

Implementation of a Province Clustering Dashboard Based on IPM and Smoking Prevalence 2024 Using K-Means with KNN Validation as a Support for Data-Driven Decision Making

Indry Widiyani^{1*}, Asep Arwan Sulaeman², Handala Simetris Harahap³

¹²³Program Studi Teknik Informatika, Universitas Pelita Bangsa, Indonesia

* Corresponding author:

Email: indry.312210060@mhs.pelitabangsa.ac.id

Abstract.

Differences in the Human Development Index (HDI) and smoking rates across provinces in Indonesia indicate variations in social conditions that need to be systematically analyzed. This study aims to implement an interactive dashboard to cluster provinces based on their HDI and smoking rates for the year 2024 using the K-Means algorithm, and to validate the clustering results using K-Nearest Neighbor (KNN). The research data were obtained from the Central Statistics Agency (BPS) and the Regional Management Information System (SIMREG-Bappenas) and processed using RapidMiner Studio. The research stages included data cleaning, normalization using Z-transformation, clustering with K-Means into three clusters, validation using KNN via cross-validation, and the implementation of a Streamlit-based dashboard. The results show that provinces in Indonesia can be grouped into three clusters with distinct characteristics based on HDI values and smoking rates. The developed dashboard presents the analysis results in the form of tables, graphs, and interactive maps, thereby facilitating data interpretation and supporting data-driven decision-making. Validation results indicate that the clustering model exhibits a high level of consistency, making it suitable as a basis for formulating policy recommendations regarding regional development and public health.

Keywords: K-Means, KNN, Dashboard, HDI, Smoking and Data Mining.

I. INTRODUCTION

In Indonesia, the level of smoking habits is high in various provinces, which has an impact on people's health and productivity, and is one of the social problems that is still a concern. In both developed and developing countries, smoking is a major health problem. Smoking is not only done by adults, but also children in their teens and even women [1]. According to data from the Central Statistics Agency (BPS) in 2024, the percentage of people aged 15 and over who smoke is still high in many provinces.

Smoking at productive age not only has a direct impact on the health of individuals, but also impacts the productivity and financing of the health sector in the long run. On the other hand, to see the extent of the success of human development, the *United Development Programme* (UNDP) has issued an indicator, namely the *Human Development Index* (HDI). The Human Development Index (HDI) greatly affects the progress of a country; if the HDI is not good, the country's progress expectation will be lower.

To achieve this goal, it is crucial to consider or pay attention to the regions in each country, especially in remote and under-HDI areas [2]. The success of development should indeed be measured not only by high economic growth, but also by the improvement of human quality [3]. Previous studies have analyzed the grouping of provinces in Indonesia using HDI and smoking habit levels. This is done using the K-Means and K-Nearest Neighbour (KNN) techniques [4].

However, because the results of the analysis are still in the form of static tables and graphs, they look non-interactive and cannot be used as a visual aid in the decision-making process by the general public or related agencies. The use of digital technology allows complex information to be presented in a more communicative and interactive form, so that it can increase users' understanding of the information conveyed [5].

Therefore, an interactive web-based dashboard using the K-Means and KNN algorithms was created to display the results of the grouping and classification of provinces based on HDI and smoking percentage. In addition, this dashboard is expected to help the process of interpretation and policy formulation related to public health and regional development. A well-designed dashboard can provide a concise overview of the process, which makes it an effective tool in the decision-making process [6].

II. METHODS

RapidMiner is a standalone program for data analysis and *data mining* that can be integrated into its own products. RapidMiner can be used on all operating systems because it is built using the Java language [7]. *Data mining* is an automated data retrieval process that has the ability to search and process data [8]. One of the methods in *data mining* that is often used is *clustering*, which is the process of grouping data into several groups or segments based on similar characteristics [9].

The K-Means *Clustering* algorithm is an *unsupervised learning* technique used to group data based on the similarity of characteristics between objects. Recent research shows that the application of *K-Means* is able to group data well and is used in conjunction with other evaluation techniques such as *silhouettes* or validity indices to measure the quality of clusters formed (e.g. in customer segmentation visualization) [10].

The *K-Nearest Neighbour* (KNN) algorithm is a *distance-based supervised learning* method used to classify or evaluate data based on proximity to a number of nearby neighbors [11]. KNN was first introduced by Cover and Hart (1967) and is known as a simple but effective algorithm [12]. Another advantage of this algorithm is that it is able to provide users with faster and more accurate recommendations with good quality [13].

In this study, KNN was used not as the main method, but as a method of validating the results of *K-Means clustering* [14].

