Journal of Artificial Intelligence and Engineering Applications

Website: https://ioinformatic.org/

15th February 2025. Vol. 4. No. 2; e-ISSN: 2808-4519

The Application of the K-Means Algorithm in Enhancing the Clustering Model for Job Seekers in Cirebon City

Laylatunna'imah[1*], Martanto[2], Arif Rinaldi Dikananda[3], Ahmad Rifa'i[4]

[1,2,3,4]*STMIK IKMI Cirebon*
ellanaimah@gmail.com [1*]

Abstract

The development of information technology opens up opportunities to improve the efficiency of job search through the grouping of job seekers based on specific characteristics. This study uses the K-Means algorithm to analyze data on job seekers in Cirebon City for 2018–2022, focusing on education level and gender. The stages of the research include (1) data selection, (2) data preprocessing, (3) transformation of attributes into numerical format, (4) data grouping using RapidMiner, and (5) evaluation of clustering results using the Davies-Bouldin Index (DBI). The results showed that the optimal number of clusters was two (K=2), with a DBI value of 0.608 which indicates good cluster separation. The first cluster consists of job seekers with a higher level of education, while the second cluster has a lower level of education. Gender did not show a significant influence. These findings provide strategic insights for governments and companies in developing data-driven policies, such as more effective training or recruitment programs. The K-Means algorithm has proven its potential in supporting strategic decision-making in workforce management and being adaptable to other regions.

Keywords: *K-Means algorithm, clustering, job seekers, education level, Cirebon City*

1. Introduction

Information and communication technology development has significantly impacted the world of work, especially in the increasingly complex job search. As a grouping method in data mining, the K-Means algorithm can improve the matching efficiency between job seekers and companies by identifying patterns in job seeker data. [1]. With the rise of job search platforms, it is important to develop models to analyze the big data generated. [2]. This study aims to explore using the K-Means algorithm in the context of job search in Cirebon City, which has unique job market dynamics. [3].

Although the K-Means algorithm is effective, there are challenges in grouping that can result in suboptimal matching. Some previous studies have not considered important variables such as specific skills and individual preferences in the grouping process [4]. This research aims to overcome these shortcomings and develop a more effective model. Several related studies, such as those conducted by [5],[6] , and [7], These points to the importance of data analysis in an economic context but do not address the application of the K-Means algorithm in job search.

This research will use the K-Means algorithm to group data on job seekers in Cirebon City based on education level and gender. Data will be obtained from official sources such as BPS and the Manpower Office to ensure the accuracy of the analysis. The grouping process includes data preprocessing, normalization, and evaluation using validity indices such as the Davies-Bouldin Index. The results of this study are expected to provide new insights into the application of the K-Means algorithm to support labor absorption strategies and improve employment policies [8]. These findings can also provide practical benefits for companies and job seekers in improving job matching jobs.

2. Literature Review

The literature review method in this study aims to analyze the application of the K-Means algorithm in clustering job seekers based on education level and gender. This process was conducted systematically through several steps, namely problem formulation, literature search, literature evaluation, and analysis and interpretation. The main issue addressed is how the K-Means algorithm can be utilized to improve the job seeker clustering model in Cirebon City. Subsequently, relevant literature was collected from databases such as Google Scholar, Scopus, and Sinta using keywords like "K-Means algorithm," "job seeker clustering," and "data mining." The selection was made by considering recency (within the last 5 years) and relevance to the research topic. From this search, 16 articles were identified as meeting the criteria for further analysis.

3. Research Methods

3.1. Methods

The following is a chart of the research method as follows:

