Volume 2 Nomor 1 | bulan Juli tahun 2024 | halaman 1-7
Inside | Jurnal Sistem Informatika Cerdas
e-ISSN: 3031-4860

Clustering Analysis of Results of Students' Industrial Work Practice Activities Using the K-Means Method

Yulian Ansori[1], Mas'ud Hermansyah[2*], M. Faiz Firdausi[3], Abdul Wahid[4], Iqbal Sabilirrasyad[5]
[1]Universitas Primagraha, Serang, Indonesia
[2345]Institut Teknologi dan Sains Mandala, Jember, Indonesia

Corresponding Author:
Mas'ud Hermansyah, Institut Teknologi dan Sains Mandala, Jember, 68121, Indonesia
Email: masudhermansyah@itsm.ac.id

Abstract

Praktik Kerja Industri (PRAKERIN) is an important program in the Vocational High School curriculum that provides students with real work experience before they enter the industrial world. Apart from technical skills, this program also develops soft skills such as teamwork, communication, and problem solving. However, evaluating PRAKERIN results often faces challenges in identifying and classifying student performance objectively and systematically. With data mining technology, it is possible to analyze the results of students' abilities after participating in PRAKERIN activities. In conducting this research, the K-Means clustering method was used to group students' competency abilities. With the K-Means clustering technique, it is hoped that teachers can adjust the learning model according to students' abilities. Based on the grouping results, it was found that grouping with 3 clusters was the most optimal grouping result with the smallest Davies Bouldin Index (DBI) value, namely 0.160. The application of the K-Means method in grouping student data based on English language ability scores can produce 3 groups of students who are competent, quite competent, and less competent.

Keywords : Clustering, K-Means, Davies Bouldin Index

1 INTRODUCTION

Vocational education, which combines theoretical learning with industrial work practice, has an important role in preparing students to enter the world of work. One of the main components of vocational education is Praktik Kerja Industri (PRAKERIN), where students are deployed directly into a real work environment to apply the knowledge and skills they have learned at school. These activities not only provide practical experience, but also enable students to understand industry dynamics and increase their readiness to face professional challenges [1]. PRAKERIN is one of the important programs in the Vocational High School curriculum which aims to provide real work experience to students before they enter the industrial world. [2]. This program not only provides relevant technical skills, but also helps students develop soft skills such as teamwork, communication, and problem-solving. However, in the context of evaluating PRAKERIN results, challenges often arise in identifying and classifying student performance objectively and systematically [3].

SMK Lab Business School Tangerang, especially the Multimedia Department, has a large number of students taking part in the PRAKERIN program in various companies. The diversity of locations and types of work carried out by students adds complexity in evaluating and classifying PRAKERIN results. Therefore, an effective method is needed to group these results to get a clearer picture of student performance and development [4].

One method that can be used to overcome this problem is Data Mining with the K-Means clustering algorithm. Data Mining, or what is also known as Knowledge Discovery in Databases (KDD), is a method for extracting knowledge from large amounts of data in order to discover new patterns, thereby producing new knowledge and information. Data Mining technology is used to explore the knowledge contained in databases [5]. K-Means is a popular clustering algorithm in data analysis which functions to group data into several clusters based on similar characteristics [6]. It is hoped that the use of the K-Means method in analyzing PRAKERIN results can help identify groups of students with similar performance, making it easier to assess and design more targeted development programs. [3].

.

This research aims to analyze and classify the results of PRAKERIN activities for students at the Tangerang Lab Business School, Multimedia Department, using the K-Means method. By clustering, it is hoped that schools can gain deeper insight into variations in student performance and the factors that influence them. In addition, the clustering results can be used as a basis for decision making in improving the curriculum and teaching methods in the future [7].

Through this research, it is hoped that certain patterns can also be found that can be used as a reference for future students in preparing themselves for the PRAKERIN program. Thus, this research not only contributes to improving the quality of education at SMK Lab Business School Tangerang, but also helps students achieve optimal results in PRAKERIN and prepares them for successful careers in the multimedia industry.

2 RESEARCH METHOD

The research was carried out using the Cross-Industry Standard Data Mining Process (CRISP-DM) approach with research material in the form of data from the results of PRAKERIN activities for class XI students majoring in Multimedia in the form of soft files in xlsx format. The data consists of 4 (four) attributes that will be used, namely: (1) Systematic Report Writing, (2) Understanding of Majors, (3) Mastery of Employment, and (4) Final Presentation.

