CRIME DATA IDENTIFICATION BASED ON EVENTS LOCATIONS IN WEST JAVA USING K-MEANS ALGORITHM

1st Khadijah Kibtiyah
Informatic Engineering Study Program
Nusa Putra University
Sukabumi, Indonesia
khadijah.kibtiyah_ti19@nusaputra.ac.id

2nd Muhamad Ridwan Nullah
Informatic Engineering Study Program
Nusa Putra University
Sukabumi, Indonesia
muhamad.ridwan_ti19@nusaputra.ac.id

3rd Aden Rahmat Ramdani
Informatic Engineering Study Program
Nusa Putra University
Sukabumi, Indonesia
aden.rahmat_ti19@nusaputra.ac.id

4th Mega Putri Utami
Informatic Engineering Study Program
Nusa Putra University
Sukabumi, Indonesia
mega.putri_ti19@nusaputra.ac.id

5th Alun Sujjada
Informatic Engineering Study Program
Nusa Putra University
Sukabumi, Indonesia
alun.sujjada@nusaputra.ac.id

6th Kamdan
Informatic Engineering Study Program
Nusa Putra University
Sukabumi, Indonesia
kamdan@nusaputra.ac.id

7th Hermanto
Informatic Engineering Study Program
Nusa Putra University
Sukabumi, Indonesia
hermanto@nusaputra.ac.id

Abstract—The problem of crime is a social problem that always demands serious attention from time to time. Moreover, according to general assumptions and some observations and research results of various parties, there is a trend of increasing development of certain forms and types of crime, both in quality and quantity. This crime does not look at place, gender, age or class. Therefore, all forms of crime must be tackled immediately because they can cause victims to suffer physical and psychological disorders. In this study, it will be explained how to collect data on the number of victims of crime that occurred in West Java based on the location of the incident which will later be processed into a data mining application program with Python language using the K-Means Clustering method. The data is calculated based on the number of cases reported by the community in each region in West Java that occurred in 2021. The K-Means Clustering method provides accuracy in grouping data and separating it by cluster categories by about 95%. In addition to calculating and grouping data based on clusters, this research is expected to help accelerate the collection of data on crime cases in West Java so that it can be handled as soon as possible by the government. As well as reducing crimes that occur in West Java with firm action based on the large data on crime in each region.

Keywords—Clustering, Data Mining, K-Means, Criminal, Python

I. INTRODUCTION

At the moment, crimes are increasingly rife, especially against women and children. This crime can happen anywhere, schools, workplaces, and family can be places for someone to commit crimes. This can certainly disturb the mentality of people who experience it. The evil that has been accepted as a child will continue to stick in the child's mind until he is an adult. Of the many cases, some of them are cured and can realize and accept themselves. But there are also many of them whose personalities change to isolate themselves from society. Others think that crime can make them win in any situation so that he becomes a criminal offender.

West Java is one of the provinces with the most crime in Indonesia. In 2021, there were around 964 cases of crime reported in West Java. Most of it occurs in family, the rest occurs in the work environment, schools, training institutions and some public facilities. This data continues to grow every year, firm handling from the government needs to be done to reduce the number of crimes that occur in West Java.

Based on the above problems, the author took the title of the research "Crime Data Identification Based on Events Locations in West Java Using K-Means Algorithm". K-Means is an algorithm used to analyze and group data into categories or clusters. The data collection method is carried out by taking a dataset on the official West Java Open Data website with crime categories according to the location of the incident in 2021. This study aims to accelerate the collection of crime data in West Java based on the number of cases in each City/Regency, so that the problem can be handled quickly and precisely. So that the number of crime cases in West Java can be reduced from year to year. In addition, assistance for victims of crime is also needed to deal with trauma and improve the victim's mentality.

II. RELATED WORK

This research was conducted by taking a dataset contained on the official West Java Open Data website, which is data on the number of victims of crime scenes in West Java in 2021. The data will later be processed using a data mining program with Python language and using the K-Means Clustering method. Then the clustering results will later know the number of crime victims in each Regency / City based on the location of the incident, which will later be divided into 2 clusters, namely low and high levels. This refers to the number of victims in each district/city, with several locations of incidents, namely family, workplaces, schools, public facilities, training institutions and others. Which is then broadly divided into family and environments (workplaces, schools, public facilities, training institutions and others).