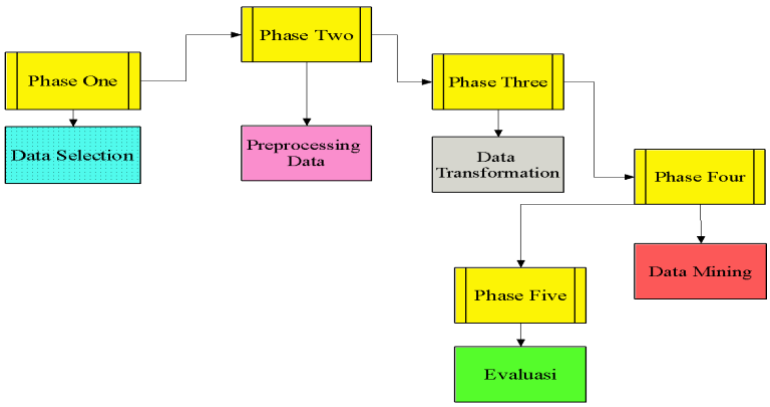

Fig. 1: Stages of Research Methods

3.2. Supporting Data

At this stage, the selection of job seeker data to be further analyzed is carried out. The data used in this study was obtained from official sources, namely the Cirebon City Open Data website, which presents information on job seekers based on education level and gender in Cirebon City. This data selection process ensures that only relevant data that match the research criteria are included in the analysis. Thus, the resulting grouping results are expected to accurately reflect job seekers' characteristics, so that they can support more optimal mapping and grouping.

Table 1: Job Seeker Dataset

No	Name of Regency/City	Education Level	Gender	Number of Job Seekers	Year
1	Kota Cirebon	SD	Laki-Laki	11	2022
2	Kota Cirebon	SD	Perempuan	46	2022
3	Kota Cirebon	SLTP	Laki-Laki	39	2022
4	Kota Cirebon	SLTP	Perempuan	68	2022
5	Kota Cirebon	SMU/Sederajat	Laki-Laki	997	2022
...	...	...	...	...	...
56	Kota Cirebon	Diploma	Perempuan	157	2018
57	Kota Cirebon	Sarjana	Laki-Laki	214	2018
58	Kota Cirebon	Sarjana	Perempuan	393	2018
59	Kota Cirebon	Pasca Sarjana	Laki-Laki	2	2018
60	Kota Cirebon	Pasca Sarjana	Perempuan	4	2018

Table 1 presents a dataset of job seekers in Cirebon City for the period 2018–2022. This dataset includes attributes: Regency/City Name (Cirebon City), Education Level (Elementary, Junior High, High School/Vocational, Diploma, Undergraduate, Postgraduate), Gender (male/female), Number of Job Seekers (by category), and Year (2018–2022). This data provides an overview of job seeker profiles based on education level and gender, which will be processed using the K-Means algorithm to form a cluster of job seeker characteristics.

4. Results and Discussion

4.1. Result

This chapter presents the results of research that has been conducted related to the use of the K-Means algorithm in the grouping of job seekers in Cirebon City. The analysis was carried out on the data of job seekers who had been clustered based on education level and gender, to develop a more effective and efficient grouping model of job seekers using the K-Means algorithm. The results of this study are expected to be able to provide an overview of the distribution of job seeker groups in Cirebon City which can be the basis for the preparation of a more effective employment strategy.

4.1.1. Preprocessing

The preprocessing stage ensures the quality of the data for clustering analysis by eliminating inconsistent values and irrelevant attributes. The steps include checking *missing values*, selecting attributes, and normalizing data to match the K-Means algorithm format. This process ensures that the data used is more accurate and consistent so that the clustering results are more optimal and relevant.

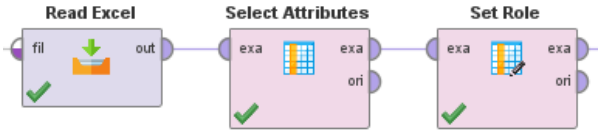

Fig. 2: Excel Read Operators, Select Attributes, and Set Roles

Figure 2 shows the Excel Read Operators, Select Attributes, and Set Roles, there are three operators used in the data preprocessing process. The following is an explanation of each operator's function:

a. This Excel Read operator functions to read Excel files that contain job seeker data. This operator imports data from external files into the processing environment so that the data is ready for the next stage of preprocessing.
b. Once the data is imported, this Select Attributes Operator serves to select attributes that are relevant to the analysis purpose. At this stage, attributes that are irrelevant or unrelated to the clustering process can be excluded from the dataset.
c. The Set Role operator sets the role or "role" of each attribute in the dataset, according to its function in the clustering process. For example, certain attributes can be set as labels or IDs. By defining the role of each attribute, the clustering process can run more purposefully and effectively.

4.1.2. Transformation

The next stage is to carry out the data transformation process to convert nominal data types into numerical data types so that they can be further processed by the K-Means algorithm in the RapidMiner software. This step is very important because the K-Means algorithm requires data in numerical format to measure the distance between data based on *centroids*.

