Transaction on Informatics and Data Science
Vol. 1, No. 1, 2024, 45-53
ISSN: 3064-1772 (online)
DOI: https://doi.org/10.24090/tids.v1i1.12231

K-Means Clustering in Relevance Grouping of Undergraduate Informatics Jobs: Case Study at the Informatics Engineering Department, Universitas Muhammadiyah Malang, Malang, Indonesia

Dikky Cahyo Hariyanto *[1], Sri Harini[1], Totok Chamidy [1]

[1] Computer Science Department, Faculty of Science & Technology, UIN Maulana Malik Ibrahim, Malang, Indonesia.

Article Information

Submitted February 6, 2024
Accepted March 11, 2024
Published April 1, 2024

Abstract

Higher education is one of the levels of study expected to produce graduates competent in the field of knowledge taken. The large number of graduates from tertiary institutions with many job opportunities causes many graduates to work in ways that do not match their majors, so there is a need to evaluate the level of success of graduates learning achievements in tertiary institutions. This research aims to analyze data on the relevance of the work of undergraduate graduates in Informatics Engineering to what they have learned by the learning outcomes in the Informatics Engineering study program at the University of Muhammadiyah Malang using K-Means clustering. Using data from questionnaires measuring graduate learning outcomes and measuring job suitability for 137 respondents who had been tested for validity, reliability, and multicollinearity, the results of this research showed that the data was formed into three clusters with the analysis that 29.92% of UMM Informatics Engineering graduates were able to meet graduate learning outcomes and obtain jobs that are relevant to what they studied, 49.63% of other graduates also got jobs that were relevant to their major even though they lacked mastery of specific skills as measured by graduate learning outcomes, and 20.45% of other graduates got jobs that were less relevant to the field of Informatics engineering.

Keywords: K-Means clustering, unsupervised learning, data mining, informatics jobs.

1. Introduction

Society expects higher education to produce graduates who are competent in their fields of study. The high interest in studying in higher education has led to the establishment of many higher education institutions. Data collected from the Central Bureau of Statistics (BPS) in 2022 shows 3107 universities in Indonesia[1].

The large number of universities impacts the number of graduates produced. From the existing phenomenon, many college graduates get jobs not based on what they learned during college. This study aims to analyze data on the relevance of the work of Informatics Engineering graduates and what they learned according to the learning achievements in the Informatics Engineering study program at the University of Muhammadiyah Malang using K-Means.

The K-means algorithm is a method used in cluster analysis and data mining because of its simplicity and computational efficiency. This algorithm aims to partition a given data set into many predetermined clusters, with each data point assigned to the cluster with the closest mean. This process helps identify clustering in the data, which is very useful for pattern recognition and data interpretation [2].

The K-means algorithm has been widely discussed in the literature and has several drawbacks. One of the major drawbacks is its sensitivity to the initial selection of cluster centroids, which can significantly impact the final clustering results [3]. The dependence on

* **Author Correspondence**: Dikky Cahyo Hariyanto: UIN Maulana Malik Ibrahim, Gajayana street No.50, Dinoyo, Lowokwaru, Malang, East Java 65144 – Indonesia. Email: dikky037@gmail.com .

the inileadvalues often leads to poor-quality clustering results, especially when the initial centroids are not selected properly.

K-means is widely known in data mining and is recognized as one of the top 10 algorithms [4]. This algorithm is used to group and cluster data in various domains, such as urban data analysis [5], government debt risk evaluation [6], student achievement assessment [7], financial market strategy mining [8], health profile clustering [9], and crime rate analysis [10]. This algorithm is known for its unsupervised process and ability to group data using a partition system [9]. On the other hand, this algorithm has also been applied to monitoring models, such as wind turbine vibration monitoring based on SCADA data [11]. The K-means algorithm has been compared with other methods, such as in the study of traffic clustering algorithms, where this algorithm is used as a benchmark for comparison [5]. In addition, this algorithm has been used for speech classification and audio feature extraction, demonstrating its versatility in various data analysis tasks [12]. Researchers have also used the algorithm for performance analysis in educational institutions, clustering data collected from private universities using the K-means method [13]. This application highlights the adaptability and effectiveness of the algorithm in various research fields.

