JOURNAL OF INFORMATION TECHNOLOGY AND ITS UTILIZATION, VOLUME 7, ISSUE 2, DECEMBER 2024
EISSN 2654-802X ; PISSN 2985-4067
DOI: https://doi.org/10.56873/JITU.7.1.5854. SUBMITTED: OCTOBER 9, 2024; REVISED: NOVEMBER 1, 2024; ACCEPTED: NOVEMBER 13, 2024

Clustering the Happiness Index of Provincials in Indonesia using K-MEANS

Heti Mulyani[1*], Ricak Agus Setiawan[2]

[1,2]Politeknik Enjinering Indorama, Purwakarta, Indonesia
heti.mulyani@pei.ac.id[1*], ricak.agus@pei.ac.id[2]
*Corresponding author

***Abstract*--Community welfare is a government goal related to the fulfillment of basic needs, education and employment, which can be measured through the happiness index. The purpose of this research is to cluster provinces in Indonesia based on their resident's happiness level. The data obtained from the Indonesian Central Bureau of Statistics website. The method used in this research is K-means clustering. There are 2 dimensions used, namely the personal dimension which includes education, employment, household income, health, housing conditions and, assets. The social dimension includes social relations, environmental conditions, security conditions, family harmony, and availability of free time. Based on the results of the study, 2 provincial groups were obtained based on the level of happiness. Testing is done using the Davies Bouldin Index (DBI). The optimal K is obtained, namely K = 2 with a DBI value of = 0.776. The first group is the happiest group including the provinces of North Maluku, Maluku, North Sulawesi, North Kalimantan, Gorontalo, Central Sulawesi, West Papua, Riau Islands, East Kalimantan. The other provinces are in the second group. The unhappiest groups are Banten, Bengkulu and Papua.**

Keywords: Happiness Level, Cluster, Province, KMeans.

I. INTRODUCTION

Community welfare can be defined as a situation when a person is able to fulfill his basic needs (clothing, food, and shelter), including the opportunity to get an education and get adequate work to improve the quality of life so that he can have a social status equal to the average community in a certain area [1]. Indonesia, with its cultural, social and economic diversity, presents unique challenges in measuring happiness levels in each of its provinces. Happiness levels are an important indicator in assessing people's well-being and can be influenced by a variety of factors such as income, health, education, environment and social relationships. To understand the variation in happiness among Indonesia's provinces, an analytical approach is needed that is able to categorize provinces based on their happiness characteristics. In addition to being a comprehensive measure of societal well-being, the happiness index also provides a more accurate picture of the quality of life of the population [2].

Several studies related to happiness measurement have been conducted by previous researchers, including Modeling the Provincial Happiness Index in Indonesia Using Spline Truncated Nonparametric Regression. In this study, the province with the lowest level of happiness was Papua with a value of 67.52 [3]. The next research is a comparison of the k-means algorithm with k-medoids in clustering the happiness level of provinces in Indonesia, in this study the variables used are satisfaction index, feelings and meaning of life and obtained 2 clustering [4]. Further research with the title of clustering provinces in Indonesia according to happiness index indicators using the average linkage method, in this study grouping provinces into 4 areas with an accuracy value of 53.2% [5].

In previous studies, happiness index measurements were made based on the dimensions of life satisfaction, social and feelings, but no one has measured using the personal dimension and the social dimension. Based on the above background, this study will analyze and cluster the measurement of provincial happiness levels in Indonesia based on 2 dimensions, namely the personal dimension which includes education, employment, household income, health, housing conditions, and assets. The social dimension includes social relationships, environmental conditions, security conditions, family harmony, and availability of free time. The technique used is K-Means clustering [6][7].

The purpose of this research is to cluster the happiness rate of provincials in Indonesia. Data is

EISSN 2654-802X ; PISSN 2985-4067
DOI: https://doi.org/10.56873/JITU.7.1.5854. SUBMITTED: OCTOBER 9, 2024; REVISED: NOVEMBER 1, 2024; ACCEPTED: NOVEMBER 13, 2024

taken from the Central Bureau of Statistics website. The data used is 2021 data. With this research, it is hoped that it will be able to become a basis for the government in making policies and making more effective decisions for the progress of the country. Not only that, the happiness index figure is also expected to help the government measure the impact of a policy on aspects of life in society and ensure that the quality of life can be better and sustainable in the long run.

II. METHOD

The method used in this research is CRISP-DM, with the stages of business understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment. The CRISP-DM method can be seen in Fig. 1 [8].