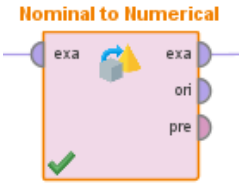

Fig. 3: Operator Nominal To Numerical

Figure 3 shows the use of *the Nominal to Numerical* operator with certain parameter settings to convert the nominal attributes of "Regency/City Name, Education Level, and Gender" into numerical form.

Row No.	nama_kabu...	tingkat_pen...	jenis_kelamin
1	KOTA CIREB...	0	0
2	KOTA CIREB...	0	1
3	KOTA CIREB...	1	0
4	KOTA CIREB...	1	1
5	KOTA CIREB...	2	0
6	KOTA CIREB...	2	1
7	KOTA CIREB...	3	0
8	KOTA CIREB...	3	1
9	KOTA CIREB...	4	0
10	KOTA CIREB...	4	1
11	KOTA CIREB...	5	0
12	KOTA CIREB...	5	1
13	KOTA CIREB...	0	0
14	KOTA CIREB...	0	1
15	KOTA CIREB...	1	0

Fig. 4: Data Transformation Results.

Figure 4 shows the results of the data conversion process using *the Nominal to Numeric* operator, where the columns that were previously nominal types such as *Education Level* and *Gender* have been successfully converted into numerical form. This result is particularly important because the K-Means algorithm, used in clustering analysis, requires numerical data to perform the calculation of the distance between the data point and *the centroid.*

4.1.3. Data Mining

At this stage, the K-Means algorithm is used to group the data of job seekers in Cirebon City based on education level and gender. This algorithm groups the data based on the distance between the data points and the centroids in each cluster, after determining the optimal

number of clusters (K). This study uses RapidMiner software, with the operator *Cluster Distance Performance* to evaluate the quality of clustering results.

Fig. 5: K-Means Algorithm

Figure 5 shows the use of the K-Means Algorithm operator to group the data of Job Seekers in Cirebon City. In this process, the Cluster Distance Performance operator is used to evaluate the quality of the cluster and determine the best number of clusters (K). In this case, the K value is set to 6 to see the grouping variations that might give the optimal result.

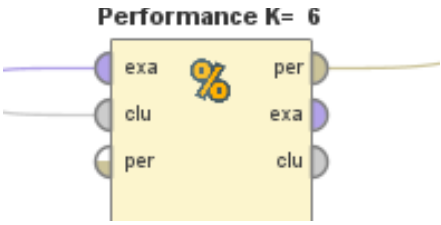

Fig. 6: Operator Cluster Distance Performance

Figure 6 shows the use of *Cluster Distance Performance* operators in RapidMiner to evaluate the quality of the clustering results of job seeker data in Cirebon City. These operators calculate the distance metric between clusters to determine the optimal number of clusters (K), ensuring clear and effective data sharing in describing job seeker patterns.

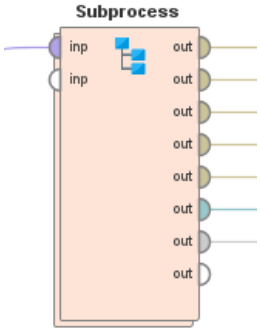

Fig. 7: SubProcess Operator

Figure 7 shows the use of *the SubProcess* operator in the RapidMiner software to run the clustering process in a structured and organized manner. This operator serves as a container that allows multiple operations or algorithms to be executed sequentially within it. In the context of clustering analysis with the K-Means algorithm, *SubProcess* is used to manage various stages of data analysis modularly.

