Journal of Computer Networks, Architecture and High Performance Computing
Volume 6, Number 3, July 2024
https://doi.org/10.47709/cnapc.v6i3.4214

Submitted : 2 July 2024
Accepted : 20 July 2024
Published : 22 July 2024

Industry Class Clustering of Software Expertise Competency at SMKN 2 Kraksaan Using Constrained K-Means Clustering Algorithm

Matlubul Khairi[1)], M. Syafiih[2)*], Ahmad Khairi[3)]
[1)2)*3)]Universitas Nurul Jadid, Indonesia
[1)]sangrato88@gmail.com, [2)*]m.syafii@unuja.ac.id, [3)]khairi@unuja.ac.id

ABSTRACT

Addressing the gap between school education and industry needs is a recurring concern, as many graduates struggle to enter the workforce due to lacking practical skills. Industry Classes aim to bridge this gap by preparing students with relevant skills and knowledge aligned with real-world industry demands. This study focuses on the application of Constrained K-Means Clustering to categorize students in the software engineering competency classes at SMKN 2 Kraksaan. This algorithm modifies traditional K-Means by utilizing Linear Programming Algorithm (LPA), ensuring each cluster meets predefined subject requirements. The research involves analyzing academic proficiency test data (TKDA) from 96 X-grade students, evaluating their abilities in analogy, series, figural, mathematical, and recall skills. Using a 3-cluster approach, each with 32 to 60 student capacity constraints, the study aims to optimize student distribution for effective learning outcomes. Evaluation through silhouette method yielded a score of 0.3199, indicating satisfactory separation between clusters with overlap to address. Cluster analysis revealed Cluster 2 as the most proficient, showcasing strengths in recall and series attributes critical for software engineering. These findings suggest that Constrained K-Means Clustering is effective in classifying students, highlighting Cluster 2 as optimal for software engineering competencies at SMKN 2 Kraksaan. Future research should focus on enhancing data quality, expanding sample size, and refining algorithms for improved clustering accuracy and effectiveness.

Keywords: Constrained K-Means Clustering; Linear Programming Algorithm (LPA); Academic Proficiency Test; Software Engineering Competency

1. INTRODUCTION

The gap between education in schools and the needs of industry is an issue of concern. Many graduates find it difficult to jump straight into the workforce due to a lack of practical skills required by the industry (Alboaouh, 2018). In an effort to address this issue, the Industrial Class concept comes as a solution to bridge the gap. The Industrial Class aims to prepare students with relevant skills and knowledge so that they are ready to face challenges in the world of work (Judijanto, Suharyono, & Wahyudi, 2024). Through collaboration between schools and companies, the Industrial Class curriculum is tailored to the real needs of the industrial world.

However, not all students can immediately join the Industrial Class program due to class limitations (Forestyanto, Syamwil, & Wijaya, 2019). These limitations can be in the form of the number of seats, teaching resources, and supporting facilities available. Therefore, student selection is an important step to ensure that only students who truly meet the criteria can join the program (Mengash, 2020). This selection also aims to maximize the effectiveness of learning, so that the selected students are those who have high interest and potential to develop in the intended industrial field (Elfrianto, Nasrun, & Arifin, 2023).

Constrained K-Means Clustering is a modified algorithm of the traditional K-Means formula, which in performing its function approaches the Linear Programming Algorithm (LPA) (Risal, Zainuddin, & Niswar, 2022). This algorithm requires each cluster to have a predetermined subject, so as to overcome the imbalance in the cluster formed by K-Means (Bibi, Alqahtani, & Ghanem, 2023). Imbalanced data is a condition in a data set where the amount of data in a certain class (majority class) is more than the amount of data in another class (minority class) (Lampert, Nickisch, & Harmeling, 2018). Thus, Constrained K-Means Clustering provides a solution to ensure a more even distribution of data in each resulting cluster. The Constrained Clustering method has an advantage in terms of less computational time compared to the traditional K-Means algorithm. This method has competitive performance with state-of-the-art

* Corresponding author

Journal of Computer Networks, Architecture and High Performance Computing
Volume 6, Number 3, July 2024
https://doi.org/10.47709/cnapc.v6i3.4214

Submitted : 2 July 2024
Accepted : 20 July 2024
Published : 22 July 2024

Constrained Clustering methods for large datasets and requires much less computational time (Huang, Yao, Hao, Peng, & Guo, 2021).