Table 1 Research Data Sources

No	Variabel	Sumber Data	Tahun
1	Persentase Penduduk Usia ≥ 15 Tahun yang Merokok	Badan Pusat Statistik (BPS)	2024
2	Indeks Pembangunan Manusia (IPM)	SIMREG Bappenas	2024

The research data consists of 38 provinces in Indonesia with two main variables, namely the percentage of the population aged 15 years and above who smoke and the value of the Human Development Index (HDI) in 2024. The data was obtained from the official publication of the Central Statistics Agency (BPS) and the Regional Management Information System (SIMREG) of Bappenas.

Table 2 Dataset Penelitian

No	Provinsi	Smoking Percentage	IPM 2024
1	ACEH	28.61	74.03
2	SUMATERA UTARA	26.69	74.02
...	...	...	...
37	PAPUA TENGAH	19.48	59.75
38	PAPUA PEGUNUNGAN	19.54	53.42

The complete dataset used in the study consists of 38 provinces according to the latest Indonesian administrative division after the expansion of Papua in 2022, and is used as input to the *clustering* process using the K-Means algorithm.

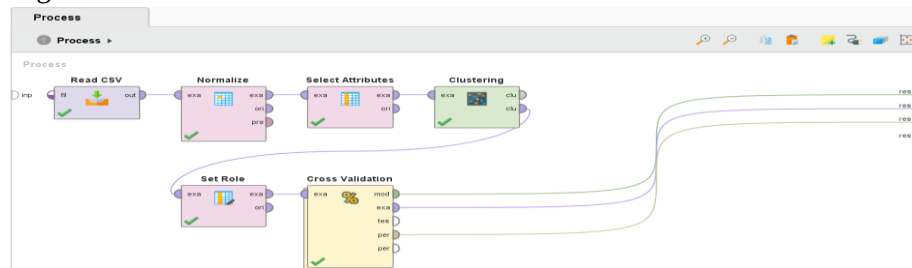


Fig. 1 Workflow on RapidMiner

The initial stage of the research is carried out by collecting and cleaning data to ensure that there is no lost data or duplication. Next, the data is normalized using the *Z-Transformation* method. Studies comparing normalization techniques show that the use of *Z-Score* can provide more optimal *clustering* results than without normalization or using other techniques because it reduces the effect of variable scale differences [15]. After the normalization process is completed, the selection of research attributes is carried out by grouping data using the K-Means algorithm with a number of clusters of three clusters. The selection of the

number of clusters aims to identify the characteristics of the province based on the level of human development and smoking habits.

The grouping results obtained from the K-Means algorithm are then used as labels in the validation process using the *K-Nearest Neighbour* (KNN) algorithm. Validation is carried out through the *Cross Validation* method in the RapidMiner software to measure the consistency and reliability of the *clustering results* that have been formed. Model evaluation was carried out using several performance metrics such as *accuracy*, *classification error*, *kappa*, *weighted recall*, *weighted precision*.

The final stage of the research is the implementation of a web-based dashboard using Streamlit. Streamlit is an *open-source framework* that can be accessed using the Python programming language [16]. *Built-in widgets* that can be used include image uploads, sliders, *text input*, and other *familiar hypertext markup language* (HTML) elements, such as *checkbox buttons* and radio buttons [17].

The *dashboard* is designed to display *clustering results* in the form of tables, graphs, *scatter plots*, cluster statistics, and provincial distribution maps interactively. In addition, the *dashboard* also provides interpretation of the results of the analysis and recommendations for policy directions based on the characteristics of each cluster. With this *dashboard*, data processing results can be presented in a more informative and easy-to-understand manner as a support for data-based decision-making.

III. RESULT AND DISCUSSION

Clustering Results Using K-Means

Row No.	id	cluster	Presentase ...	IPM 2024
1	1	cluster_0	0.314	0.319
2	2	cluster_0	-0.185	0.317
3	3	cluster_0	1.052	0.408
4	4	cluster_0	0.161	0.466
5	5	cluster_0	0.613	0.202
6	6	cluster_0	0.938	-0.017
7	7	cluster_0	1.445	0.194
8	8	cluster_0	1.674	-0.112
9	9	cluster_0	0.148	0.183
10	10	cluster_1	-0.505	1.084
11	11	cluster_1	-1.259	2.076

Fig. 2 Clustering Results on RapidMiner

Display some of the results of the *clustering process* using the K-Means algorithm on RapidMiner after the data has gone through the normalization stage. Each province obtained a cluster label that indicated the grouping results based on the similarity of Human Development Index (HDI) values and smoking percentage. The values in the smoking percentage and HDI columns displayed are the result of normalization using the *Z-Transformation* method so that there are positive and negative values. A positive value indicates that the data is above average, while a negative value indicates that the data is below average.