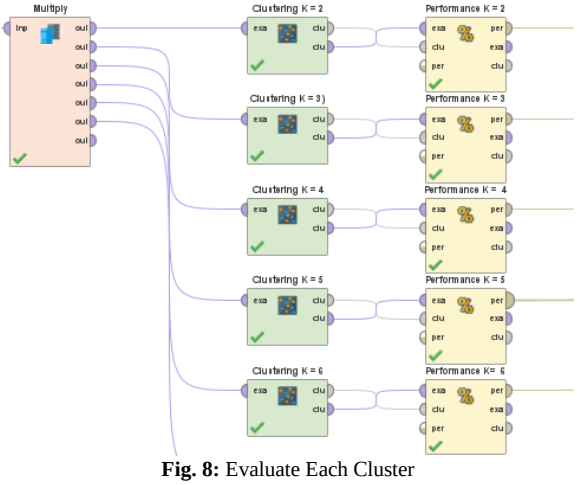

Fig. 8: Evaluate Each Cluster

Furthermore, in the Sub Process Operator, a multiplication process is carried out on the data input which is then forwarded to various stages of clustering with different variations of K values, namely from K=2 to K=6. Each clustering process produces an output in the form of cluster results which are then evaluated through the Performance stage for each K value used. In the Performance evaluation process, there are several evaluation metrics used to assess the quality of the clustering results of each K value. Each K value will be compared with each other to get optimal results in determining the number of clusters that are most suitable for the data used.

4.1.4. Evaluation

At this stage, an evaluation of the clustering results is carried out to determine the optimal number of clusters (K), using *the Davies-Bouldin Index* (DBI) metric as an indicator of cluster quality.

Table 2: DBI Values of Each Cluster

K	***Davies Bouldin Index***
2	0.608
3	0.707
4	0.805
5	0.848
6	0.898

Table 2 shows the *Davies-Bouldin Index (DBI)* values for each number of clusters (K) tested. Based on the table, the best cluster is achieved at K=2 with a value of 0.608, which indicates excellent cluster separation.

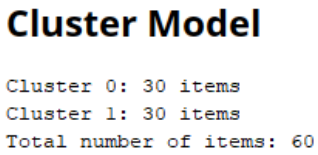

Fig. 9: Model Cluster K=2

Figure 9 shows the distribution of the amount of data in each cluster formed from the clustering process with K=2, where there are a total of 60 items divided into 2 clusters. Cluster 0 and Cluster 1 have the same number of 30 items. This distribution provides an overview of how much the amount of data varies in each cluster and helps in assessing the quality of cluster separation in the clustering process that has been carried out.

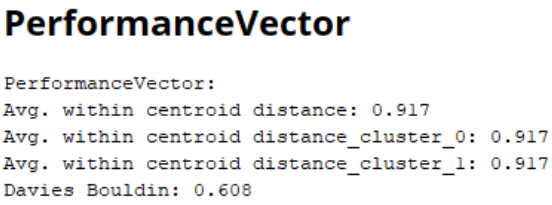

Fig. 10: Performance Vector

Figure 10 shows the results of the evaluation of the clustering of job seeker datasets in Cirebon City using *Performance Vector*. The main indicator, which is the average within centroid distance, is 0.917 for Cluster 0 and Cluster 1. The Davies-Bouldin Index value of 0.608 indicates a fairly good quality of cluster separation. These results illustrate an effective data division even though the average distance in the cluster is the same for both clusters.

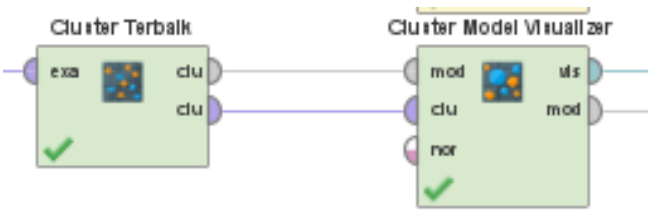

Fig. 11: Best Cluster & Cluster Operator Model Visualizer

Figure 11 shows the best cluster results and the Cluster Model Visualizer operator. After carrying out the clustering process, the next step is to visualize the results of the best clusters.

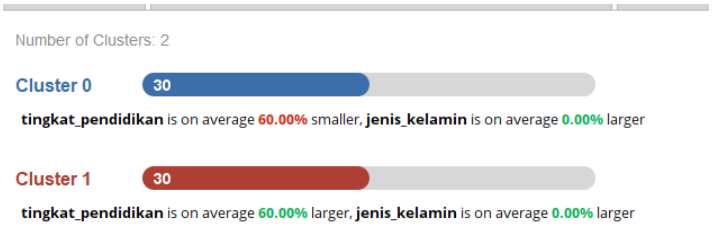

Fig. 12: Cluster Visualization

Figure 12 shows the results of clustering job seeker data in Cirebon City into two clusters, each containing 30 data. Cluster 0 had an average education level 60% higher than Cluster 1, while the gender variable did not show a significant difference. The level of education is the main differentiating factor between the two clusters.