2.1 Data Mining

Data mining is a technique for digging up valuable information hidden in very large data sets, so that interesting patterns can be discovered that were previously unknown. Another definition states that data mining is the process of searching and analyzing large amounts of data with the aim of finding meaning from patterns and rules. Data mining techniques have been around for a long time and various algorithm theories have been widely discussed in the literature [8] [9]. Data mining is actually one of the stages in the process of searching for knowledge in a database known as Knowledge Discovery in Database (KDD). KDD involves integration techniques and scientific discovery. Interpretation and visualization of patterns produced by KDD data sets is a non-trivial process for searching for and identifying patterns in data, where these patterns must be valid, useful and understandable. This process includes data cleaning and data integration stages (cleaning and integration). This stage is used to eliminate inconsistent and noisy data from data obtained from various databases with different formats or platforms, which are then integrated into a data warehouse. [10].

Data mining involves selecting and transforming data in a database using various data mining processing techniques. Algorithms in data mining are very diverse, so it is important to choose an algorithm that suits the problem you want to solve. Searching for models and presenting knowledge in this process involves checking whether the resulting information model corresponds to reality or previously established rules. The final stage of KDD is implementing knowledge with a model that is easy for users to understand. Data mining has many uses, including:

1. Present a model that allows selecting sets of items with appropriate definitions.
2. Perform estimates and grouping, except that the estimated variable focuses more on calculations than grouping.
3. Separation is used to find patterns or describe and distinguish groups of set items, or guess groups of set items that do not yet have a group.
4. Separation for collecting records, monitoring, or viewing groups of subjects that have something in common. A cluster is a combination of records that have similarities and differences with records in other clusters.
5. Associations in data mining are symbols that exist at a certain time. In the business world, this is often referred to as an analysis platform.

The data mining process involves several interactive stages, where users can be involved directly or through a knowledge base. These stages include: Data Cleaning, Data Integration, Data Selection, Data Transformation, Mining Process, and Pattern Evaluation. [11]. These stages are interactive and often involve users directly or through knowledge base intermediaries.

2.2 Clustering

Clustering is a technique in data mining and machine learning that is used to group a set of objects or data into groups (clusters) based on the similarity or proximity between them. Objects in one cluster have characteristics that are more similar to each other than to objects in other clusters. The clustering process does not require previous data labels (unsupervised learning) and is often used to discover hidden structures in data.

The most widely used clustering method is the K-Means clustering method. The main weakness of this method is that the results are sensitive to the selection of initial cluster centers and the calculation of local solutions to achieve optimal conditions. Cluster analysis is a multivariate technique which has the main aim of grouping objects based on their characteristics. Cluster analysis classifies objects so that each object that is most closely similar to another object is in the same cluster [12].

2.3 Algoritma K-Means

The K-Means algorithm is a group analysis method that focuses on partitioning. In this method, N observation objects are divided into K groups or clusters, where each observation object has a group with an average or mean. The K-Means algorithm is part of the application of data mining clustering which is used to group data into several groups. These groups are formed based on certain criteria, and data corresponding to the group is collected into one cluster. Each cluster has a center point or centroid. The following are the stages of the K-Means algorithm [13]:

1. Choose how many (k) Clusters are expected in the dataset
2. Randomly select a Centroid
3. Calculate the shortest distance for each data with Centroid. Use the Euclidean distance(d) formula to calculate the shortest distance to the centroid. 2 (two) points that are not the same can be measured using this Euclidean distance method [14]. The formula is:

   $$de = \sqrt{(xi - si)^2 + (yi - ti)^2} \tag{1}$$

   Information:

   (x,y) = object coordinates

   (s, t) = centroid coordinates

   i = many objects
4. Recalculate Cluster points with the latest Cluster membership. The average of all data in the cluster is the cluster center. The formula for calculating this is:

   $$V_{ij} = \frac{1}{Ni} = \sum_{k=0}^{Ni} Xkj \tag{2}$$

   Information:

   Vij = Average centroid in the i-th Cluster for the j-th variable

   Ni = Number of members of the ith Cluster

   i, k = Index of Clusters

   j = Index variable

   Xkj = The kth data value of the jth variable for the Cluster
5. A new cluster (new Centroid) is used to recalculate each object. This stage is the initial opening of a new iteration. If the cluster still has members moving, then return to step c. If cluster members do not move clusters again, the clustering process is complete [15].