This study aims to conduct clustering using the Constrained K-Means Clustering algorithm to ensure a balanced and targeted distribution of students in the Industrial Class program. Through this approach, it is expected that each selected group of students has a fair and proportional composition according to the specified criteria, so as to optimize the effectiveness of learning. The Constrained K-Means Clustering algorithm will help overcome the problem of data imbalance by ensuring each cluster has an even number of students, and facilitate a more efficient selection process. This research also aims to test the performance and superiority of the algorithm in the context of student selection, so as to provide recommendations for the application of this method in the Industrial Class program of the software engineering expertise program at SMKN 2 Kraksaan.

2. LITERATURE REVIEW

The research of Francesco Alesiani, Gulcin Ermis & Konstantinos Gkiotsalitis (2022) entitled "Constrained Clustering for the Capacitated Vehicle Routing Problem (CC-CVRP)" discusses solving the Capacitated Vehicle Routing Problem (CVRP) on a large scale faced by logistics, shipping and e-commerce planners.CVRP is an NP-Hard problem that is difficult to solve optimally for large problem instances.This research utilizes the Clustering for the CC-CVRP approach, which is an efficient version of clustering that considers the constraints of the original problem to transform it into a more solvable version.This approach results in a clustered vehicle routing problem with fewer decision variables. This approach successfully reduces the computational complexity associated with solving CVRP. This research provides a better solution for large-scale instances of the CVRP problem in a short period of time.

Uraiwan Buatoom, Waree Kongprawechnon and Thanaruk Theeramunkong (2020) in a study that discusses one of the problems that occur when using K-Means is that the document clustering problem with K-Means is NP-hard, which means it is difficult to find a globally optimal solution. K-Means tends to find the best local solution, which may not always achieve globally optimal results. Moreover, in the context of this study, the use of distribution-based term weighting as a distance constraint may also affect the performance of K-Means in correctly clustering documents. Compared with conventional TFIDF, distribution-based term weighting improves centroid-based, seeded k-means, and k-means methods with error reduction rates of 22.45%, 31.13%, and 58.96% respectively. The experimental results demonstrate the effectiveness of term weighting in document clustering using six different text collections.

Research conducted by Andi Alviadi Nur Risal entitled "School Zoning System for Student Admission using Constrained K-Means Algorithm" analyzes school zoning based on the closest distance between student domicile and school location using Constrained K-Means Algorithm. The dataset used is 22 school locations and 2248 student location data. the method used is Constrained K-Means to group prospective new students into each school. The Constrained K-Means algorithm works based on the value of K as the cluster center closest to the value of N (cluster members) with the Linear Programming Algorithm (LPA) approach so that each cluster has a balanced N member. Based on the results of trials conducted, the Constrained K-Means algorithm has an accuracy of 95.35% compared to the K-Means algorithm with the accuracy achieved between cluster members and the cluster center of only 73.93%.

This research addresses the gap between industry needs and school education by using Constrained K-Means clustering algorithm to group students into industry classes based on their competencies. Unlike traditional clustering approaches that tend to produce the best local solution and are not globally optimal, Constrained K-Means ensures that each cluster has a balanced composition and is in line with industry needs. Thus, this research not only optimizes the distribution of students in industrial classes, but also provides a fairer and more efficient solution for the selection of students who will join the industrial class program, especially in the field of software engineering competence at SMKN 2 Kraksaan.

This research has several similarities and differences with related research that has been done before. One of the similarities is the use of clustering techniques to group entities, be it students, documents, or logistical data. Like the research of Francesco Alesiani et al. (2022) who used clustering to solve vehicle routing problems, and Andi Alviadi Nur Risal who applied Constrained K-Means in school zoning, this research also adopted the Constrained K-Means method to group students based on their competencies. Similar to Andi Alviadi Nur Risal's research, this research focuses on clustering by considering certain constraints, so that each cluster has a balanced and relevant composition.

However, there are significant differences in the application context and methods used. Research by Francesco

* Corresponding author

Journal of Computer Networks, Architecture and High Performance Computing
Volume 6, Number 3, July 2024
https://doi.org/10.47709/cnapc.v6i3.4214

Submitted : 2 July 2024
Accepted : 20 July 2024
Published : 22 July 2024

Alesiani et al. (2022) focuses on the vehicle routing problem in a logistics context, while Uraiwan Buatoom et al. (2020) focused on document clustering using distribution-based term weighting. In contrast, this research and Andi Alviadi Nur Risal's research focus on the educational context, specifically in clustering students based on distance and competency. In addition, this study uses a local dataset from SMKN 2 Kraksaan, which includes local student and school location data, in contrast to the logistics and document datasets used in other studies.

3. METHOD

This research uses the constrained K-means clustering algorithm to group new students in the software engineering industry class. The research stages can be seen in Figure 1.

