Professional Clustering Based on the Graduates Profile Using K-Means Method

Susi 1,*, Alyauma Hajjah 1

1 Institut Bisnis dan Teknologi Pelita Indonesia; Jl. Jend. Ahmad Yani No.78-88 Pekanbaru; e-mail: susi177234@gmail.com, alyauma.hajjah@lecturer.pelitaindonesia.ac.id

* Corespondence Author: e-mail: alyauma.hajjah@lecturer.pelitaindonesia.ac.id

Received: 25 January 2021; Revised: 09 February 2021; Accepted: 28 February 2021; Available online: 24 March 2021

Abstract

The Information Systems department of Pelita Indonesia Institute of Business and Technology produces graduates with education knowledge to face the professional work, but the majority of students after graduating do not work according to their graduate profiles/educational background. This research aimed to help students in determining the appropriate graduate profile. The proposed system was built using the K-means method for the classification process of graduate profiles; hence, the results can be used as recommended profession to be taken. The data used in this study is the data of students of class 2016, and these results are compared with their current professional record data with the aim of knowing the percentage of professional suitability obtained with the current profession. After this research is tested, the results of the classification of the graduate profile can be obtained where there are 10 students in the administrator database cluster, 11 students of the web design and developer cluster, one student of the cluster information system manager,
and 8 students of the cluster system analysis. The percentage of suitability was 43.33%. This program is designed using the PHP programming language.

**Keywords**: clusters, k-means, graduate profiles, jobs, professions

1. **Introduction**

The development of education in Indonesia has increased rapidly. An important things that becomes a success factor of education in a university is the profession skill that is obtained after graduating. Undergraduate graduates from higher education are expected to work according to their educational background. However, the level of conformity between the competence of undergraduate graduates and the needs of the world of work is still low.

Currently, most educational institutions are taking advantage of advances in information and communication technology in order to improve the quality of services provided. Good data organization and management will provide benefits for educational institutions, especially in the efficiency of existing resources. In addition, the speed in producing and distributing information greatly supports the quality of service. After graduating from a college, getting a diploma certificate is a temporary success, because it is just one requirement to apply for jobs. The biggest challenge for graduates is the unavailability of information as a support for choosing any profession to work in accordance with the majors they have and according to the value of the courses obtained.

To solve this problem, the K-means clustering method will be implemented. The input used in the K-means method is the value of each related subject, graduate profile, and student data. After getting the results of the professional classification, they will then be compared with the professions undertaken by the graduates with a tracer study. Tracer study is a method used by some universities to obtain feedback from graduates. This research was conducted to help students or graduates to determine their profession using the k-means method. With the proposed system, a college can get more accurate classification of the student profession.
Various studies related to the k-means method have been carried out, for example a study conducted by (Kusumantara et al., 2019) where the K-Means method is used to determine student achievement groupings, a research by (Dzulhaq & Imani 2015) in determining the concentration for students using the sample as the main requirement to obtain data that describes as the focus of the research using the Slovin formula. The K-Means method is the well-known clustering algorithm and has been widely used in various fields because of its simplicity, easy to implement and has the ability to cluster large data (Agusta, 2007).

2. Research Methods
2.1 System Development Life Cycle
Data are collected from the IBT Pelita Indonesia system which is currently running. The professional classification system still runs manually, no online system is used to help campuses or students in finding professional classifications based on graduate profiles, hence, students face problems in finding their appropriate professions based on graduate profiles.

![SDLC Diagram](image.png)

Figure 1. SDLC

After the first phase, an Information System development project is determined. In second phase, initiation actions are carried out related to the commencement of development. The initial action is to examine more deeply
the problems experienced at IBT Pelita Indonesia, i.e. the output, the forms and types of data, information workflow, and procedures. The boundaries and the scope of the system is described as well as data management, distribution, and data security in Information System process.

In the third phase, the proposed system is divided into sub-phases consisted of a sub-phase, i.e. determination of the computer-based system criteria, a structuring sub-phase, and designing an online computer-based system model as an option.

In the fourth phase, a description of all the functions have been selected in the analysis stage are described separately from certain computer specifications, based on functions and detailed specifications of all system elements (data, process, input and output). In this logic design, coding or rules are made to determine the profile of graduates and also the appropriate profession.

In the fifth phase, changes are made from the logic design stage to the use of certain technologies in detail, such as designing user-friendly data input forms and reports (login forms, data input, etc.) as well as some reports needed in this system.

The sixth phase of implementation can be interpreted as a process to ensure the implementation of a policy and the achievement of that policy. In this phase, a system including the operating system installation is carried out in the field of study section, using the proposed application program, and provide training to the potential users.

The last stage, a periodic maintenance to check whether the system is running well should be conducted. The process of monitoring, evaluation, and changes (improvements) to the system are carried out if needed. Latest version of software or with updates for documentation, training and support. Changes will be made if there is an error, so that the software and hardware must be adjusted again to accommodate the desired changes.