The resulting cluster labels are then used as a basis for analyzing the characteristics of each provincial group as well as the validation process using the *K-Nearest Neighbour* (KNN) algorithm. The *complete clustering results* are then returned to the original data value (denormalized) to make it easier to interpret and use in the preparation of policy recommendations.

The clustering results resulted in three groups of provinces with the following characteristics:

1. Cluster 0: HDI Close to Average & Relatively High Smoking Percentage

Table 3: Cluster 0

No	Provinsi	Smoking Percentage	IPM 2024	Cluster
1	ACEH	28.61	74.03	0
2	SUMATERA UTARA	26.69	74.02	0
3	SUMATERA BARAT	31.45	74.49	0
4	RIAU	28.02	74.79	0
5	JAMBI	29.76	73.43	0
6	SUMATERA SELATAN	31.01	72.30	0
7	BENGKULU	32.96	73.39	0
8	LAMPUNG	33.84	71.81	0
9	KEP. BANGKA BELITUNG	27.97	73.33	0

10	JAWA BARAT	32.98	74.43	0
11	JAWA TENGAH	29.13	73.88	0
12	JAWA TIMUR	28.72	74.09	0
13	BANTEN	31.31	74.48	0
14	NUSA TENGGARA BARAT	32.40	70.93	0
15	NUSA TENGGARA TIMUR	27.51	67.39	0
16	KALIMANTAN BARAT	28.06	70.13	0
17	KALIMANTAN TENGAH	28.02	72.73	0
18	KALIMANTAN UTARA	25.93	73.02	0
19	SULAWESI UTARA	26.93	75.03	0
20	SULAWESI TENGAH	28.98	71.56	0
21	GORONTALO	31.61	71.23	0
22	SULAWESI BARAT	26.99	68.20	0
23	MALUKU	29.43	71.57	0
24	MALUKU UTARA	29.72	71.03	0
25	PAPUA BARAT	25.90	67.02	0
26	PAPUA BARAT DAYA	25.00	68.63	0
27	PAPUA SELATAN	32.28	67.90	0

2. Cluster 1: High HDI & Relatively Low Smoking Percentage

Table 4: Cluster 1

No	Provinsi	Smoking Percentage	IPM 2024	Cluster
1	KEP. RIAU	25.46	77.97	1
2	DKI JAKARTA	22.56	83.08	1
3	DI YOGYAKARTA	25.18	81.55	1
4	BALI	19.22	77.76	1
5	KALIMANTAN SELATAN	23.07	73.03	1
6	KALIMANTAN TIMUR	23.99	78.83	1
7	SULAWESI SELATAN	24.66	74.05	1
8	SULAWESI TENGGARA	24.80	73.48	1
9	PAPUA	22.11	73.00	1

3. Cluster 2: Low HDI & Unique Characteristics

Table 5: Cluster 2

No	Provinsi	Smoking Percentage	IPM 2024	Cluster
1	PAPUA TENGAH	19.48	59.75	2
2	PAPUA PEGUNUNGAN	19.54	53.42	2

Dashboard Implementation

The *clustering* results are visualized using a Streamlit-based dashboard to facilitate data analysis and interpretation.



Fig. 3 Main Dashboard

Smoking Percentage This page displays the title of the application, the methods used, namely K-Means and *K-Nearest Neighbour* (KNN), as well as a brief description of the purpose of the dashboard. In the middle there is a file upload feature that functions to enter the clustered dataset in CSV or XLSX format before the visualization process is carried out.

Data Hasil Clustering

	provinsi	cluster	persentase_merokok	ipm_2024
0	ACEH	cluster_0	0.314	0.319
1	SUMATERA UTARA	cluster_0	-0.185	0.317
2	SUMATERA BARAT	cluster_0	1.052	0.408
3	RIAU	cluster_0	0.161	0.466
4	JAMBI	cluster_0	0.613	0.202
5	SUMATERA SELATAN	cluster_0	0.938	-0.017
6	BENGKULU	cluster_0	1.445	0.194
7	LAMPUNG	cluster_0	1.674	-0.112
8	KEP. BANGKA BELITUNG	cluster_0	0.148	0.183
9	KEP. RIAU	cluster_1	-0.505	1.084

Fig. 4 Clustering Data

This table contains information for each province along with cluster labels obtained from the grouping process. In addition, the value of the variables used in the study was also displayed, namely the percentage of smoking and HDI in 2024 that have gone through a normalization process using the *Z-Transformation* method. Through this table, users can see in detail the results of grouping each province into specific clusters.