K-Means Clustering is one of the most widely used algorithms in unsupervised learning, particularly for partitioning a dataset into distinct groups based on inherent patterns. The goal of this algorithm is to minimize intra-cluster variance while maximizing inter-cluster differences. Each data point is assigned to the nearest cluster centroid based on Euclidean distance, forming a cluster where the mean of the points represents the centroid. The algorithm iteratively updates centroids to minimize the overall variance in the dataset[14].

Starting with an initial set of k means $m_1^{(1)}, \ldots, m_k^{(1)}$, the algorithm alternates between two steps[15]:

Step 1 for assignment allocate each observation to the cluster whose mean is closest, based on the smallest squared Euclidean distance.

$$S_i^{(t)} = \left\{ x_p : \left\| x_p - m_i^{(t)} \right\|^2 \leq \left\| x_p - m_j^{(t)} \right\|^2 \forall j, 1 \leq j \leq k \right\} \quad (1)$$

Each x_p is assigned to exactly one $S^{(t)}$, even if it could potentially be assigned to multiple clusters.

Step 2 for recomputing the means (centroids) for the observations allocated to each cluster.

$$m_i^{(t+1)} = \frac{1}{\left| S_i^{(t)} \right|} \sum_{x_j \in S_i^{(t)}} x_j \quad (2)$$

The objective function in k-means is the within-cluster sum of squares (WCSS). With each iteration, the WCSS decreases, resulting in a nonnegative, monotonically decreasing sequence. This ensures that the k-means algorithm always converges, though not necessarily to the global optimum.

The algorithm is considered to have converged when the assignments stop changing or, alternatively, when the WCSS stabilizes. However, the algorithm is not guaranteed to find the optimal solution. The k-means algorithm generally uses Euclidean distance for cluster assignment, but variations like spherical k-means and k-medoids allow for alternative distance measures to improve flexibility and convergence.

K-means has been widely used in various applications in data mining, demonstrating its significance and flexibility in clustering and analyzing data in various domains. This effectiveness makes the author use this algorithm to analyze data on the relevance of college graduate jobs, especially the Informatics Engineering study program at the University of Muhammadiyah Malang.

2. Method

The data collection method used in this study was through a questionnaire measuring graduates' learning achievements and the suitability of the jobs they get. The instrument was adopted from a questionnaire made by [16]. The questionnaire was distributed to graduates of the Informatics Engineering Undergraduate Program, University of Muhammadiyah Malang (UMM) from 2018 to 2021, totaling 836 graduates. The data from filling out the questionnaire, as many as 137 respondents with 32 question items representing elements of attitude, knowledge, general skills, and special skills, were processed using the Knowledges Discovery in Database process, and validity, reliability, and multicollinearity tests were carried out before entering the computation and the K-Means algorithm was applied. Improvements were made to analyze the results of the data grouping. The characteristics of the respondents filling out the questionnaire, as shown in Table 1.

Table 1.
Respondent data characteristics.

Characteristics	Frequency	Percentage
Gender		
Man	103	75.18%
Woman	34	24.82
Total	137	100%
Employment Waiting Period		
<1 Month	42	30.65%
<3 Months	26	18.97%
<6 Months	24	17.51%
<12 Months	22	16.05%
>12 Months	23	16.82%
Total	137	100%

Next, we conduct validity and reliability testing to assess the data quality and determine if the instruments in this questionnaire can accurately measure the intended variables. The instrument is considered quality and can be accounted for if its validity and reliability have been proven [17]. The following are the results of the questionnaire test that has been carried out:

Tabel 2.
The results of the validity test of the 32 questionnaire instruments.