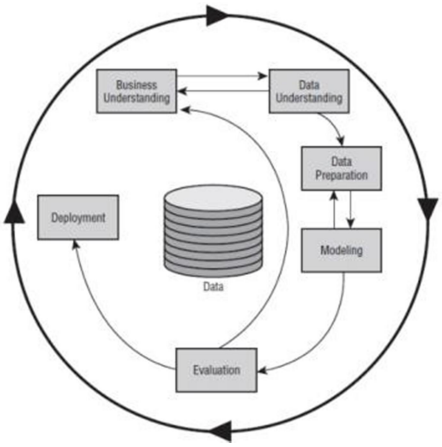

Fig. 1. CRISP-DM Cycle [7]

The following is an explanation of the CRISP-DM Framework:

A. Business Understanding

Business understanding is to understand what the client really wants to achieve from a business perspective, determine the project objectives and necessities in detail within the business scope. Then, the knowledge is translated into a data mining problem. Indonesia is a country that has many provinces with different levels of happiness. So it needs to be grouped so that the government can find out which provinces need attention and service improvement.

B. Data Understanding

Data understanding is the phase to identify and collect relevant data for the project, including what data is existed, how it is gathered, and how it can be used to reach the business goals. The data was taken from Badan Pusat Statistik or Central Bureau of Statistics (BPS). It contains two dimensions: personal and social. The personal dimension contains 5 subdimensions, namely: education, employment, household income, health, housing conditions, and assets. While the social dimension has 5 subdimensions, namely: social relations, environmental conditions, security conditions, family harmony, availability of free time.

C. Data Preparation

Data preparation is the phase to clean, integrates, and prepares the data for analysis. This process includes processing missing data, dealing with outliers, and converting the data format if necessary.

D. Modelling

Using certain models, such as regression, classification, clustering, or other machine learning models to develop data analysis and evaluation is known as the modeling step [8]. The steps for the K-Means method can be described as follows [9]:

1. Specify the number of groups (K).
2. Calculate the group center or centroid value of the data in each group.
$$x = \frac{1}{\mathrm{M}} = \sum_{j=1}^{M} Xj \quad (1)$$
(M is the amount of data in a group.)
3. Allocate each data to the nearest centroid/average using the Euclidian below:
$$d = \sqrt{\sum xi^2 - yi^2} \quad (2)$$
4. Group the data according to the closest distance.
5. Update the centroid value using:
$$ck = \frac{1}{nk} \sum di \quad (3)$$
6. Repeat steps 2 to 5 until the members of each cluster have not changed.

E. Evaluation

The Davies-Bouldin index approach is used for evaluation in this study. Whether the cluster results are evident from the quantity and proximity of the cluster result data, evaluation using DBI has an internal cluster evaluation method. The goal of the DBI measuring method is to reduce the intra-cluster distance while increasing the inter-cluster distance. The best

EISSN 2654-802X ; PISSN 2985-4067
DOI: https://doi.org/10.56873/JITU.7.1.5854. SUBMITTED: OCTOBER 9, 2024; REVISED: NOVEMBER 1, 2024; ACCEPTED: NOVEMBER 13, 2024

cluster architecture is indicated by a smaller DBI value [10]. The steps to calculate the Davies-Bouldin Index are as follows [11]:

1. Formulate the sum of squares within clusters (SSW) as a cohesiveness measure in an i-th cluster using the following formula:

$$SSWi = \frac{1}{mi}\sum_{j=1}^{mi} d(\mathrm{xj, ci}) \quad (4)$$

mi is the number of data in the i-th cluster, and $d\ (x, c)$ is the distance of data x to centroid c.

2. Calculate the sum of squares between clusters (SSB) by calculating the separation between centroids ci and cj using the formula below:

$$SSBi, j = d(\mathrm{ci, cj}) \quad (5)$$

(, cj) is the distance between centroid ci and centroid cj.

3. Find Rij, a metric that expresses how well the comparison value between the i-th and j-th clusters is. The cohesion and dissociation components provide the value. A cluster with the maximum spacing and the least amount of cohesiveness is considered good.

$$Rij = \frac{SSWi+SSW}{SSBi,j} \quad (6)$$

4. Calculate the Davies Bouldin Index (DBI) using formula:

$$DBI = \frac{1}{K}\sum_{j=1}^{K} max(Rij) \quad (7)$$

F. Deployment

Implement the model by developing *dashboards* or other tools to monitor the performance of the model and create reports and visualizations that are easy to understand. In this research, a dashboard of cluster results will be built to make it easier to see data visualization.