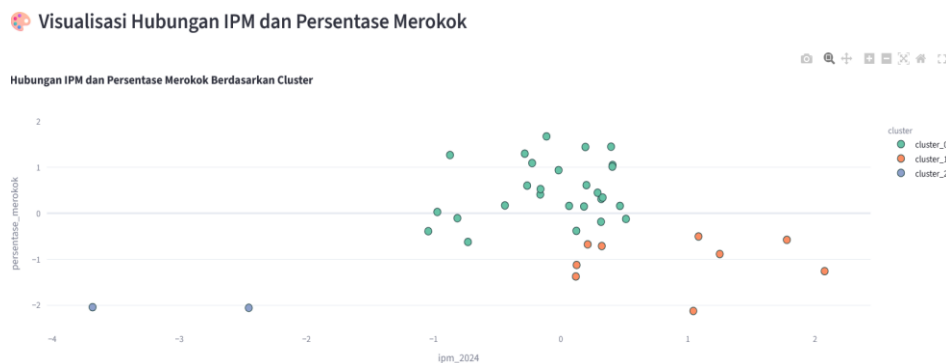


Fig. 5 Scatter Plot Graph

Based on the graph, it can be seen that the provinces in cluster_0 have a relatively dominant distribution and are in the range of HDI values and smoking percentages that are close to the average. Meanwhile, cluster_1 tend to be in provincial groups with different characteristics than other clusters. The cluster_2 have a smaller number of members and appear to be separate from other groups, showing quite different characteristics compared to other provinces in the dataset.

Statistik Rata-rata per Cluster

cluster	Rata-rata IPM 2024	Rata-rata Persentase Merokok
0 cluster_0	-0.0694	0.4944
1 cluster_1	0.8902	-1.0276
2 cluster_2	-3.0685	-2.052

Fig. 6 Average cluster Statistics

Based on the statistical table, cluster_0 have an average HDI value of -0.0694 and an average smoking percentage of 0.4944. This shows that the provinces in cluster_0 have HDI values that are close to the overall average with a relatively higher percentage of smoking than the average dataset. Furthermore, cluster_1 had an average HDI of 0.8902 and an average smoking percentage of -1.0276.

This value shows that the provinces in cluster_1 tend to have better HDI values with a relatively lower percentage of smoking compared to the overall provincial average. The cluster_2 have an average HDI of -3.0685 and an average smoking percentage of -2.0520. This value shows very different characteristics compared to other clusters, so cluster_2 can be categorized as a special group that is not the same as the majority of other provinces.

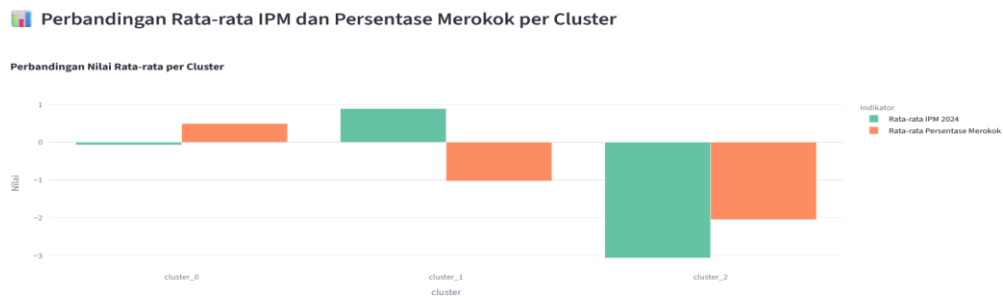


Fig. 7 Bar Chart

The bar graph at the bottom is used to clarify the comparison of the average values between *clusters*. The difference in bar height in each indicator shows that there is a significant variation in characteristics between the provincial groups that are formed.

