means clustering," *Indonesian Journal of Electrical Engineering and Computer Science (IJEECS)*, vol. 16, no. 2, 2019, doi: 10.11591/ijeecs.v16.i2.pp752-758.

[17] D. W. Shin, B. J. Ko, J. C. Cheong, W. Lee, S. Kim, and J. Y. Kim, "Impurity profiling and chemometric analysis of methamphetamine seizures in Korea," *Analytical Science and Technology*, vol. 33, no. 2, 2020, doi: 10.5806/AST.2020.33.2.98.

[18] S. Surono, K. W. Goh, C. W. Onn, and F. Marestiani, "Developing an optimized recurrent neural network model for air quality prediction using K-means clustering and PCA dimension reduction," *International Journal of Innovative Research and Scientific Studies*, vol. 6, no. 2, 2023, doi: 10.53894/ijirss.v6i2.1427.

[19] S. Suárez, M. Cañizares, and W. Carvajal, "Attitudes and beliefs of high-performance cuban athletes about doping," *Cuadernos de Psicologia del Deporte*, vol. 22, no. 2, 2022, doi: 10.6018/cpd.485361.

[20] E. D. Jiru, B. G.Wordofa, and M. Redi-Abshiro, "Improved principal component analysis and linear discriminant analysis for the determination of origin of coffee beans using," *SINET: Ethiopian Journal of Science*, vol. 45, no. 1, 2022, doi: 10.4314/sinet.v45i1.1.

[21] S. Sivavelu and V. Palanisamy, "Gaussian kernelized feature selection and improved multilayer perceptive deep learning classifier for software fault prediction," *Indonesian Journal of Electrical Engineering and Computer Science (IJEECS)*, vol. 30, no. 3, 2023, doi: 10.11591/ijeecs.v30.i3.pp1534-1547.

[22] S. Cortés, S. Burgos, H. Adaros, B. Lucero, and L. Quirós-Alcalá, "Environmental health risk perception: Adaptation of a population-based questionnaire from latin america," *International Journal of Environmental Research and Public Health*, vol. 18, no. 16, 2021, doi: 10.3390/ijerph18168600.

[23] O. Oloruntobi, S. Mokhtar, M. Z. Shah, and K. Mokhtar, "Significant factors affecting public transport use for leisure travel and tourism," *AIMS Environ Science*, vol. 10, no. 1, 2023, doi: 10.3934/environsci.2023004.

[24] M. Zulkifilu and A. Yasir, "About some data precaution techniques for K-means clustering algorithm," *UMYU Scientifica*, vol. 1, no. 1, 2022, doi: 10.56919/usci.1122.003.

[25] C. Zhu, C. U. Idemudia, and W. Feng, "Improved logistic regression model for diabetes prediction by integrating PCA and K-means techniques," *Informatics in Medicine Unlocked*, vol. 17, 2019, doi: 10.1016/j.imu.2019.100179.

[26] G. Liu, F. Ji, W. Sun, and L. Sun, "Optimization design of short-circuit test platform for the distribution network of integrated power system based on improved K-means clustering," *Energy Reports*, vol. 9, 2023, doi: 10.1016/j.egyr.2023.04.319.

[27] J. Xiao, J. Lu, and X. Li, "Davies bouldin index based hierarchical initialization K-means," *Intelligent Data Analysis*, vol. 21, no. 6, 2017, doi: 10.3233/IDA-163129.

[28] D. Festa *et al.*, "Unsupervised detection of InSAR time series patterns based on PCA and K-means clustering," *International Journal of Applied Earth Observation and Geoinformation*, vol. 118, 2023, doi: 10.1016/j.jag.2023.103276.

## BIOGRAPHIES OF AUTHORS

**Ratih Hafsarah Maharrani** is a graduate of Master of Informatics Engineering, Dian Nuswantoro University, Semarang. Currently working at Cilacap State Polytechnic, Department of Computer and Business with a focus on the Cyber Security Engineering study program since 2019. With the existing educational background, several studies have been developed where his research focuses on machine learning and decision support systems. She can be contacted at email: ratih.hafsarah@pnc.ac.id.

**Prih Diantono Abda'u** he is a graduate of the Master of Informatics Engineering from Amikom University Yogyakarta with a concentration in Information Systems Audit, has now been pursuing his career at Cilacap State Polytechnic since 2019. Serving in the Computer and Business Department with a focus on the D3 Informatics Engineering study program, he has shown his dedication in developing students' understanding of the world of technology. With a strong educational background, Abda'u focuses his research on two main areas: software engineering and information systems audit. His expertise in combining aspects of software development and information system evaluation creates a solid foundation, enriches the academic environment and makes a valuable contribution to the development of information technology at Cilacap State Polytechnic. He can be contacted at email: abdau@pnc.ac.id.

**Muhammad Nur Faiz** he is a graduated from Ahmad Dahlan University Yogyakarta with a concentration in Cyber Security Engineering, digital forensics, and machine learning, has been working at Cilacap State Polytechnic since 2019. Serving in the Computer and Business Department with a focus on the D4 Cybersecurity Engineering study program, he has shown his dedication in developing an understanding of the importance of information security and technology ethics. He has published several scientific articles on cybersecurity that combine machine learning and digital forensics. He can be contacted at email: faiz@pnc.ac.id.