Question No	rxy	r table	Status
1	0.468448597	0.1678	valid
2	0.54375537	0.1678	valid
3	0.708375551	0.1678	valid
4	0.684413143	0.1678	valid
5	0.638263094	0.1678	valid
6	0.667439207	0.1678	valid
7	0.632527803	0.1678	valid
8	0.737420947	0.1678	valid

Question No	rxy	r table	Status
9	0.70561445	0.1678	valid
10	0.635590463	0.1678	valid
11	0.669549295	0.1678	valid
12	0.725092416	0.1678	valid
13	0.766010932	0.1678	valid
14	0.661780393	0.1678	valid
15	0.587108171	0.1678	valid
16	0.491418681	0.1678	valid
17	0.541332054	0.1678	valid
18	0.511319347	0.1678	valid
19	0.47795073	0.1678	valid
20	0.584123228	0.1678	valid
21	0.588728978	0.1678	valid
22	0.572717395	0.1678	valid
23	0.572019477	0.1678	valid
24	0.61745001	0.1678	valid
25	0.580622927	0.1678	valid
26	0.664173648	0.1678	valid
27	0.669499905	0.1678	valid
28	0.61977652	0.1678	valid
29	0.620612121	0.1678	valid
30	0.612712438	0.1678	valid
31	0.601405894	0.1678	valid
32	0.61793568	0.1678	valid

Source: Data processing results

Table 3.
Reliability test results.

Number of Item Variants	Total Variance	Reliability Coefficient (r11)	Interpretation
24.008	31.810	0.950	Very Reliable

Source: Data processing results

Table 4.
Dataset after data integration process.

Data to	IPK	Rx1	Rx2	Rx3	Rx4	Ry
1	3,2	75	71	75	53	100
2	3,7	75	75	50	50	84
3	3,82	81	100	90	81	75
4	3,44	81	75	85	50	66
5	3,53	69	54	65	61	34
....	....	....	....	....	....	....
137	3,65	75	67	40	39	47

Source: Data processing result

Table 5.
Results of multicollinearity test.

Attribute	r	r2	Tolerance	VIF	Status
rIPKrx1	0.3099	0.0960	0.9040	1.1062	non-multicollinearity
rIPKrx2	0.1956	0.0383	0.9617	1.0398	non-multicollinearity
rIPKrx3	0.1103	0.0122	0.9878	1.0123	non-multicollinearity
rIPKrx4	-0.0272	0.0007	0.9993	1.0007	non-multicollinearity
rRx1rx2	0.8164	0.6665	0.3335	2.9989	non-multicollinearity
rRx1rx3	0.6160	0.3795	0.6205	1.6116	non-multicollinearity
rRx1rx4	0.4512	0.2036	0.7964	1.2556	non-multicollinearity
rRx2rx3	0.7136	0.5092	0.4908	2.0377	non-multicollinearity
rRx2rx4	0.4889	0.2390	0.7610	1.3140	non-multicollinearity
rRx3rx4	0.4821	0.2324	0.7676	1.3028	non-multicollinearity

Table 2 shows that the results of the validity test of the 32 questionnaire instruments used in this study are declared valid where the rxy results, when compared with the r table value obtained in the r table value list with a significance level of 0.05 with a 2-sided test where the value is 0.1678. From the results of this comparison, the instrument is declared valid.

Table 3 shows that the results of the reliability test of the reliability coefficient values (r11) show a 0.950, and the value is above 0.80, where it can be concluded that the results of the Cronbach Alfa reliability test on the questionnaire in this study have a very high level of reliability or are very reliable.

Furthermore, data processing is carried out where the data will be integrated according to the elements of learning achievement into attributes Rx1, Rx2, Rx3, Rx4, and the level of suitability of the work they get becomes Ry. One more attribute is added in the form of GPA, which is filled in by the respondents when filling out the questionnaire, as shown in the following table.

After that, a multicollinearity test was carried out to see how strong the relationship/correlation was between the variables by looking for the tolerance value and VIF (variance inflation factor).

Table 5 shows the results of multicollinearity testing where the VIF value is found to be <10 between attributes. So, it can be concluded that the dataset is declared non-multicollinear.

3. Result and Analysis

Through the existing dataset, the optimal K value is sought. Before that, the initial computation process carried out in RStudio first involves the data standardization process to equate unequal data units. Next, it will enter the stage of finding how many optimal clusters from the dataset used.