This research will later be implemented with the Google Colab program using Python language with K-Means Clustering algorithm in analyzing a data. Its function is to get the information needed to group various types of data based on their categories. Research with the K-Means Clustering method is expected to be useful for the authorities as a consideration and data reference to solve crime problems in West Java.

III. Methodology

In identifying crime data based on incident locations in West Java Province in 2021, researchers used several stages and methods. The following are the stages and methods of the research conducted.

Fig. 1. Research Flow

3.1 Problem Identification

Identification of the problem is done by looking at and analyzing the crime data available on the government's official website Open Data Jabar. After it was concluded that the data was incomplete, development was carried out so that the data was easier to read and more concise. So that the countermeasures carried out can be more targeted based on the highest category and must be addressed immediately.

3.2 Study Literature

Study Literature are carried out by reviewing and looking for methods that can be done to solve the problems faced. Review of methods and theories obtained from the internet, books, and previous scientific works to complement the guidelines for research concepts. So that later this research has a clear scientific foundation based on the theories in the source.

3.3 Data Collection and Data Processing

At this stage, data related to research conducted by children are collected. Furthermore, the data will be processed according to the existing scheme. The schema is created by designing, collecting, processing and processing data on Google Colab using the K-Means Clustering algorithm with Python language. The final result of this study is to determine the low and high clusters of crime cases in West Java based on the location of their occurrence, with two categories given, namely family and environment.

3.4 Program Development

The stage of making the program is prepared with the waterfall method which contains the stages of activities carried out during the research in a concise manner.

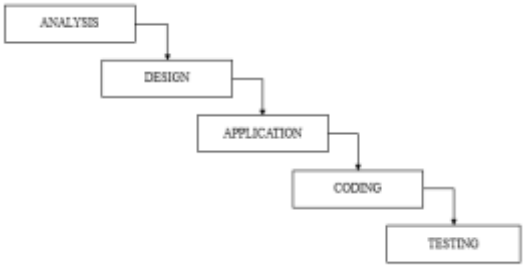


Fig. 2. Stages of Making a Waterfall Method Program

1) Analysis, which is looking for as much information as possible and analyzing which one is more suitable for use in the program to be created.

2) Design, carried out by pouring ideas and system design from the identification of problems made.

3) Creation, is the process of making programs according to needs and designs that have been carried out.

4) Coding, done using the python programming language on Google Colab.

5) Testing, the finished program will be tested using the black box method to find out whether the program has run according to the design at the beginning.

3.4 Write Report

The report is prepared by including all the elements of the research that has been done. Namely the objectives, benefits, and research activities carried out by researchers. Reporting is done by taking into account the theories used as the basis and reference in research.

IV. Results and Discussion

The result of the research activities carried out is a data in excel or xlsx format that has been processed using the K-Means Clustering algorithm. The data is a development of a raw dataset provided by the government from the official West Java Open Data website. This data displays the category of high and low crime rates in an area. The data can later be used to analyze which areas are most affected and must be addressed as soon as possible. So that it is hoped that crime in areas in West Java province can be suppressed.

4.1 Research Activity

A. Data Collection

The data used in the research of crime cases in West Java was obtained from the official West Java Open Data website entitled "Number of Violence Cases Based on the Scene of Violence in West Java". On the official website, the data is still raw data that contains many variables that must be converted and some of them must be eliminated when it enters the data processing stage.

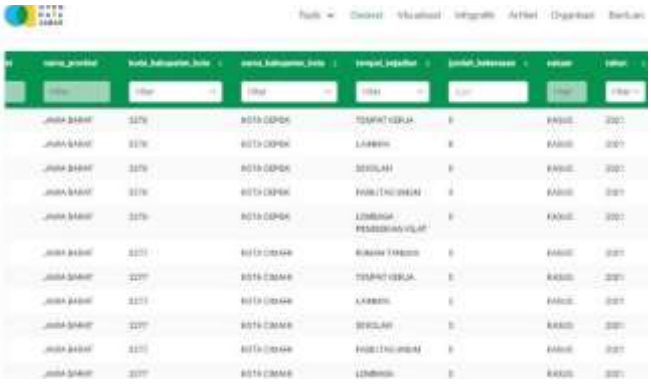


Fig. 3. Dataset in website Open Data Jabar

B. *Data Processing*

Data is processed in Microsoft Excel to tidy up data and make it easier to group data. Furthermore, the data will be processed on Google Colab using Python language using the K-Means Clustering method. Then the data is saved into CSV format to make it easier to read when imported into a python program that will be created on Google Colab.