Peta Persebaran Cluster Provinsi di Indonesia

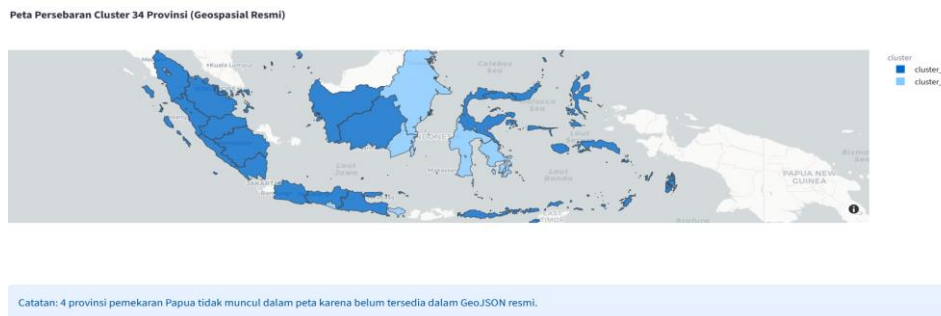


Fig. 8 Cluster Distribution Map

Displays a visualization of the distribution of *provincial clustering results* in Indonesia in the form of an interactive map. This visualization was developed using geospatial data (GeoJSON) and aims to facilitate the identification of cluster distribution patterns based on geographic area. The four provinces resulting from the expansion of Papua, namely South Papua, Central Papua, Mountainous Papua, and Southwest Papua cannot be displayed on the map due to the limitations of the GeoJSON data used, but they are still listed in the clustering results table. Because until now there is no official geospatial data containing the division of 38 provinces as a result of the expansion of Papua.

Interpretasi Awal

- Hasil clustering menunjukkan bahwa provinsi di Indonesia dapat dikelompokkan berdasarkan kemiripan karakteristik pada indikator Indeks Pembangunan Manusia (IPM) dan persentase penduduk yang merokok.
- Visualisasi scatter plot memperlihatkan bagaimana posisi setiap provinsi terhadap kedua indikator yang digunakan. Jarak antar titik menunjukkan tingkat kemiripan atau perbedaan karakteristik antar provinsi.
- Statistik rata-rata per cluster memberikan gambaran umum mengenai kondisi masing-masing kelompok, sehingga memudahkan identifikasi cluster dengan tingkat IPM yang lebih tinggi atau lebih rendah serta tingkat merokok yang relatif lebih besar atau lebih kecil.
- Peta persebaran cluster menunjukkan distribusi geografis hasil pengelompokan.
- Hasil pengelompokan tidak dimaksudkan untuk menentukan peringkat provinsi, melainkan untuk mengidentifikasi kesamaan karakteristik sehingga dapat digunakan sebagai dasar analisis dan perumusan kebijakan yang lebih tepat sasaran.
- Informasi yang disajikan pada dashboard ini dapat membantu pemerintah, peneliti, maupun pemangku kepentingan dalam memahami kondisi pembangunan manusia dan perilaku kesehatan masyarakat secara lebih terstruktur dan berbasis data.

Fig. 9 Initial Interpretation

Displays the initial interpretation feature provided on the *dashboard* to help users understand the *clustering results* more easily. This section serves as a summary of information that explains the meaning of the visualization that has been displayed beforehand, whether in the form of tables, graphs, or distribution maps.

Rekomendasi Arah Kebijakan

Rekomendasi berikut disusun berdasarkan karakteristik masing-masing cluster yang diperoleh dari hasil pengelompokan provinsi menggunakan algoritma K-Means. Rekomendasi ini bersifat indikatif dan dapat digunakan sebagai bahan pertimbangan dalam penyusunan kebijakan pembangunan daerah dan kesehatan masyarakat.

> Cluster 0: IPM Mendekati Rata-Rata & Persentase Merokok Relatif Tinggi

> Cluster 1: IPM Tinggi & Persentase Merokok Relatif Rendah

> Cluster 2: IPM Rendah & Karakteristik Unik

Fig. 10 Policy Direction Recommendations

Cluster 0: IPM Mendekati Rata-Rata & Persentase Merokok Relatif Tinggi

Karakteristik: Provinsi dalam kelompok ini berada di fase transisi berkembang. Ekonomi dan infrastruktur mulai stabil, namun peningkatan kualitas hidup tertahan akibat beban pengeluaran masyarakat yang tinggi pada konsumsi rokok.

Arah Kebijakan: Pengendalian Konsumsi & Proteksi Anggaran Rumah Tangga

- Optimalisasi Pajak Daerah: Menegakkan alokasi pemanfaatan (earmarking) minimal 50% dari Pajak Rokok Daerah dan Retribusi Daerah (PRD-RD) secara ketat khusus untuk pelayanan kesehatan promotif dan preventif.
- Penetapan Kawasan Tanpa Rokok (KTR): Menerbitkan dan mempertegas Peraturan Daerah (Perda) mengenai KTR di 7 tatanan (fasilitas kesehatan, sekolah, tempat bermain anak, tempat ibadah, angkutan umum, tempat kerja, dan tempat umum).
- Edukasi Pengalihan Belanja Mandiri: Melakukan kampanye sosialisasi makroekonomi mikro (keluarga) untuk mengalihkan porsi pengeluaran rokok menjadi investasi pendidikan anak serta jaminan kesehatan mandiri (BPJS).