III. Result and Discussion

The next step in this research is to explain about the results and discussion. In this research, it follows the CRSIP-DM and data is taken from the Central Bureau of Statistics (BPS) source. It consists of the provincial happiness index in 2021 from 34 provinces, with measurements using 2 dimensions mentioned above. Table I shows the dataset of the results of measuring the happiness dimensions of the provincials in Indonesia. For clarity, this study shows an exploration of the data obtained (provincial data based on personal level) in Fig. 2.

TABLE I
Dataset of Happiness Dimensions of Provinces in Indonesia

No.	Province	Education	...	Total Social	End Total
1.	Aceh	62.99	…	81.01	75.50
2.	North Sumatera	61.57	…	79.44	74.24
3.	West Sumatera	62.56	…	79.34	74.49
4.	Riau	61.00	…	80.95	75.58
5.	Jambi	64.29	…	81.40	76.70
6.	South Sumatera	62.06	…	80.68	75.32
7.	Bengkulu	58.07	…	78.40	72.83
8.	Lampung	62.44	…	80.22	74.98
9.	Bangka Belitung Islands	62.07	…	81.62	76.44
10.	Riau Islands	68.54	…	81.80	77.72
11.	DKI Jakarta	68.67	…	78.15	75.25
12.	West Java	61.88	…	79.13	74.17
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
34.	Papua	61.89	…	77.50	73.23
35.	Indonesia	62.79	…	80.07	75.16

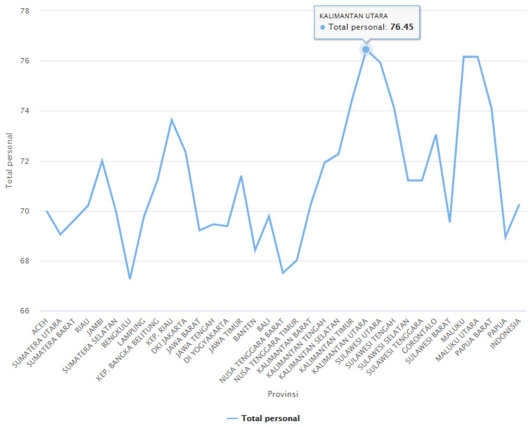

Fig. 2. Data visualization by personal level

Based on Fig. 2, it can be seen that the highest life satisfaction based on the personal dimension is in North Kalimantan. Meanwhile, the lowest personal level happiness is in Bengkulu Province. The social level happiness can be seen in Fig. 3. It is obtained that the highest social level of life satisfaction is in North Maluku province. Meanwhile, the lowest is in Banten Province.

The next step is data preparation. At this stage, the data cleaning process is carried out. For provinces, there is still the name Indonesia, so this data needs to be deleted, because the provinces in Indonesia only include 34 provinces. Fig. 4 shows the distribution of provincial happiness data in Indonesia.

EISSN 2654-802X ; PISSN 2985-4067
DOI: https://doi.org/10.56873/JITU.7.1.5854. SUBMITTED: OCTOBER 9, 2024; REVISED: NOVEMBER 1, 2024; ACCEPTED: NOVEMBER 13, 2024