2.4 RapidMiner

RapidMiner is a software platform designed for data analysis, data mining, machine learning, and predictive analytics. RapidMiner provides a variety of tools and functions that allow users to crunch data, build models, and perform in-depth analysis without having to write code manually. Here are some of the main features and uses of RapidMiner:

1. Visualization Tools: RapidMiner offers an intuitive graphical user interface (GUI), allowing users to drag-and-drop components to create analytical workflows.
2. Data Integration: The platform can integrate data from various sources, including databases, spreadsheets, and cloud services, making it easy for users to access and combine the necessary data.

3. Data Pre-processing: RapidMiner provides various tools for data cleaning, transformation, and feature selection, which are important for ensuring high data quality before further analysis.
4. Machine Learning Algorithms: Supports a variety of machine learning algorithms, both for classification, regression, clustering, and association, allowing users to build powerful predictive models.
5. Model Evaluation: RapidMiner has tools for evaluating and validating models that have been built, including cross-validation and model performance analysis.
6. Model Deployment: Once the model is built and validated, RapidMiner allows users to implement the model in a production environment for use in real applications.

2.5 Davies–Bouldin Index

Clustering evaluation aims to assess how good the quality of the clustering results is. In this research, the evaluation method used is the Davies Bouldin Index to determine the most optimal number of clusters. The Davies Bouldin Index (DBI), introduced by David L. Davies and Donald W. Bouldin in 1979, is a method used to measure the validity or optimal number of clusters in a grouping method. In DBI, cohesion is defined as the closeness of data to the center point of the cluster it follows. Evaluation with the Davies Bouldin Index uses an internal cluster evaluation scheme, where the quality of the clustering results is seen from the quantity and closeness between data in the cluster [16]. To achieve good clustering results, inter-cluster distance should be high and intra-cluster distance should be low and therefore, lower DBI values are required to show good clustering results.

3 RESULTS AND ANALYSIS

In this research, the steps taken include collecting a dataset, selecting relevant attributes, building a clustering model using the K-Means method for grouping data, and evaluating the model using the Davies Bouldin Index.

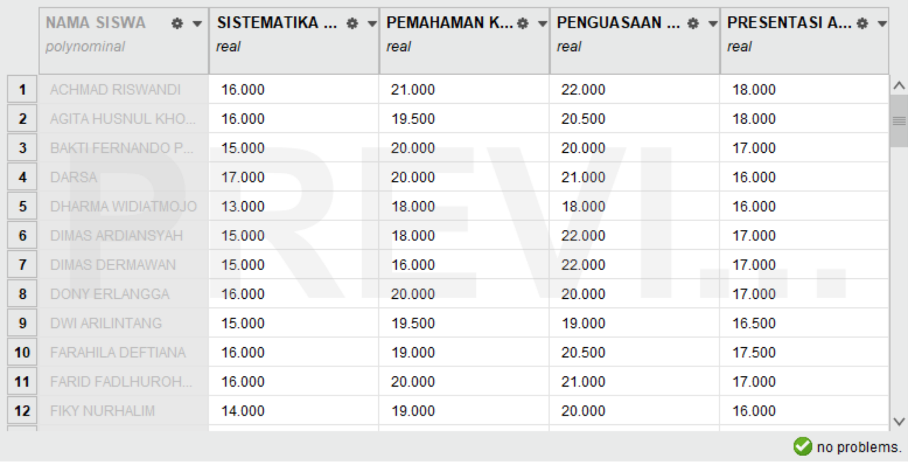
3.1 Dataset

The dataset used in this clustering process comes from 2023 students at the Tangerang Vocational School Lab Business School, Department of Multimedia, with a total of 72 student data. Each student is assessed based on 4 main criteria: systematic report writing, understanding of the major, mastery of the field of work, and final presentation. This dataset is then processed to form groups that have similar characteristics using the clustering method. These criteria were chosen because they were considered to represent important aspects in assessing student performance and abilities in the multimedia field.

3.2 Preprocessing Data

Cleaning aims to reduce noise which can affect calculations in data analysis. Noise in data can be inconsistencies, errors, or missing values that can cause analysis results to be inaccurate or biased. Therefore, the cleaning process is a crucial step before the data is used for further analysis. In this cleansing process, various techniques such as deduplication removal, filling of missing values, and data error correction are applied to ensure the data is clean and ready to use.