Figure 2.
Cluster data results.

```
K-means clustering with 3 clusters of sizes 68, 41, 28

Cluster means:
         IPK        rx1        rx2        rx3        rx4         ry
1  0.04359803 -0.1389142 -0.1482809  0.1336863 -0.2475787  0.1360706
2  0.05909966  0.9359578  1.0431686  0.7942361  0.9305440  0.6733269
3 -0.19241973 -1.0331467 -1.1673862 -1.4876554 -0.7613198 -1.3164001

Clustering vector:
  [1] 1 1 2 1 3 2 2 1 3 1 3 2 1 2 1 1 1 1 1 2 3 2 1 3 3 3 1 2 1 2 1 3 2 3 3 1 2 2 2 1 1 2 2 2 1 2 3 1 2 1 1 3
 [53] 1 3 1 1 1 1 2 2 1 3 1 1 2 2 2 1 1 1 3 1 1 1 1 1 1 1 1 3 1 1 1 1 3 1 1 1 1 2 1 1 2 1 3 2 2 1 3 1 3 2 1 2
[105] 1 1 1 1 1 2 3 2 1 3 3 3 1 2 1 2 1 3 2 3 3 1 2 2 2 1 1 2 2 2 1 2 3

Within cluster sum of squares by cluster:
[1] 190.7223 125.5454 133.7238
 (between_SS / total_SS =  44.9 %)

Available components:

[1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss" "betweenss"    "size"
[8] "iter"         "ifault"
```

Source: Data processing results on RStudio

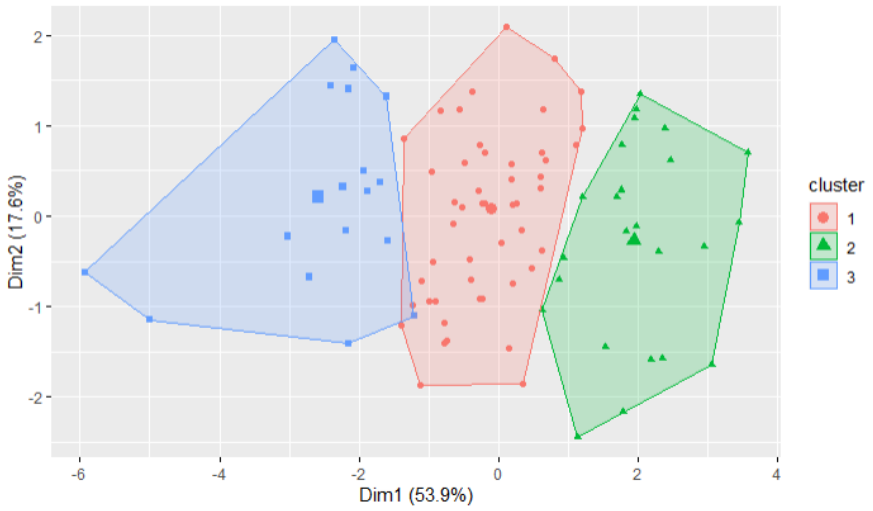
As shown in Figure 1, shows that through the Elbow model in RStudio, the optimal K value is in cluster 2; this is indicated by the steep angle created and with a more stable decrease in the Sum of Square value at the 3rd K value. Likewise, the number of optimal clusters is shown using the Silhouette model, where the optimal cluster for the data used in this study is 2. However, in this case, the author decided to see the cluster results in more detail and make it easier to analyze, so the author chose to use 3 clusters.

Figure 2 shows the results of the computational processing that has been done where the data has been divided into three clusters with the composition of the first cluster with as many as 68 members, the second cluster with as many as 41 members, and the third cluster as many as 28 members. The cluster success rate from this process was also found to be 44.9%. The visualization results of the distribution of members from this clustering, as shown in Figure 3.

From the results of data grouping that has been done using the K-Means algorithm, it was found that the first cluster with 68 members (49.63%) of the total respondents has the characteristics of members who have high values in the three elements of learning achievement of graduates from the UMM Informatics Engineering study program represented by attributes Rx1, Rx2, Rx3 (attitude, general skills, and knowledge), while for attribute Rx4 (specific skills) members of this cluster partly get sufficient values. In addition, members of this cluster, through the Ry value as an attribute for measuring job suitability, get high values where it can be said that this first cluster is a cluster of graduates who get jobs that are relevant to the field of Informatics Engineering even though their mastery of their specific skills is lacking.