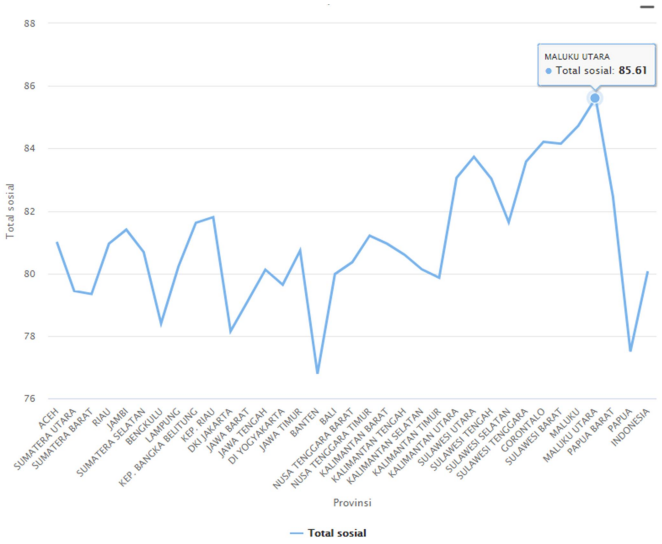

Fig. 3. Data visualization by social level

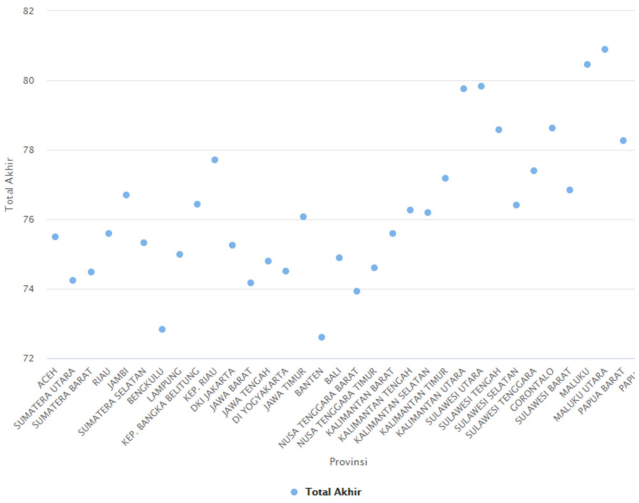

Fig. 4. Indonesian happiness data distribution

The next step is modelling data. The model used for clustering in this study uses the K-Means technique. Determination of the K value is based on the Elbow technique [12][13], where the most optimal K value is K = 2. The application used to analyze K-Means is rapidminer [14][15]. The results for data clustering can be seen in Fig. 6.

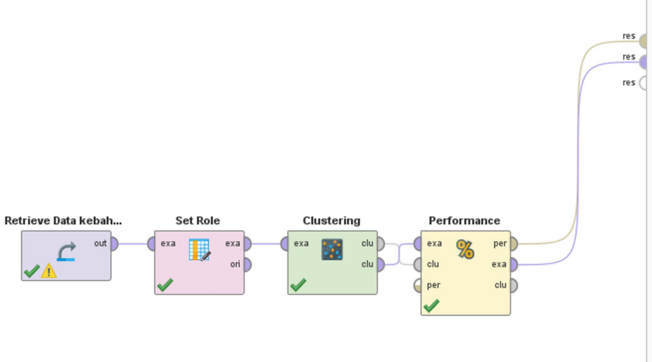

Fig. 5. K-Means using Rapidminer

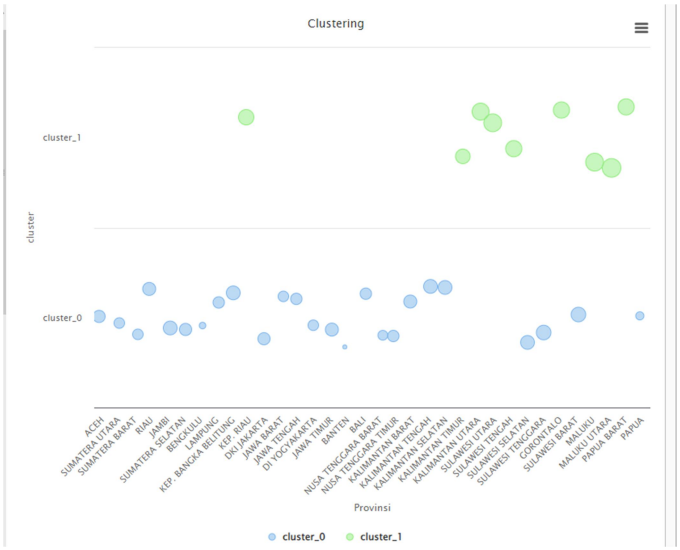

Fig. 6. Data clustering results with K-Means

Based on Fig. 6, it can be seen that the cluster results 2 groups of provinces; the first group consists of North Maluku, Maluku, North Sulawesi, North Kalimantan, Gorontalo, Central Sulawesi, West Papua, Riau Islands, and East Kalimantan. While the second group are Banten, Bengkulu, Papua, West Nusa Tenggara, West Java, North Sumatra, West Sumatra, Yogyakarta Special Province (DI Yogyakarta), East Nusa Tenggara, Central Java, Bali, Lampung, DKI Jakarta, South Sumatra, Aceh, Riau, West Kalimantan, East Java, South Kalimantan, Central Kalimantan, South Sulawesi, Bangka Belitung Islands, Jambi, West Sulawesi, Southeast Sulawesi.