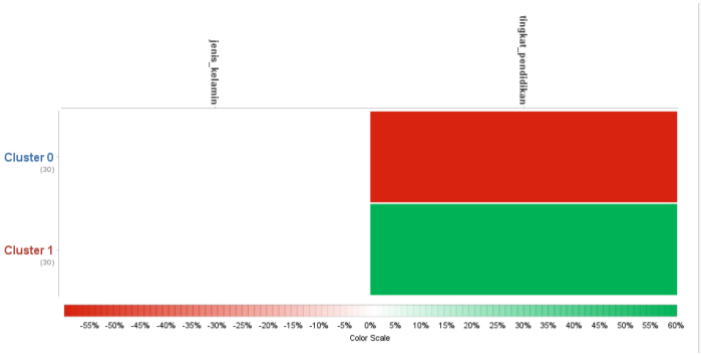

Fig. 13: Heatmap visualization

Figure 13 compares the characteristics of Cluster 0 and Cluster 1 in the dataset of job seekers in Cirebon Regency based on education level and gender. Cluster 0 has a 60% higher education rate, marked in green, while Cluster 1 is 60% lower, marked in red. The gender variable did not show significant differences in the two clusters. The level of education is the main distinguishing factor.

Cluster	tingkat_pendidikan	jenis_kelamin
Cluster 0	1	0.500
Cluster 1	4	0.500

Fig. 14: Centroid of Each Cluster

Figure 14 shows the centroid of job seeker clusters in Cirebon City based on education level and gender. Cluster 1 has a higher average education level (4) than Cluster 0 (1). Meanwhile, the gender variable was valued at 0.500 in both clusters, signaling a balanced distribution. This clustering mainly differentiates based on education level.

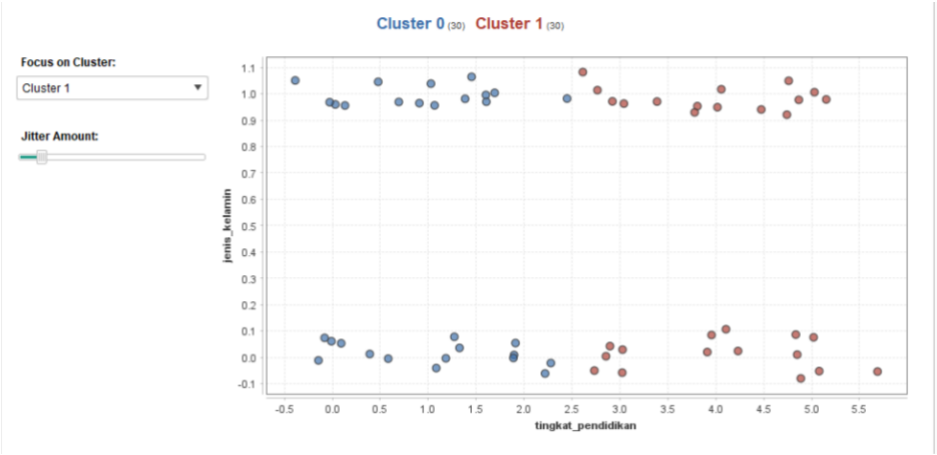

Fig. 15: Visualization *of the Scatter Plot*

Figure 15 shows the distribution of job seeker data in Cirebon City in two clusters based on education level (x-axis) and gender (y-axis). The blue dots represent Cluster 0, with higher education levels (3–5.5), while the red dots indicate Cluster 1, which is lower (0–2.5). The distribution of sex was evenly distributed in both clusters, showing job seekers of both genders in each group. This visualization highlights the key differences of clusters based on education level.

4. Discussion

The results of this study show the grouping of job seekers in Cirebon City into two clusters based on education level and gender, with a Davies-Bouldin Index (DBI) value of 0.608 at K=2. This value reflects the fairly good quality of the clusters, where the separation between clusters is relatively clear.
Overall, this study contributes to the literature on the grouping of job seekers by identifying education level as a key characteristic in cluster formation, although gender remains relevant to understanding the overall distribution of job seekers. Compared to other studies, the approach provided comprehensive results but still had room for further development, such as the addition of new variables or the application of other evaluation methods.