2.2. K-Means Methods

The K-Means algorithm is an iterative clustering algorithm that partitions data sets into a predefined number of K clusters. The K-Means algorithm is
simple in the implementation, relatively fast, adaptable, and have been widely used in practice. Historically, K-Means has been one of the most important algorithms in data mining. The following are the steps for the K-Means (Sibuea & Safta, 2017):

a. Determining the number of clusters (k) to be formed.

\[ SSE = \sum_{i=1}^{k} \sum_{x_i \in S_i} \| x_i - C_i \|^2 \]  

(1)

b. Find k cluster center point (centroid) early done in a manner random. Determination on centroid early done random or random of available objects as many as k cluster, to count centroid the next cluster, the following formula is used:

\[ v = \frac{\sum_{i=1}^{k} x_i}{n} ; i = 1, 2, 3, \ldots, n \]  

(2)

c. Calculates the distance from each object to each - respectively centroid from each cluster use Euclidean Distance, with the following formula:

\[ d(x, y) = \| x - y \| = \sqrt{\sum_{i=1}^{M} (x_i - y_i)^2} ; i = 1, 2, 3, \ldots, n \]  

(3)

d. Allocate from each object into centroid the closest. For allocating objects into each-respectively cluster at the time iteration is generally done with the hard k-means way in which each object is explicit stated as a member cluster by measuring the proximity distance its nature to the center point cluster the

e. Do iteration and then determine the position of the centroid by using equations.

f. Repeat step three if in position centroid new is not the same

3. Results and Discussion

3.1. Data Processing

At this stage, data processing is carried out to get results that useful in determining the classification of the profession. In general, the K-Means algorithm has some typical steps as follows (Sibuea & Safta, 2017): i) Determine the number of clusters (k), ii) Randomly determine the centroid, iii)
Calculate the centroid distance, iv) After iterating, does the centroid change, if yes, recalculate the data distance from the centroid. If not, then it's done. Iteration is stopped, and v) Grouping the data based on the closest distance. Therefore, the first thing to do is determine the number of k in this study to determine $k = 4$ and the determination of the centroid. At the first time, the centroid is done randomly and then followed by calculating the distance between the centroids.

| Courses | SIDN | MPS | AOK | RPL | ASD2 | SBD | PBD | PGR | PWM | C |
|---------|------|-----|-----|-----|------|-----|-----|-----|-----|---|
|         | 57004| 2   | 3   | 3   | 3    | 3   | 3   | 3   | 3   | C1|
|         | 57011| 3   | 2   | 2   | 3    | 2   | 3   | 3   | 3   | C2|
|         | 57007| 3   | 3   | 3   | 3    | 3   | 3   | 3   | 3   | C3|
|         | 57016| 3   | 3   | 3   | 3    | 3   | 2   | 3   | 3   | C4|

Source: Research Result (2021)

In this study the K-Means method was used to allocate data into a cluster that is closest to the center point of each cluster. To find out which cluster is closest to the data, it is necessary to calculate the distance of each data from the center point of each cluster using equation (3). The first iteration process can be seen in the Table 2.

| Courses | SIDN | C1 | C2 | C3 | C4 | Cluster Distance |
|---------|------|----|----|----|----|------------------|
|         | 57001| 1  | 2  | 0  | 1  | 3                |
|         | 57002| 1.414 | 1.732 | 1 | 0 | 4                |
|         | 57003| 1.732 | 2 | 1.414 | 1.732 | 3                |
|         | 57004| 0 | 2.236 | 1 | 1.414 | 1                |
|         | 57005| 2.44 | 1.732 | 2.645 | 2.44 | 2                |
|         | 57006| 1 | 2 | 0 | 1 | 3                |
|         | 57007| 1 | 2 | 0 | 1 | 3                |
|         | 57008| 2.44 | 1.732 | 2.236 | 2 | 2                |
|         | 57009| 2 | 1.732 | 1.731 | 2 | 4                |
|         | 57010| 3.162 | 2.236 | 3.316 | 3.162 | 2                |
|         | 57011| 2.236 | 0 | 2 | 1.732 | 2                |
|         | 57012| 1.414 | 2.236 | 1 | 1.414 | 3                |
|         | 57013| 2.236 | 2.44 | 2 | 2.236 | 3                |
|         | 57014| 1.414 | 1.732 | 1 | 0 | 4                |
From the calculation results of the first iteration, it is known that the data with Student ID Number (SIDN) 57001 is closest to cluster 3 with a distance of = 0. Then determine the new centroid for the second iteration. The new centroid is obtained from taking student scores with the cluster of each student. For example, the first score of students is added up by the value of the two students until the last data, then the result is divided by the number of students.