The next step is evaluation model. Evaluation of the K-means model in this study was carried out by calculating the Davies Bouldin Index value. The results of DBI measurement with rapidminer can be seen in Fig. 7. Based on it, the DBI value is -0.776. The smaller the DBI value is close to 0, the better the cluster results. Because the DBI value is -0.776, the cluster results are quite good.

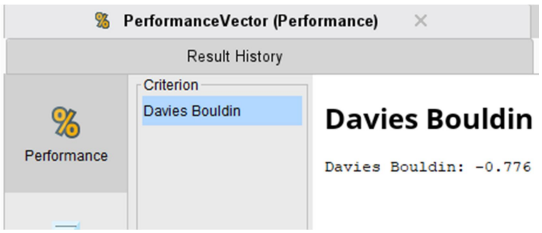

Fig. 7. Evaluation of K-Means with DBI

The last step is to build a dashboard to display the cluster result data. Dashboard can be seen in Fig. 8. It shows cluster result data based on social,

EISSN 2654-802X ; PISSN 2985-4067
DOI: https://doi.org/10.56873/JITU.7.1.5854. SUBMITTED: OCTOBER 9, 2024; REVISED: NOVEMBER 1, 2024; ACCEPTED: NOVEMBER 13, 2024

personal, education, and total clusters. The dashboard can display data for each cluster and the entire cluster.

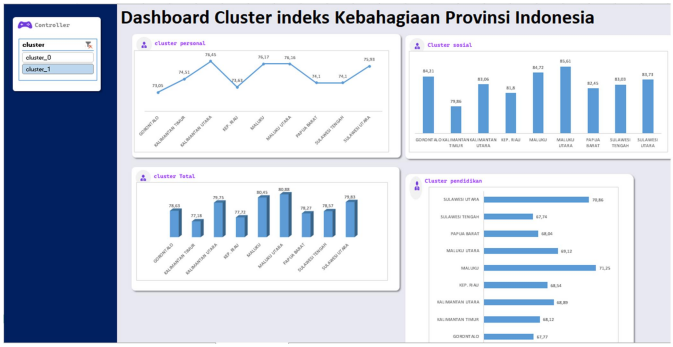

Fig. 8. Happiness index dashboard for Indonesia province

IV. Conclusion

Based on the results of this study, it is found that the provincial happiness cluster in Indonesia is divided into 2 clusters. Of the 34 provinces obtained, 9 provinces belong to cluster 1, including North Maluku, Maluku, North Sulawesi, North Kalimantan, Gorontalo, Central Sulawesi, West Papua, Riau Islands, East Kalimantan. For cluster 2, there are 25 provinces including Banten, Bengkulu, Papua, West Nusa Tenggara, West Java, North Sumatra, West Sumatra, DI Yogyakarta, East Nusa Tenggara, Central Java, Bali, Lampung, DKI Jakarta, South Sumatra, Aceh, Riau, West Kalimantan, East Java, South Kalimantan, Central Kalimantan, South Sulawesi, Bangka Belitung Islands, Jambi, West Sulawesi. The cluster evaluation results show a Davies Bouldin Index value of -0.776, indicating that the cluster results are quite good.

V. References

[1] K. Hamidah and A. Voutama, "Analisis Faktor Tingkat Kebahagiaan Negara Menggunakan Data World Happiness Report dengan Metode Regresi Linier," Explor. IT J. Keilmuan dan Apl. Tek. Inform., vol. 15, no. 1, pp. 1–7, 2023, doi: 10.35891/explorit.v15i1.3874.

[2] F. O. Dayera and M. B. Palungan, "G-Tech : Jurnal Teknologi Terapan," G-Tech J. Teknol. Terap., vol. 8, no. 1, pp. 186–195, 2024, [Online]. Available: https://ejournal.uniramalang.ac.id/index.php/g-tech/article/view/1823/1229

[3] A. F. D. Rositawati and I. N. Budiantara, "Pemodelan Indeks Kebahagiaan Provinsi di Indonesia Menggunakan Regresi Nonparametrik Spline Truncated," J. Sains dan Seni ITS, vol. 8, no. 2, 2020, doi: 10.12962/j23373520.v8i2.45160.