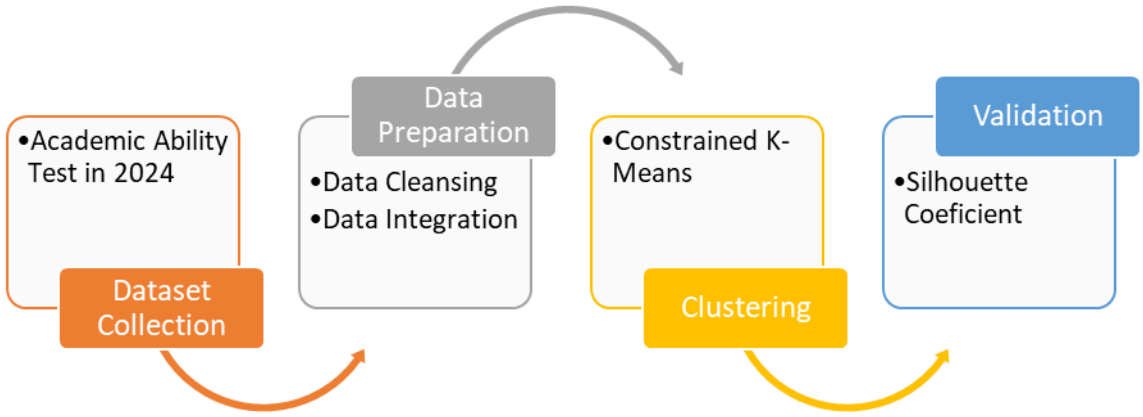

Fig.1 Model Proposed

Dataset

This research dataset is the result of the basic academic ability test (TKDA) of class X students in the software expertise competency of SMKN 2 Kraksaan. Basic academic tests include analogy skills, sequence skills, figural skills, math skills, and memory skills. The data obtained is 96 data as shown in Figure 2.

	NO	NAME	CLASS	FIGURAL	SEQUENCE	REMEMBERING	MATH	ANALOGY
0	1	ADE RAKA MUHAIMIN	X-RPL-1	NaN	66.67	86.67	20.0	53.33
1	2	ADE RAKA MUHAIMIN	X-RPL-1	30.0	66.67	93.33	NaN	NaN
2	3	AFANDI SAHRONI	X-RPL-3	30.0	66.67	86.67	20.0	26.67
3	4	AFIF WAHYU FAIR SAFITRI	X-RPL-2	50.0	80.00	80.00	40.0	46.67
4	5	AHMAD DANIYAL HAKIM	X-RPL-3	60.0	66.67	66.67	30.0	13.33
...	...	...	...	...	...	...	...	...
91	92	ULIN NIKMAH FARAH KHOIRUN NISA	X-RPL-1	60.0	80.00	100.00	40.0	13.33
92	93	VIQI INDRA MAULANA	X-RPL-1	70.0	80.00	66.67	50.0	40.00
93	94	WIWIK	X-RPL-1	50.0	60.00	86.67	10.0	33.33
94	95	YUANDA FAJAR APRIANDA	X-RPL-3	70.0	46.67	93.33	NaN	NaN
95	96	YUNITA PUSPA NINGTIAS	X-RPL-3	30.0	66.67	73.33	40.0	NaN

Fig. 2 Dataset TKDA Result

Data Preparation

In the process of addressing issues found within a dataset, cleansing techniques involve analyzing the quality of data by modifying, correcting, or removing data that does not meet the research requirements (Noor, Kusumasari, &

* Corresponding author

Journal of Computer Networks, Architecture and High Performance Computing
Volume 6, Number 3, July 2024
https://doi.org/10.47709/cnapc.v6i3.4214

Submitted : 2 July 2024
Accepted : 20 July 2024
Published : 22 July 2024

Hasibuan, 2019). Once the data has been corrected, it is then fed into the modeling process with the expectation of producing a robust model.

Clustering

This stage involves creating a clustering model to group industry competency classes for software expertise at SMKN 2 Kraksaan using the Constrained K-Means clustering algorithm. Constrained K-Means is a modified algorithm based on the traditional K-Means formula, which incorporates the Linear Programming Algorithm (LPA). It enforces that each cluster must contain predefined subjects (Melnykov, & Melnykov, 2020). The steps for determining the K-Means algorithm with the Linear Programming Algorithm (LPA) approach are as follows (Baumann, 2019):

a. Select initial cluster centers (k) either through random sampling or from the Constrained K-Means solution.
b. Use the Linear Programming Algorithm (LPA) procedure to find the optimal clustering with constraints that require each cluster to contain a minimum of 2 subjects.
c. Update the cluster centers based on the results of the Linear Programming Algorithm (LPA).
d. d. If there are still changes in cluster membership, repeat steps b and c.
e. e. Repeat steps a - e for a number of initial object datasets. The cluster solution with the minimum objective value will be the final solution.