Meanwhile, the second cluster with 41 members (29.92%) of the total respondents has the characteristics where they fulfill the four elements of learning achievement of graduates set by the UMM Informatics Engineering study program and get jobs relevant to what they learned during college.

Figure 3.
Visualization of clustering data results.



Source: Data processing results on RStudio

On the other hand, the third cluster, with 28 members (20.45%) of the total respondents, has characteristics of members who meet the learning outcomes of graduates for the three elements (Rx1, Rx2, Rx3). Still, their special skills element (Ry) is low, and they get less relevant jobs in Informatics Engineering.

4. Conclusion

Through this study, it can be concluded that K-Means divides the data into three clusters against the job relevance data used. By grouping the cluster results, it can be seen that 29.92% of UMM Informatics Engineering graduates were able to meet the learning outcomes of graduates and get jobs that are relevant to what they learned, 49.63% of other graduates also got jobs that were relevant to their majors even though they did not master the specific skills measured by the learning outcomes of graduates, and 20.45% of other graduates got jobs that were less relevant to the field of Informatics engineering.

In terms of performance, the K-Means algorithm applied in this study with a total of 3 clusters formed gave a success rate of 44.9%, which is to the statement [3] where random centroid selection is one of the main weaknesses of this algorithm. So further research and development of the K-means algorithm version are needed.

The implications of this study in real-world cases suggest that educational institutions, like the University of Muhammadiyah Malang (UMM), can use K-Means clustering to assess and improve the relevance of their programs to job market demands. By analyzing graduate job placement in relation to their learning outcomes, universities can identify areas where specific skills are lacking and adjust their curriculum to better align with industry needs. Additionally, the results show that while a significant portion

of graduates secure relevant jobs, there remains a gap in skills mastery, emphasizing the importance of targeted skill development. Furthermore, the algorithm's performance limitation, due to random centroid selection, indicates that industries utilizing K-Means clustering for job-market analysis or other data-driven tasks should be cautious and consider improvements or alternative methods for more accurate clustering.

Future work on this topic could focus on improving the K-Means algorithm by integrating advanced techniques such as K-Means++ to address the limitations of random centroid selection and improve clustering accuracy. Additionally, incorporating more diverse variables like industry-specific demands, soft skills, and long-term career progression could enhance the depth of analysis. Exploring alternative distance metrics, such as Manhattan or cosine similarity, may offer further insights into clustering performance. Real-time data analysis systems could be developed to continuously monitor and adjust to evolving job market conditions. Furthermore, applying the clustering approach to other academic programs and conducting longitudinal studies would provide a broader understanding of job relevance over time, offering significant value to educational institutions and industries alike.

References

[1] Badan Pusat Statistik, "Jumlah Perguruan Tinggi1, Dosen, dan Mahasiswa2 (Negeri dan Swasta) di Bawah Kementerian Pendidikan, Kebudayaan, Riset, dan Teknologi Menurut Provinsi, 2022." [Online]. Available: https://www.bps.go.id/id/statistics-table/3/Y21kVGRHNXZVMEl3S3pCRlIyMHJRbnB1WkVZemR6MDkjMw==/jumlah-perguruan-tinggi--tenaga-pendidik-dan-mahasiswa-negeri-dan-swasta--di-bawah-kementerian-riset--teknologi-dan-pendidikan-tinggi-kementerian-pendidikan-dan-kebudayaan-menurut-provinsi--2022.html?year=2022

[2] M. Ahmed, R. Seraj, and S. M. S. Islam, "The k-means Algorithm: A Comprehensive Survey and Performance Evaluation," *Electronics*, vol. 9, no. 8, p. 1295, Aug. 2020, doi: 10.3390/electronics9081295.

[3] A. Maghawry, Y. Omar, and A. Badr, "Self-Organizing Map vs Initial Centroid Selection Optimization to Enhance K-Means with Genetic Algorithm to Cluster Transcribed Broadcast News Documents," *Int. Arab J. Inf. Technol.*, vol. 17, no. 3, pp. 316–324, May 2020, doi: 10.34028/iajit/17/3/5.