In the data cleansing process carried out on the student dataset of SMK Lab Business School Tangerang, Multimedia Department, it was found that there were no null data entries. This means all data entries are complete and there are no missing values. Thus, all data from 72 students can be used in analysis without the need for imputation or data deletion. The existence of complete data facilitates the clustering process and increases the reliability of analysis results, because each student's data makes a full contribution to the formation of clusters. The preprocessing data can be seen in Figure 1.



	NAMA SISWA polynominal	SISTEMATIKA ... real	PEMAHAMAN K... real	PENGUASAAN ... real	PRESENTASI A... real
1	ACHMAD RISWANDI	16.000	21.000	22.000	18.000
2	AGITA HUSNUL KHO...	16.000	19.500	20.500	18.000
3	BAKTI FERNANDO P...	15.000	20.000	20.000	17.000
4	DARSA	17.000	20.000	21.000	16.000
5	DHARMA WIDIATMOJO	13.000	18.000	18.000	16.000
6	DIMAS ARDIANSYAH	15.000	18.000	22.000	17.000
7	DIMAS DERMAWAN	15.000	16.000	22.000	17.000
8	DONY ERLANGGA	16.000	20.000	20.000	17.000
9	DWI ARILINTANG	15.000	19.500	19.000	16.500
10	FARAHILA DEFTIANA	16.000	19.000	20.500	17.500
11	FARID FADLHUROH...	16.000	20.000	21.000	17.000
12	FIKY NURHALIM	14.000	19.000	20.000	16.000

✓ no problems.

Figure 1. Data Preprocessing

3.3 Modelling

At this stage, modeling was carried out using the RapidMiner tool with the K-Means method. The model used adds the Replace Missing Values operator to delete data that has missing or null values and the Performance operator to calculate the Index Davies Bouldin (IDB) value. The smallest value of IDB indicates the most optimal number of clusters.

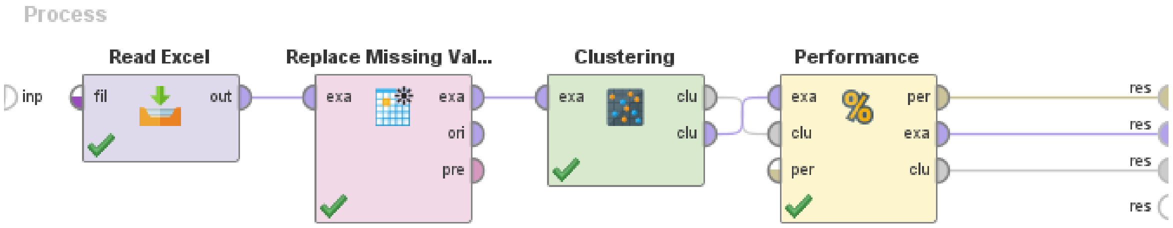


Figure 2. Clustering Model with RapidMiner

The steps and stages in the modeling and evaluation process can be seen in Figure 2 as follows:

1. Create a clustering model using the K-Means method where the number of clusters modeled is 2 – 10. The clustering method uses the K-Means method.
2. Each cluster created is evaluated with the Cluster Distance Performance operator to determine the DBI value of each cluster.
3. The smallest DBI value shows the best results and shows the optimal number of clusters.

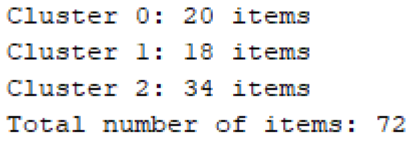
3.4 Evaluasi

The data processing process is carried out by modeling using the K-Means method, where the smallest DBI value is sought to determine the optimization of the number of clusters. The number of clusters and their DBI values are shown in Table 1 below:

Table 1. DBI Values in Each Cluster

Cluster	DBI Values
2	0,338
3	0,160
4	0,186
5	0,214
6	0,349

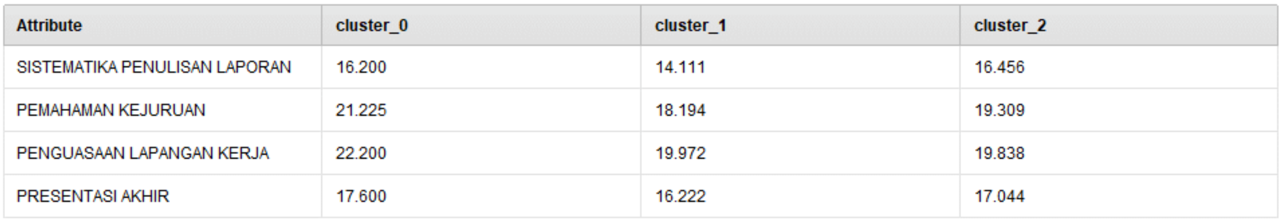
From the processing results, it is known that for optimization the number of clusters is 3 with a DBI value of 0.160. So for this data, the grouping carried out by optimizing the number of clusters is 3. The number of members of each cluster is shown in Figure 3 below:



```
Cluster 0: 20 items
Cluster 1: 18 items
Cluster 2: 34 items
Total number of items: 72
```

Gambar 3. Jumlah Anggota Tiap Cluster

From the picture above it can be explained that cluster 0 has 20 students, cluster 1 has 18 students and cluster 2 has 34 students. And the centroid value for each cluster is shown in Figure 4 as below:



Attribute	cluster_0	cluster_1	cluster_2
SISTEMATIKA PENULISAN LAPORAN	16.200	14.111	16.456
PEMAHAMAN KEJURUAN	21.225	18.194	19.309
PENGUASAAN LAPANGAN KERJA	22.200	19.972	19.838
PRESENTASI AKHIR	17.600	16.222	17.044

Figure 4. Centroid Value

From the values in the table above, it is known that each cluster for student categories is based on the potential results of PRAKERIN activities, namely that existing students are divided into:

1. Cluster 0: the category of competitors in their PRAKERIN activities
2. Cluster 1: die categorie wat kompeteer vir die PRAKERIN
3. Cluster 2: the competition category for PRAKERIN competitors

4 CONCLUSION

The application of the K-Means method in grouping students' PRAKERIN results data resulted in three groups: competent, moderately competent and less competent students. The results of this clustering provide a clear picture of students' ability levels based on predetermined criteria. With this grouping, school teachers can more easily plan and organize learning strategies and practicums that suit the needs of each group. For example, students who fall into the less competent group can be given additional guidance, while competent students can be given higher challenges to develop their abilities further.

Based on the grouping results, it was found that dividing student data into three clusters was the most optimal result. This is shown by the smallest Davies Bouldin Index (DBI) value, namely 0.160. A low DBI value indicates that the cluster formed has a good level of cohesion within the cluster and clear separation between different clusters. In other words, the K-Means method successfully groups students in an efficient and effective way, providing a strong basis for teachers to take appropriate pedagogical steps based on accurate and structured data.

REFERENCES

[1] J. B. Sukoco, N. I. Kurniawati, R. E. Werdani, and A. Windriya, “Pemahaman Pendidikan Vokasi Di Jenjang Pendidikan Tinggi Bagi Masyarakat,” *J. Pengabdi. Vokasi*, vol. 01, no. 01, pp. 23–26, 2019.

[2] J. F. Dinita, K. Setyaningsih, and R. Kanada, “Pelaksanaan Praktik Kerja Industri (Prakerin) Bagi Siswa Jurusan Bisnis Daring & Pemasaran di SMK Negeri 3 Palembang,” *J. Law, Adm. Soc. Sci.*, vol. 4, no. 4, pp. 544–555, 2024, doi: 10.54957/jolas.v4i4.832.

[3] R. Fahrozi, A. M. Siregar, and T. Al Mudzakir, “Clustering Penempatan Praktek Kerja Lapangan Siswa Sekolah Menengah Kejuruan TI Muhammadiyah Cikampek Menggunkan Algoritma K-Means dan Algoritma Topsis,” *Sci. Student J. Information, Technol. Sci.*, vol. 2, no. 1, pp. 237–246, 2021.

[4] B. Hasmaulina, M. Maryaningsih, and S. Sapri, “Penerapan Data Mining Untuk Membentuk Kelompok

Belajar Menggunakan Metode Clustering Di SMK Negeri 3 Seluma," *JUKOMIKA (Jurnal Ilmu Komput. dan Inform.*, vol. 4, no. 2, pp. 57–71, 2022, doi: 10.54650/jukomika.v4i2.368.