Evaluation

This research uses the Silhouette calculation method to determine the quality of the clustering model. The Silhouette method measures how similar each object is to other objects in its own cluster compared to objects in other clusters (Mulyani, Setiawan, & Fathi, 2023). The results of the Silhouette calculation provide a value that helps in evaluating how well the data has been grouped in the clustering model used.

4. RESULT

Data Analysis Tool Using Python Programming with pandas, numpy, k_means_constrained, and sklearn Libraries. In the data preparation stage, detection of missing values and duplicate data is performed on each attribute.

```
NO           0
NAMA         0
KELAS        0
FIGURAL      3
DERET        1
MENGINGAT    3
MATEMATIKA   7
ANALOGI      4
dtype: int64
```

Fig. 3 Missing Value Detection

The data preparation results in the use of 5 attributes for clustering: figural, series, recall, mathematics, and analogy, totaling 96 data points (rows). In the modeling stage, Constrained K-Means Clustering with 3 clusters is used, with a minimum constraint of 32 and a maximum of 60. The choice of 3 clusters aligns with X-grade class groups with a capacity of 32 students. The clustering determination in the model can be seen in Figure 4.

* Corresponding author

Journal of Computer Networks, Architecture and High Performance Computing
Volume 6, Number 3, July 2024
https://doi.org/10.47709/cnapc.v6i3.4214

Submitted : 2 July 2024
Accepted : 20 July 2024
Published : 22 July 2024

```python
# Menjalankan Constrained K-Means dengan 3 cluster
kmeans = KMeansConstrained(n_clusters=3, size_min=32, size_max=60, random_state=0)
kmeans.fit(x)
labels = kmeans.labels_

# Menambahkan label cluster ke data asli
data['cluster'] = labels
```

Python

Fig. 4 Modeling with Constrained K-Means

The above model produces a total of 3 clusters with each cluster containing 32 data. The distribution of industry class clustering evenly is shown in Figure 5.

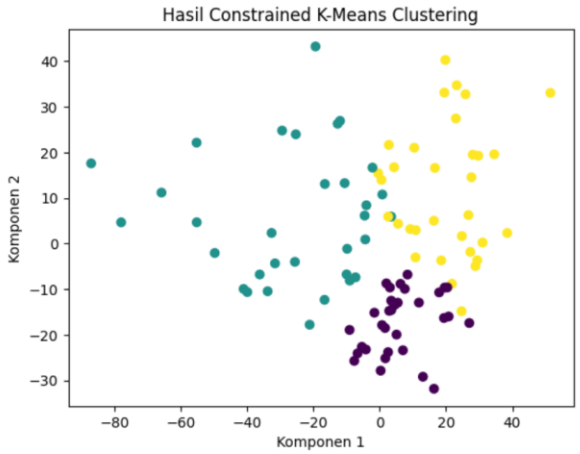

Fig. 5 Data distribution of clustering results

The distribution in Figure 5 evaluated using the silhouette method produces a value of 0.3199, which indicates that although there is a fairly good separation between clusters, there is still overlap between clusters that needs to be considered. This indicates that some learners have similar characteristics to members of other clusters. From the clustering results, the average score on each attribute is shown in table 1.

Table 1
Average Cluster Value of Each Attribute

Cluster	Figural	Squence	Remembering	Math	Analogy	Mean
Cluster 0	50	67	91	20	38	53
Cluster 1	48	53	54	26	38	44
Cluster 2	56	75	86	52	38	61

Cluster 0 shows strengths in the Remembering attribute, but weaknesses in Math. This may indicate that learners in Cluster 0 have good memory ability but need improvement in math skills. Cluster 1 has the lowest mean score among the three clusters, suggesting that learners in this cluster may need more support and additional learning in various attributes. Cluster 2 is the best performing cluster overall, especially in the attributes of Sequences and Remembering. This suggests that learners in Cluster 2 have good analytical and memory skills.