Fig. 11 Cluster 0 Policy Direction Recommendations

Cluster 1: IPM Tinggi & Persentase Merokok Relatif Rendah

Karakteristik: Kelompok provinsi ideal (Benchmark). Masyarakatnya mandiri secara ekonomi, memiliki literasi kesehatan yang baik, dan sarana publik memadai.

Arah Kebijakan: Retensi Kebijakan Sehat & Keberlanjutan Pembangunan (Sustainability)

- Standarisasi Kawasan Sehat Internasional: Mendorong kota-kota di provinsi ini untuk meraih sertifikasi kawasan sehat atau Healthy Cities dari lembaga kesehatan dunia.
- Program Insentif Daerah: Memberikan penghargaan atau insentif fiskal (seperti dana hasil daerah) bagi kota/kabupaten di dalam wilayahnya yang berhasil mempertahankan angka prevalensi merokok di bawah target nasional.
- Pusat Edukasi Nasional: menjadikan model kebijakan di cluster ini sebagai rujukan atau pilot project bagi provinsi-provinsi di Cluster 0 dan Cluster 2 untuk melakukan studi banding.

Fig. 12 Cluster 1 Policy Direction Recommendations

Cluster 2: IPM Rendah & Karakteristik Unik

Karakteristik: Merupakan wilayah khusus (Anomali/Outlier). Pembangunan manusia tertinggal jauh secara nasional akibat kendala geografis ekstrem, status Daerah Otonomi Baru (DOB), dan merokok dipengaruhi oleh faktor iklim dingin serta adat budaya.

Arah Kebijakan: Intervensi Asimetris, Pendekatan Kultural, & Akselerasi Infrastruktur Dasar

- Pendekatan Kebijakan Asimetris (Khusus): Mengintegrasikan kampanye kesehatan masyarakat ke dalam skema Dana Otonomi Khusus (Otsus) agar penanganannya memiliki anggaran yang kuat dan fleksibel.
- Edukasi Berbasis Tokoh Adat dan Agama: Intervensi pembatasan rokok tidak akan efektif jika menggunakan pendekatan hukum formal/denda uang. Sosialisasi bahaya rokok (terutama bagi remaja usia 15 tahun ke atas) wajib melibatkan Kepala Suku atau pemuka agama setempat dengan menyisipkannya ke norma sosial komunal.
- Perlindungan Dana Bantuan Sosial: Mengunci atau memberikan syarat ketat pada penyaluran bantuan sosial non-tunai (seperti bantuan pangan) agar sepenuhnya dikonversi menjadi pemenuhan protein (susu, telur, ikan) guna mengentaskan kemiskinan ekstrem dan stunting, alih-alih disalahgunakan untuk membeli rokok.
- Akselerasi Dimensi Dasar IPM: Fokus anggaran dipusatkan secara radikal pada perbaikan fasilitas kesehatan dasar (puskesmas pembantu pembawa oksigen/alat deteksi ISPA) dan pemerataan guru di distrik terpencil untuk mendorong angka harapan hidup dan lama sekolah.

Fig. 11 Cluster 2 Policy Direction Recommendations

Displays the Policy Direction Recommendations feature that is organized based on the characteristics of each cluster. This section provides a more applicable interpretation of possible policies that can be considered by decision-makers. Each recommendation is presented in the form of *an expandable section (accordion)* so that users can choose the cluster they want to learn without having to see all the explanations at once.

Thus, *the dashboard* not only functions as a medium for visualizing the results of data analysis, but also as a *decision support* tool that can help the government, researchers, and stakeholders in designing human development and public health policies in a more targeted manner.

Validation Results Using KNN

```
PerformanceVector
accuracy: 94.17% +/- 12.45% (micro average: 94.74%)
ConfusionMatrix:
True: cluster_0 cluster_1 cluster_2
cluster_0: 27 0 2
cluster_1: 0 9 0
cluster_2: 0 0 0
classification_error: 5.83% +/- 12.45% (micro average: 5.26%)
ConfusionMatrix:
True: cluster_0 cluster_1 cluster_2
cluster_0: 27 0 2
cluster_1: 0 9 0
cluster_2: 0 0 0
kappa: 0.869
ConfusionMatrix:
True: cluster_0 cluster_1 cluster_2
cluster_0: 27 0 2
cluster_1: 0 9 0
cluster_2: 0 0 0
weighted_mean_recall: 60.00% +/- 14.05% (micro average: 66.67%), weights: 1, 1, 1
ConfusionMatrix:
True: cluster_0 cluster_1 cluster_2
cluster_0: 27 0 2
cluster_1: 0 9 0
cluster_2: 0 0 0
weighted_mean_precision: 57.76% +/- 16.40% (micro average: 64.37%), weights: 1, 1, 1
ConfusionMatrix:
True: cluster_0 cluster_1 cluster_2
cluster_0: 27 0 2
cluster_1: 0 9 0
cluster_2: 0 0 0
```