4.2 Research Project Management

At this stage, the program will begin to be built and tested to ensure the program runs according to the initial design.

A. *Flowchart Diagram*

The following flowchart diagram displays a description of the activities performed when data processing and processing takes place.

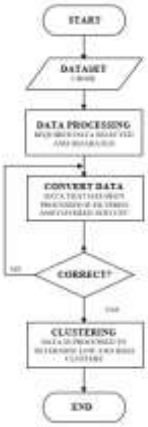


Fig. 4. Flowchart Diagram

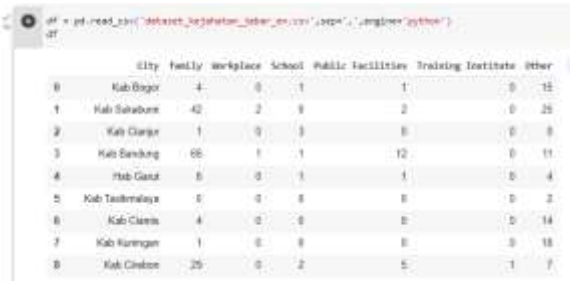
B. *System Design*

- Library, is a set of code that has certain functions and can be called into other programs. In this program, the libraries used are pandas and matploitlib. Pandas is a library that is often used to analyze data and build machine learning. While matplotlib is used for data visualization.

```
import pandas as pd
import matplotlib.pyplot as plt
```

Fig. 5. Library

- Import Data, CSV datasets are imported into the program with DF variables (dataframes with comma separators and Python programming language engine.



	City	Family	Workplace	School	Public Facilities	Training Institute	Other
0	Kab Bogor	4	0	1	1	0	15
1	Kab Sukabumi	42	2	9	2	0	25
2	Kab Cianjur	1	0	3	0	0	0
3	Kab Bandung	66	1	1	12	0	11
4	Kab Garut	6	0	1	1	0	4
5	Kab Tasikmalaya	0	0	0	0	0	2
6	Kab Ciamis	4	0	0	0	0	14
7	Kab Kuningan	1	0	0	0	0	18
8	Kab Cirebon	25	0	2	5	1	7

Fig. 6. Imported Dataset

- Describe Data, dataset that has been imported is then described to find out the sum, mean, standard deviation, lowest value, 25% value, 50% value, 75% value and the highest value in one column of the dataset.
- Search Null Character, the next step is to look for empty or unfilled data in a column in the dataset that has been imported for elimination.
- Combining Data and Creating New Columns, the West Java crime dataset has 7 variables and columns. Therefore, data that has the same category needs to be combined so that there are not too many variables. A new column is created with the name "Environment" which contains data from workplaces, schools, public facilities, training institutions and others. Next, the family and neighborhood columns are summed again and then stored in the "Total" column.
- Displaying data in the form of a scatter plot, data that has been imported and added columns Environment and Total are then displayed in the form of a scatter plot chart. The data taken is crime data within the scope of family and environment.

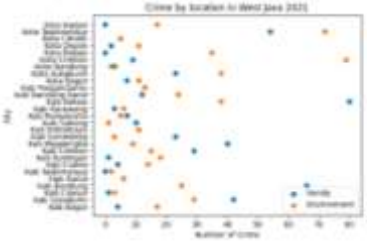


Fig. 7. Scatter Plot Dataset

- Data Training, then data on violence in the Family and Environment will be converted into an array to conduct variable training.
- Feature Scaling, Feature scaling is used to normalize data to be on the same scale of values with each other.
- Clustering, After feature scaling is complete, then proceed to the next stage, namely feature clustering. Clustering is done using K- Means with 2 clusters or n_clusters=2. Then display the y_cluster of x_train data or data that has been trained into the form of an array Then it is known that there are clusters 0 and 1 with integer data types. Then display the clustering result data.
- Find Centroid, find centroid point from data that has been previously entered into the program.
- Data Visualization, data that has been clustered and determined centroid points are then visualized in the

form of scatter plots with 2 centroid points and 2 clusters.