5. DISCUSSIONS

Based on the analysis of the average value of various attributes, Cluster 2 is the best choice to be proposed as the industry class for software competency. This cluster has the highest average score (61) compared to Cluster 0 (53) and

* Corresponding author

Journal of Computer Networks, Architecture and High Performance Computing
Volume 6, Number 3, July 2024
https://doi.org/10.47709/cnapc.v6i3.4214

Submitted : 2 July 2024
Accepted : 20 July 2024
Published : 22 July 2024

Cluster 1 (44). Cluster 2 excels in the Remembering (86) and Series (75) attributes, which are essential for understanding and applying programming concepts and algorithms. The moderately high Math score (52) also indicates good analytical skills, which are crucial in software development. With consistently high scores across a range of attributes, Cluster 2 shows even and stable competencies, making it the most prepared group for the software field.

6. CONCLUSION

Overall, Constrained K-Means Clustering is effective in grouping students based on basic academic ability tests using 3 clusters. Evaluation results using the silhouette method with a value of 0.3199 show that there is still overlap between clusters. It is recommended to increase the amount and quality of data used and develop adaptive algorithms to improve the accuracy and effectiveness of clustering. With these steps, it is expected that more precise modeling and more optimal results can be achieved in the future.

7. REFERENCES

Alboaouh, K. (2018). The Gap Between Engineering Schools and Industry: A Strategic Initiative. 2018 IEEE Frontiers in Education Conference (FIE), 1-6. https://doi.org/10.1109/FIE.2018.8659314.\

Alesiani, F., Ermis, G., & Gkiotsalitis, K. (2022). Constrained clustering for the capacitated vehicle routing problem (cc-cvrp). Applied artificial intelligence, 36(1), 1995658.

Baumann, P. (2019). A Binary Linear Programming-Based K-Means Approach for the Capacitated Centered Clustering Problem. 2019 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM), 335-339. https://doi.org/10.1109/IEEM44572.2019.8978840.

Bibi, A., Alqahtani, A., & Ghanem, B. (2023). Constrained clustering: General pairwise and cardinality constraints. IEEE Access, 11, 5824-5836.

Buatoom, U., Kongprawechnon, W., & Theeramunkong, T. (2020). Document clustering using K-means with term weighting as similarity-based constraints. Symmetry, 12(6), 967.

Elfrianto, H., Nasrun, M. S., & Arifin, M. (2023). Buku Ajar Manajemen Pendidikan. umsu press.

Forestyanto, Y., Syamwil, R., & Wijaya, M. (2019). School Constraints in Recruitment and Implementation of Industrial Classes. , 4. https://doi.org/10.15294/JVCE.V4I1.21766.

Huang, P., Yao, P., Hao, Z., Peng, H., & Guo, L. (2021). Improved constrained k-means algorithm for clustering with domain knowledge. Mathematics, 9(19), 2390.

Judijanto, L., Mayasari, N., Baruno, Y. H. E., Tasrip, T., & Rusdi, M. (2024). Analisis Pengaruh Kemitraan Sekolah-Industri dan Program Magang terhadap Keterampilan Kerja dan Kesiapan Karier Siswa SMK di Jawa Tengah. Jurnal Multidisiplin West Science, 3(03), 378-388.

Lampert, T., Dao, T., Lafabregue, B., Serrette, N., Forestier, G., Crémilleux, B., Vrain, C., & Gançarski, P. (2018). Constrained distance based clustering for time-series: a comparative and experimental study. Data Mining and Knowledge Discovery, 32, 1663 - 1707. https://doi.org/10.1007/s10618-018-0573-y.

Melnykov, I., & Melnykov, V. (2020). A Note on the Formal Implementation of the K-means Algorithm with Hard Positive and Negative Constraints. Journal of Classification, 37, 789-809. https://doi.org/10.1007/s00357-019-09349-x.

Mengash, H. (2020). Using Data Mining Techniques to Predict Student Performance to Support Decision Making in University Admission Systems. IEEE Access, 8, 55462-55470. https://doi.org/10.1109/ACCESS.2020.2981905.

Mulyani, H., Setiawan, R., & Fathi, H. (2023). Optimization of K Value in Clustering Using Silhouette Score (Case Study: Mall Customers Data). Journal of Information Technology and Its Utilization. https://doi.org/10.56873/jitu.6.2.5243.

Noor, S., Kusumasari, T., & Hasibuan, M. (2019). Data Cleansing with PDI for Improving Data Quality. Proceedings of the International Conference on Creative Economics, Tourism and Information Management. https://doi.org/10.5220/0009868102560261.

Risal, A. A. N., Zainuddin, Z., & Niswar, M. (2022). School Zoning System for Student Admission using Constrained K-Means Algorithms. In 2022 IEEE International Conference on Communication, Networks and Satellite (COMNETSAT) (pp. 174-178). IEEE.

* Corresponding author