[5] Y. Elda, S. Defit, Y. Yunus, and R. Syaljumairi, "Klasterisasi Penempatan Siswa yang Optimal untuk Meningkatkan Nilai Rata-Rata Kelas Menggunakan K-Means," *J. Inf. dan Teknol.*, vol. 3, pp. 103–108, 2021, doi: 10.37034/jidt.v3i3.130.

[6] J. Butarbutar, N. B. Nugroho, and W. R. Maya, "Implementasi Data Mining Untuk Mengelompokkan Daerah Rawan Tindakan Kriminal Di Kota Medan Menggunakan Metode K-Means," *J. CyberTech*, vol. 4, no. 3, pp. 1–12, 2021, [Online]. Available: https://ojs.trigunadharma.ac.id/

[7] S. Anwar, T. Suprapti, G. Dwilestari, and I. Ali, "Pengelompokkan Hasil Belajar Siswa dengan Metode Clustering K-Means," *JURSISTEKNI (Jurnal Sist. Inf. dan Teknol. Informasi)*, vol. 4, no. 2, pp. 60–72, 2022.

[8] S. Sumiyatun, Y. Cahyadi, and E. Iskandar, "Data Mining Untuk Memprediksi Status Kelulusan Mahasiswa," *J. Inform. Komputer, Bisnis dan Manaj.*, vol. 21, no. 3, pp. 11–19, 2023, doi: 10.61805/fahma.v21i3.3.

[9] P. M. S. Tarigan, J. T. Hardinata, H. Qurniawan, M. Safii, and R. Winanjaya, "Implementasi Data Mining Menggunakan Algoritma Apriori Dalam Menentukan Persediaan Barang (Studi Kasus: Toko Sinar Harahap)," *G-Tech J. Teknol. Terap.*, vol. 7, no. 1, pp. 119–126, 2023, doi: 10.33379/gtech.v7i1.1938.

[10] I. Zulfa, R. Rayuwati, and K. Koko, "Implementasi Data Mining untuk Menentukan Strategi Penjualan Buku Bekas dengan Pola Pembelian Konsumen Menggunakan Metode Apriori," *Tek. J. Sains dan Teknol.*, vol. 16, no. 1, p. 69, 2020, doi: 10.36055/tjst.v16i1.7601.

[11] K. F. Mauladi and P. H. Susilo, "Klasterisasi Virus Covid-19 Di Wilayah Kabupaten Lamongan Dengan Metode K-Means Clustering," *JIPI (Jurnal Ilm. Penelit. dan Pembelajaran Inform.*, vol. 6, no. 2, pp. 325–335, 2021, doi: 10.29100/jipi.v6i2.1999.

[12] R. Kurniawan, M. M. M. Mukarrobin, and M. Mahradianur, "Klasterisasi Tingkat Pendidikan Di Dki Jakarta Pada Tingkat Kecamatan Menggunakan Algoritma K-Means," *Technol. J. Ilm.*, vol. 12, no. 4, p. 234, 2021, doi: 10.31602/tji.v12i4.5633.

[13] M. Hermansyah, N. A. Prasetyo, Y. Ansori, M. F. FIrdausi, and A. Wahid, "Implementation of K-Means Clustering for Analysis Students English Proficiency," *J. Educ. Sci. Technol.*, vol. 1, no. 6, pp. 31–35, 2023.

[14] A. Aditya, I. Jovian, and B. N. Sari, "Implementasi K-Means Clustering Ujian Nasional Sekolah Menengah Pertama di Indonesia Tahun 2018/2019," *J. Media Inform. Budidarma*, vol. 4, no. 1, p. 51, 2020, doi: 10.30865/mib.v4i1.1784.

[15] L. Fimawahib and E. Rouza, "Penerapan K-Means Clustering pada Penentuan Jenis Pembelajaran di Universitas Pasir Pengaraian," *INOVTEK Polbeng - Seri Inform.*, vol. 6, no. 2, p. 234, 2021, doi: 10.35314/isi.v6i2.2096.

[16] E. Muningsih, I. Maryani, and V. R. Handayani, "Penerapan Metode K-Means dan Optimasi Jumlah Cluster dengan Index Davies Bouldin untuk Clustering Propinsi Berdasarkan Potensi Desa," *J. Sains dan Manaj.*, vol. 9, no. 1, p. 96, 2021, [Online]. Available: www.bps.go.id