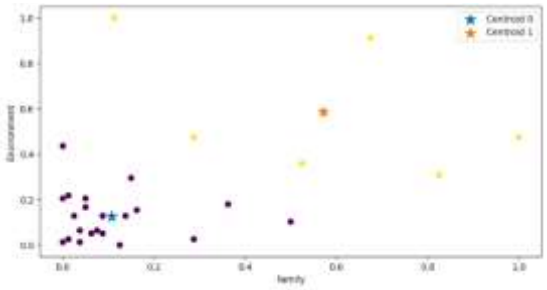


Fig. 8. Centroid and Cluster Scatter Plot

- Find for the Best Number of K, next is to find the best number of clusters (K) recommended to add. Data is taken from the value of k-means clustering and data from training or x_train then display the data in the form of a diagram. So it is known that the recommended number of clusters is 2, marked by a graph that angles the value 2.

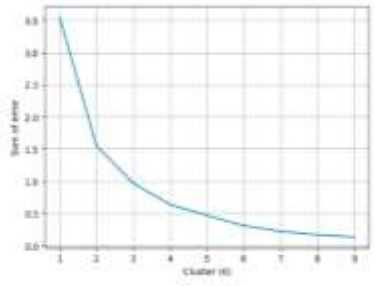


Fig. 9. Best Cluster Number

- Replace Label, the cluster label that previously contained only 0 and 1 was later changed to "Low" and "High"".

Fig. 10. Replace Label

- Displaying Data into Graphic Form, data that has been processed is then displayed in 2 different graphic forms, namely bars and lines.

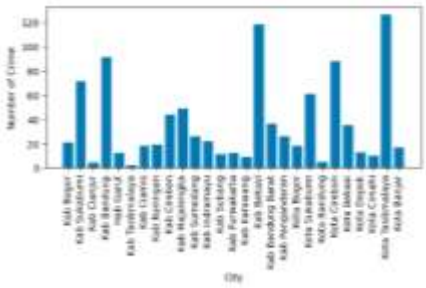


Fig. 11. Bar Graphic Form

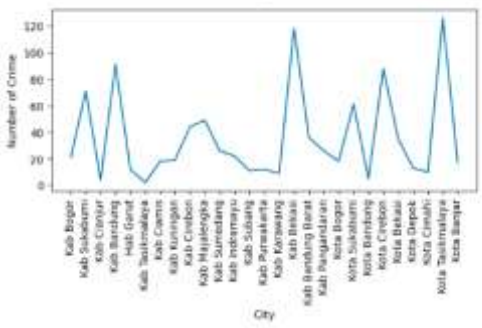


Fig. 12. Line Graphic Form

- Edit Data Export, the data is edited by selecting which columns to display and delete before being exported into an excel file.
- Export Data, export data into excel or xlsx files. And the new file will appear on the side panel of Google Colab with the name West Java Crime Case Clustering.xlsx. Data can be downloaded and opened using Microsoft Excel or similar applications.