Fig. 12 Performance Evaluation Results on RapidMiner

Table 6 Performance Evaluation Results

Metric	Value
Accuracy	94,17%
Classification Error	5,83%
Kappa	0,869
Weighted Recall	60,00%
Weighted Precision	57,78%

After the *clustering* process using the K-Means algorithm is completed, the next stage is to validate using the *K-Nearest Neighbor* (KNN) algorithm with *the Cross Validation* method. The cluster label of the

K-Means result is used as a class label in the validation process. Validation is carried out to measure the level of consistency of *clustering results* that have been formed.

An accuracy value of 94.17% was obtained. This value indicates that most of the data was successfully reclassified into the appropriate cluster. The *classification error* rate of 5.83% indicates that only a small portion of the data is subject to misclassification.

A Kappa value of 0.869 indicates an excellent degree of agreement between the classification results and the *cluster labels* used as a reference. In addition, the *weighted mean recall* of 60.00% and the *weighted mean precision* of 57.78% show the model's ability to recognize and group data in each cluster.

Overall, the evaluation results show that the KNN model is able to validate the results of *K-Means clustering* well, so that the grouping results obtained can be used as a basis for analysis and support for data-based decision-making.

IV. DISCUSSION

The results of the study show that the K-Means method is able to group provinces in Indonesia based on HDI and smoking percentage effectively. The existence of differences in characteristics between clusters shows that:

The level of human development is not always directly proportional to smoking behavior. Social and cultural factors can influence these patterns.

The use of KNN as a validation method shows that the *clustering results* have a fairly good level of consistency, so that they can be used as the basis for further analysis. In addition, the *developed dashboard* is able to present the results of analysis visually and interactively, making it easier for users to understand data patterns.

V. CONCLUSION

Based on the results of research that has been conducted on the implementation of the provincial grouping dashboard based on the Human Development Index (HDI) and the percentage of smoking in 2024 using the K-Means method and *K-Nearest Neighbor* (KNN) validation, the following conclusions can be obtained:

1. This study successfully utilizes the 2024 Human Development Index (HDI) data from SIMREG-Bappenas and the 2024 smoking population data from the Central Statistics Agency (BPS) as the basis for analyzing provincial groupings in Indonesia.
2. The *K-Means* method was successfully used to group provinces based on HDI characteristics and smoking percentage into three *clusters*, while the *K-Nearest Neighbour* (KNN) method was successfully used as a validation of *clustering results*. The validation results showed an accuracy level of 94.17%, so the grouping results obtained can be said to have a good level of consistency.
3. The *dashboard* was successfully developed using the Python-based Streamlit framework by displaying various visualizations such as *clustering results tables*, *scatter plots*, cluster average statistics, provincial distribution maps, and policy direction recommendations that make it easier for users to understand the results of the analysis.
4. This study focuses on the design and implementation of a provincial grouping dashboard based on HDI and smoking percentage without comparing performance with other *clustering methods*. The *resulting dashboard* is able to present information interactively and can be used as a support for data-driven decision-making.

REFERENCES

- [1] Suharmanto, W. S. Utami, N. Pratiwi, and F. Muhammad, "Penerapan Data Mining Menggunakan Algoritma K-Means Untuk Clustering Perokok Usia Lebih dari 15 Tahun," Bull. Inf. Technol. BIT, vol. 4, no. 4, pp. 501–507, Dec. 2023, doi: 10.47065/bit.v4i4.1067.
- [2] S. Hartanto, M. S. Adzan, D. Z. Haq, and D. C. R. Novitasari, "Analisis Indeks Pembangunan Manusia Di Jawa