[4] X. Wu *et al.*, "Top 10 algorithms in data mining," *Knowl. Inf. Syst.*, vol. 14, no. 1, pp. 1–37, Jan. 2008, doi: 10.1007/s10115-007-0114-2.

[5] N. Wang, G. Guo, B. Wang, and C. Wang, "Traffic clustering algorithm of urban data brain based on a hybrid-augmented architecture of quantum annealing and brain-inspired cognitive computing," *Tsinghua Sci. Technol.*, vol. 25, no. 6, pp. 813–825, Dec. 2020, doi: 10.26599/TST.2020.9010007.

[6] L. ChaoYing, W. X. Da, and Z. E. Hui, "Research on modeling of government debt risk comprehensive evaluation based on multidimensional data mining," *Soft Comput.*, vol. 26, no. 16, pp. 7493–7500, Aug. 2022, doi: 10.1007/s00500-021-06478-7.

[7] Z. Wang, "Higher Education Management and Student Achievement Assessment Method Based on Clustering Algorithm," *Comput. Intell. Neurosci.*, vol. 2022, no. 1, pp. 1–10, Jul. 2022, doi: 10.1155/2022/4703975.

[8] L. Sun, "Research on Mining Balanced Competition Strategy in Financial Market Based on Computer Data Mining Method," *Mob. Inf. Syst.*, vol. 2022, pp. 1–8, Jul. 2022, doi: 10.1155/2022/6202890.

[9] M. A. Wahyu Saputra and S. Harini, "Java Island Health Profile Clustering using K-Means Data Mining," *Int. J. Inf. Commun. Technol.*, vol. 8, no. 1, pp. 1–9, Jul. 2022, doi: 10.21108/ijoict.v8i1.606.

[10] H. D. Tampubolon, D. Gultom, L. Y. Hutabarat, F. I. R.H Zer, and D. Hartama, "Penerapan Algoritma K-Means untuk Mengetahui Tingkat Tindak Kejahatan Daerah Pematangsiantar," *J. Teknol. Inf.*, vol. 4, no. 1, pp. 146–151, Jun. 2020, doi: 10.36294/jurti.v4i1.1263.

[11] Z. Zhang and A. Kusiak, "Monitoring Wind Turbine Vibration Based on SCADA Data," *J. Sol. Energy Eng.*, vol. 134, no. 2, May 2012, doi: 10.1115/1.4005753.

[12] D. Kumalasari, A. B. W. Putra, and A. F. O. Gaffar, "Speech classification using combination virtual center of gravity and k-means clustering based on audio feature extraction," *J. Inform.*, vol. 14, no. 2, p. 85, May 2020, doi: 10.26555/jifo.v14i2.a17390.

[13] A. Triayudi, Iksal, and R. Haerani, "Data Mining K-Means Algorithm for Performance Analysis," *J. Phys. Conf. Ser.*, vol. 2394, no. 1, p. 012031, Dec. 2022, doi: 10.1088/1742-6596/2394/1/012031.

[14] A. K. Jain, "Data clustering: 50 years beyond K-means," *Pattern Recognit. Lett.*, vol. 31, no. 8, pp. 651–666, Jun. 2010, doi: 10.1016/j.patrec.2009.09.011.

[15] David and MacKay, *"Chapter 20. An Example Inference Task: Clustering."* Cambridge University Press, 2003.

[16] R. Asih, D. Alonzo, and T. Loughland, "The critical role of sources of efficacy information in a mandatory teacher professional development program: Evidence from Indonesia's underprivileged region," *Teach. Teach. Educ.*, vol. 118, p. 103824, Oct. 2022, doi: 10.1016/j.tate.2022.103824.

[17] C. A. W. Heryanto, C. S. F. Korangbuku, M. I. A. Djeen, and A. Widayati, "Pengembangan dan Validasi Kuesioner untuk Mengukur Penggunaan Internet dan Media Sosial dalam Pelayanan Kefarmasian," *Indones. J. Clin. Pharm.*, vol. 8, no. 3, Sep. 2019, doi: 10.15416/ijcp.2019.8.3.175.