5. Conclusions

Based on the results of the research that has been carried out, it can be concluded that the research objectives in grouping job seeker data in Cirebon City with the K-Means algorithm approach have been successfully achieved. Here are the main conclusions that can be drawn from this study:

a. This study succeeded in classifying data on job seekers in Cirebon City using the K-Means algorithm through the RapidMiner software. Job seeker data can be grouped into two main clusters based on education level. Cluster 1 contains job seekers with a higher level of education, while Cluster 0 includes job seekers with a lower level of education. These results show that education level is the dominant variable in cluster formation, while gender does not significantly affect the cluster distribution.
b. The determination of the optimum value for the number of clusters was carried out using the Davies-Bouldin Index (DBI) validation method. The results of the analysis show that the number of optimal clusters is 2, with a DBI value of 0.608. This value indicates a fairly good quality of grouping, where the separation between clusters is relatively clear and the homogeneity within the clusters is quite high.

6. Suggestions

Based on the results of the research that has been carried out, several suggestions can be considered for the development of further research and the application of the results of this research. These suggestions are expected to help improve the accuracy of grouping job seeker data.

a. This research can be further developed by adding other variables such as work experience, age, or the industry sector of interest to get a more holistic understanding of job seekers. In addition, the incorporation of other algorithms such as K-Means++ or hierarchical methods can be compared to obtain a more optimal clustering.
b. Relevant agencies, such as the Manpower Office, are advised to improve the quality and coverage of job seeker data. More complete and accurate data will support better analysis and more effective data-driven decision-making.

References

[1] H. Rahmani, A. Mohammad Rahmani, and W. Groot, "A predictive analytics solution matching job seekers' talent and employers' demands based on machine learning," 2023.
[2] Y. A. Skobtsov, D. M. Obolensky, V. I. Shevchenko, and O. V. Chengar, "Building And Analysing A Skills Graph Using Data From Job Portals," *Proc. III Int. Conf. Econ. Soc. Trends Sustain. Mod. Soc. – (ICEST-III 2022), 19-21 May, Krasn. Sci. Technol. City Hall, Krasn. Russ. Fed.*, vol. 127, pp. 147–162, 2022, doi: 10.15405/epsbs.2022.08.17.
[3] E. B. Wijaya, A. Dharma, D. Heyneker, and J. Vanness, "Comparison of the K-Means Algorithm and C4.5 Against Sales Data," *SinkrOn*, vol. 8, no. 2, pp. 741–751, 2023, doi: 10.33395/sinkron.v8i2.12224.
[4] M. Pokharel, J. Bhatta, and N. Paudel, "Comparative Analysis of K-Means and Enhanced K-Means Algorithms for Clustering," *NUTA J.*, vol. 8, no. 1–2, pp. 79–87, 2021, doi: 10.3126/nutaj.v8i1-2.44044.
[5] A. Wardhana, B. Kharisma, and M. N. F. Sofyan, "Dampak Penerimaan Dan Pengeluaran Pemerintah Daerah Terhadap Pendapatan Perkapita Antar Kabupaten Jawa Barat," *Ekon. dan Bisnis*, vol. 8, no. 2, pp. 131–141, 2021, doi: 10.35590/jeb.v8i2.3474.
[6] J. Beno, A. . Silen, and M. Yanti, "Analisis Struktur Kovarians pada Indikator Terkait Kesehatan di Kalangan Lansia yang Tinggal di Rumah dengan Fokus pada Persepsi Subjektif tentang Kesehatan," *Braz Dent J.*, vol. 33, no. 1, pp. 1–12, 2022.
[7] E. Suwandana, "Apresiasi Dan Evaluasi Peraturan Menteri Pendayagunaan Aparatur Negara Dan Reformasi Birokrasi Tentang Jabatan Fungsional Widyaiswara," *J. Kewidyaiswaraan*, vol. 7, no. 1, pp. 246–254, 2022, doi: 10.56971/jwi.v7i1.205.
[8] Y. Zhang *et al.*, "Self-adaptive k-means based on a covering algorithm," *Complexity*, vol. 2018, 2018, doi: 10.1155/2018/7698274.