- Timur Tahun 2022-2023 Berdasarkan Indikator Menggunakan Metode Fuzzy C-Means,” *INTEK J. Inform. Dan Teknol. Inf.*, vol. 7, no. 2, pp. 45–54, Nov. 2024, doi: 10.37729/intek.v7i2.5358.
- [3] A. Hardana and J. Nasution, “Pengaruh Rasio Keuangan Pemerintah Daerah terhadap Indeks Pembangunan Manusia,” *Glob. Financ. Account. J.*, vol. 6, no. 1, p. 52, Apr. 2022, doi: 10.37253/gfa.v6i1.6452.
- [4] B. A. Lubis, I. Widiyani, and E. Rilvani, “Pengelompokan Provinsi Merokok Di Indonesia Berdasarkan Ipm Menggunakan K-Means Dan Knn”.
- [5] A. A. Permatasari, D. C. Lolita, and C. C. Chotimah, “Peran Media Digital Dalam Upaya Promosi Kesehatan Untuk Meningkatkan Kesehatan Masyarakat : Tinjauan Literatur,” *Zaitun J. Ilmu Kesehat.*, vol. 11, no. 1, p. 1, Feb. 2023, doi: 10.31314/zijk.v11i1.2033.
- [6] M. Y. Bahtiar, A. Wahyudin, and A. Anisyah, “Perancangan Dashboard Interaktif Untuk Mengoptimisasi Analisis Hasil Audit Mutu Internal (AMI) Dengan Metode Pureshare,” *J. Komput. Teknol. Inf. Sist. Inf. JUKTISI*, vol. 4, no. 2, pp. 863–876, Aug. 2025, doi: 10.62712/juktisi.v4i2.550.
- [7] D. Hidayati, Y. Yahya, and Muh. A. Juniarta Hidayat, “Implementasi Data Mining Menggunakan Algoritma K-Means untuk Pengelompokan Obat di Puskesmas Kerongkong Kecamatan Suralaga,” *J. Print. J. Pengemb. Rekayasa Inform. Dan Komput.*, vol. 1, no. 2, pp. 78–88, Dec. 2023, doi: 10.29408/jprinter.v1i2.22006.
- [8] S. P. Dewi, N. Nurwati, and E. Rahayu, “Penerapan Data Mining Untuk Prediksi Penjualan Produk Terlaris Menggunakan Metode K-Nearest Neighbor,” *Build. Inform. Technol. Sci. BITS*, vol. 3, no. 4, pp. 639–648, Mar. 2022, doi: 10.47065/bits.v3i4.1408.
- [9] T. Setyadji, R. A. Ramadhani, and L. S. Wahyuniar, “Implementasi Data Mining Untuk Menentukan Pelanggan Potensial Menggunakan Algoritma Pengklasteran K-Means Dan K-Nearest Neighbors,” vol. 9.
- [10] D. A. Imanuel and G. Alfian, “Visualisasi Segmentasi Pelanggan Berdasarkan Atribut RFM Menggunakan Algoritma K-Means Untuk Memahami Karakteristik Pelanggan pada Toko Retail Online,” *J. Teknol. Inf. Dan Ilmu Komput.*, vol. 12, no. 2, pp. 283–292, Apr. 2025, doi: 10.25126/jtiik.2025128619.
- [11] J. Han, M. Kamber, and D. Mining, “Concepts and techniques,” *Morgan Kaufmann*, vol. 340, no. 1, pp. 94104–103205, 2006.
- [12] T. Cover and P. Hart, “Nearest neighbor pattern classification,” *IEEE Trans. Inf. Theory*, vol. 13, no. 1, pp. 21–27, Jan. 1967, doi: 10.1109/TIT.1967.1053964.
- [13] H. Hartatik and R. Rosyid, “Pengaruh User Profiling Pada Rekomendasi Sistem Menggunakan K Means Dan Knn,” *J. Inf. Syst. Manag. JOISM*, vol. 2, no. 1, pp. 13–18, Jul. 2020, doi: 10.24076/JOISM.2020v2i1.199.
- [14] C. C. Aggarwal, *Data Mining: The Textbook*. Cham: Springer International Publishing, 2015. doi: 10.1007/978-3-319-14142-8.
- [15] S. E. Saqila, I. P. Ferina, and A. Iskandar, “Analisis Perbandingan Kinerja Clustering Data Mining Untuk Normalisasi Dataset,” *J. Sist. Komput. Dan Inform. JSON*, vol. 5, no. 2, p. 356, Dec. 2023, doi: 10.30865/json.v5i2.6919.
- [16] O. S. H. Raharusun, A. Hasibuan, and G. C. Rorimpandey, “Rice Production Prediction System Using Multiple Linear Regression for Food Security Optimization in Minahasa Regency,” *J. Minfo Polgan*, vol. 14, no. 2, pp. 2287–2294, Oct. 2025, doi: 10.33395/jmp.v14i2.15340.
- [17] Widi Hastomo, Nur Aini, Adhitio Satyo Bayangkari Karno, and L.M. Rasdi Rere, “Metode Pembelajaran Mesin untuk Memprediksi Emisi Manure Management,” *J. Nas. Tek. Elektro Dan Teknol. Inf.*, vol. 11, no. 2, pp. 131–139, May 2022, doi: 10.22146/jnteti.v11i2.2586.