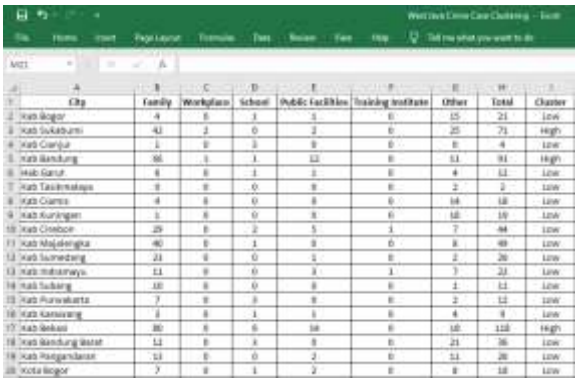


Fig. 13. Data Result in Excel

C. *Blackbox System Testing*

TABLE I. BLACKBOX TEST

Scenario	Blackbox System Testing			
	Test Case	***Expected Result***	***Test Results***	***Information***
Import dataset	Import data accordingly	System no errors and data can be imported	As expected	Valid
Read the crime dataset Jabar	Run the read dataset command	The dataset is well readable	As expected	Valid
Import dataset others	Import dataset number of traffic accidents	The dataset is unreadable and cannot enter the next process	As expected	Valid
Sum data and create new columns	Sum all columns and create a new column	Columns can be created and the amount of data is appropriate	As expected	Valid
Display data in the form of scatter plots	Run the command create scatter plot diagram	Scatter plot diagrams can form	As expected	Valid

Scenario	Blackbox System Testing			
	Test Case	***Expected Result***	***Test Results***	***Information***
	with domestic crime and neighborhood data			
Run training data	Run data training commands with Family and environmental crime data	No error messages on the program	As expected	Valid
Run feature scaling	Run the feature scaling command with the data that has been trained	The scale of the data changes to one equal scale	As expected	Valid
Clustering data	Run the k-means clustering command	Data can be grouped into 2 clusters	As expected	Valid
Find a centroid point	Run centroid point search command	Centroid point found	As expected	Valid
Data visualization of clustering results	Run the clustering data visualization command	Visualization comes out in the form of grouped scatter plots	As expected	Valid
Find and display the best K count	Run the inertias command	The best K amount can be found	As expected	Valid
Replace cluster labels with low and high labels	Run replace command	Cluster descriptions can change from 0 and 1 to low and high	As expected	Valid
Create bar and line charts	Run plt bar command and plot	Data graphs can be formed	As expected	Valid
Edit data before exporting	Choose which columns to import	Only the selected columns will be displayed when the data is exported	As expected	Valid
Export data into excel format	Run the export data to excel command	Data can be exported and downloaded to a computer	As expected	Valid

REFERENCES

[1] H. W. Bagas N, E. Mailoa, H, D, Purnomo, "Deteksi Buah untuk Klasifikasi Berdasarkan Jenis dengan Algoritma CNN Berbasis YOLOv3", J. RESTI (Rek. Sis. dan Tek. Inf.), vol . 4, no. 3, pp 476 - 481, 2020 doi: 10.29207/resti.v4i3.1868

[2] G. Gustientiedina, M. H. Adiya, and Y. Desnelita, "Penerapan Algoritma K-Means Untuk Clustering Data Obat-Obatan", J. Nas. Teknol. dan Sist. Inf., vol. 5, no. 1, pp. 17– 24, 2019, doi: 10.25077/teknosi.v5i1.2019.17-24.

[3] K. Fatmawati and A. P. Windarto, "Data Mining: Penerapan Rapidminer Dengan K- Means Cluster Pada Daerah Terjangkit Demam Berdarah Dengue (Dbd) Berdasarkan Provinsi", Comput. Eng. Sci. Syst. J., vol. 3, no. 2, p. 173, 2018, doi: 10.24114/cess.v3i2.9661.

[4] T. Santoso and E. A. Zulfa, "Kriminologi", Raja Grafindo Persada, Jakarta, 2003. Hal.21

[5] Restu, "Pengertian Kekerasan: Jenis, Ciri, Penyebab, dan Contoh", 2021, [Online]. Available: https://www.gramedia.com/literasi/pengertian-kekerasan/

[6] A. D. B. Raharja, "Machine Learning: Pengertian, Cara Kerja, dan 3 Metodenya", 2022, [Online]. Available: https://www.ekrut.com/media/apa-itu-machine-learning

[7] B. Tandika, "Bahasa Pemrograman Python: Yuk, Pelajari Arti, Fungsi, dan Keunggulannya", 2022, [Online]. Available: https://glints.com/id/lowongan/apa-itu- bahasa-pemrograman-python/

[8] O. D. Jabar, "Apa itu Open Data Jabar", 2022, [Online]. Available: https://opendata.jabarprov.go.id/id/tentang

[9] P. Purwoko, "Membuat Proyek Machine Learning dengan Python - Part 1", 2022, [Online]. Available: https://medium.com/easyread/membuat-proyek-machine-learning-dengan-python-part-1-8e8a03095636