[4] C. F. Palembang, M. Y. Matdoan, and S. P. Palembang, "Perbandingan Algoritma K-Means dan K-Medoids dalam Pengelompokkan Tingkat Kebahagiaan Provinsi di Indonesia," J. Multidisiplin Ilmu, vol. 01, no. 5, pp. 830–839, 2022, [Online]. Available: https://journal.mediapublikasi.id/index.php/bullet/article/download/1135/550

[5] N. Wayan, R. Damayanthi, N. Luh, P. Suciptawati, K. Jayanegara, and E. N. Kencana, "Pengelompokan Provinsi di Indonesia Menurut Indikator Indeks Kebahagiaan Menggunakan Metode Average Linkage," Innov. J. Soc. Sci. Res., vol. 3, pp. 8859–8872, 2023.

[6] S. N. Mayasari and J. Nugraha, "Implementasi K-Means Cluster Analysis untuk Mengelompokkan Kabupaten/Kota Berdasarkan Data Kemiskinan di Provinsi Jawa Tengah Tahun 2022," KONSTELASI Konvergensi Teknol. dan Sist. Inf., vol. 3, no. 2, pp. 317–329, 2023, doi: 10.24002/konstelasi.v3i2.7200.

[7] F. U. Fitra Ramdhani, "Pengelompokan Provinsi di Indonesia Berdasarkan Karakteristik Kesejahteraan Rakyat Menggunakan Metode K-Means Cluster," GAUSSIAN, vol. 4, no. 2015, pp. 875–884, 2015.

[8] U. N. U. Y. Yudiana, "Prediksi Customer Churn Menggunakan Metode CRISP-DM pada Industri Telekomunikasi sebagai Implementasi Mempertahankan Pelanggan," J. Pengemb. Teknol. Inf. dan Ilmu Komput., vol. 4, no. 10, pp. 3571–3579, 2020.

[9] L. Y. Hutabarat, I. Gunawan, I. Purnamasari, M. Safii, and W. Saputra, "Penerapan Algoritma K-Means Dalam Pengelompokan Jumlah Penduduk Berdasarkan Kelurahan Di Kota Pematangsiantar", IKOMTI (Jurnal Ilmu Komputer dan Teknologi), vol. 2, no. 2, pp. 20–26, 2021.

[10] F. Sandova, R. Kurniawan, and T. Supratati, "Penerapan Data Mining Menggunakan Metode K-Means Clustering pada Penjualan Tas di Asia Toserba Cirebon," JATI (Jurnal Mhs. Tek. Inform.), vol. 8, no. 1, pp. 245–251, 2024, doi: 10.36040/jati.v8i1.8330.

[11] Y. Sopyan, A. D. Lesmana, and C. Juliane, "Analisis Algoritma K-Means dan Davies Bouldin Index dalam Mencari Cluster Terbaik Kasus Perceraian di Kabupaten Kuningan," Build. Informatics, Technol. Sci., vol. 4, no. 3, pp. 1464–1470, 2022, doi: 10.47065/bits.v4i3.2697.

[12] N. A. Maori and E. Evanita, "Metode Elbow dalam Optimasi Jumlah Cluster pada K-Means Clustering," Simetris J. Tek. Mesin, Elektro dan Ilmu Komput., vol. 14, no. 2, pp. 277–288, 2023, doi: 10.24176/simet.v14i2.9630.

[13] V. A. Ekasetya and A. Jananto, "Klusterisasi Optimal Dengan Elbow Method untuk Pengelompokan Data Kecelakaan Lalu Lintas di Kota Semarang," J. Din. Inform., vol. 12, no. 1, pp. 20–28, 2020, doi: 10.35315/informatika.v12i1.8159.

[14] M. R. Nahjan, N. Heryana, and A. Voutama, "Implementasi Rapidminer dengan Metode Clustering K-Means untuk Analisa Penjualan pada Toko Oj Cell," JATI (Jurnal Mhs. Tek. Inform.), vol. 7, no. 1, pp. 101–104, 2023, doi: 10.36040/jati.v7i1.6094.

[15] I. Azhami and R. Fauziah, "Penerapan Rapidminer pada Data Mining Klastering (Kasus: Distribusi Persentase Rumah Tangga Menurut Kabupaten/Kota

EISSN 2654-802X ; PISSN 2985-4067
DOI: https://doi.org/10.56873/JITU.7.1.5854. SUBMITTED: OCTOBER 9, 2024; REVISED: NOVEMBER 1, 2024; ACCEPTED: NOVEMBER 13, 2024

dan Bahan Bakar untuk Memasak),” KESATRIA J. Penerapan Sist. Inf. (Komputer Manajemen), vol. 1, no. 2, pp. 52–58, 2020, doi: 10.30645/kesatria.v1i2.20.