Tabel 3. Second Centroid

| SIDN  | C1   | C2   | C3   | C4   | Cluster Distance |
|-------|------|------|------|------|------------------|
| 57015 | 1.414| 1.732| 1    | 0    | 4                |
| 57016 | 1.414| 1.732| 1    | 0    | 4                |
| 57017 | 1    | 2    | 1.414| 1    | 4                |
| 57018 | 2.236| 2    | 2    | 1    | 4                |
| 57019 | 2.828| 3    | 3    | 2.44 | 4                |
| 57020 | 1.732| 1.414| 1.414| 1    | 4                |
| 57021 | 2.44 | 2.236| 2.645| 2    | 4                |
| 57022 | 2.236| 1.414| 2    | 1.732| 2                |
| 57023 | 2.44 | 2.236| 2.645| 2    | 4                |
| 57024 | 1.732| 1.414| 2    | 1.732| 2                |
| 57025 | 2.268| 1.414| 2    | 1.732| 2                |
| 57026 | 2    | 1    | 1.732| 1.414| 2                |
| 57027 | 2.236| 0    | 2    | 1.732| 2                |
| 57028 | 1.414| 1.732| 1    | 1.414| 3                |
| 57029 | 1.414| 1.732| 1    | 1.414| 3                |
| 57030 | 1.732| 1.414| 1.414| 1.732| 4                |

Source: Research Result (2021)

After getting the new Centroid, the second iteration runs using the same calculation in the first iteration with the centroid. The results of the calculation of the 2nd iteration show a change in cluster distance when compared to the first iteration. There is a change in cluster distance in the data with SIDN 57009,
57022, 57030. Because there are still changes in data, it is followed by the 3rd iteration. Just like before, first determine the centroid then do the iteration calculation process.

| SIDN  | C1  | C2  | C3  | C4    | C    |
|-------|-----|-----|-----|-------|------|
| 57001 | 1   | 2   | 0.663 | 1.49655 | 3    |
| 57002 | 1.414 | 1.732 | 1.2 | 0.71581 | 4    |
| 57003 | 1.732 | 2.121 | 1.2 | 1.87413 | 3    |
| 57004 | 0    | 2.061 | 1.2 | 1.58505 | 1    |
| 57005 | 2.44 | 1.322 | 2.2 | 1.99171 | 2    |
| 57006 | 1    | 2   | 0.663 | 1.49655 | 3    |
| 57007 | 1    | 2   | 0.663 | 1.49655 | 3    |
| 57008 | 2.44 | 1.118 | 1.907 | 1.84972 | 2    |
| 57009 | 2    | 1.5  | 1.280 | 1.99171 | 3    |
| 57010 | 3.162 | 1.658 | 2.905 | 2.82257 | 2    |
| 57011 | 2.236 | 0.866 | 1.685 | 1.61347 | 2    |
| 57012 | 1.414 | 2.061 | 1.019 | 1.69588 | 3    |
| 57013 | 2.236 | 2   | 1.4966 | 2.14553 | 3    |
| 57014 | 1.414 | 1.732 | 1.2 | 0.71581 | 4    |
| 57015 | 1.414 | 1.732 | 1.2 | 0.71581 | 4    |
| 57016 | 1.414 | 1.732 | 1.2 | 0.71581 | 4    |
| 57017 | 1    | 1.802 | 1.56 | 0.88607 | 4    |
| 57018 | 2.236 | 2   | 2.107 | 0.88607 | 4    |
| 57019 | 2.828 | 2.598 | 2.939 | 2.08100 | 4    |
| 57020 | 1.732 | 1.414 | 1.2 | 0.88607 | 4    |
| 57021 | 2.44 | 1.936 | 2.416 | 1.33608 | 4    |
| 57022 | 2.236 | 1.581 | 1.56 | 1.434521 | 4    |
| 57023 | 2.44 | 1.936 | 2.416 | 1.33608 | 4    |
| 57024 | 1.732 | 1   | 1.8  | 1.43452 | 2    |
| 57025 | 2.236 | 1   | 1.74 | 1.66886 | 2    |
| 57026 | 2    | 0.866 | 1.496 | 1.33608 | 2    |
| 57027 | 2.236 | 0.866 | 1.68 | 1.61347 | 2    |
| 57028 | 1.414 | 1.7320 | 0.663 | 1.58505 | 3    |
| 57029 | 1.414 | 1.73 | 0.663 | 1.58505 | 3    |
| 57030 | 1.732 | 1.732 | 1.019 | 1.82498 | 3    |

Source: Research Result (2021)

Because in the 2nd and 3rd iterations the cluster distance did not change, the iteration was stopped and the final results obtained (Table 5).
3.2. Professional Classification

The profession used in this study has been adjusted to the profile of graduates on Information Systems Study Program IBT Pelita Indonesia. The professional classification based on the graduate profile will be shown in the following table.

### Table 5. Results of the Graduate Profile

| SIDN   | C  | Graduate Profile               |
|--------|----|--------------------------------|
| 57001  | 3  | Database Administrator         |
| 57002  | 4  | Web Design And Developer       |
| 57003  | 3  | Database Administrator         |
| 57004  | 1  | Information System Manager     |
| 57005  | 2  | System Analyst                 |
| 57006  | 3  | Database Administrator         |
| 57007  | 3  | Database Administrator         |
| 57008  | 2  | System Analyst                 |
| 57009  | 3  | Database Administrator         |
| 57010  | 2  | System Analyst                 |
| 57011  | 2  | System Analyst                 |
| 57012  | 3  | Database Administrator         |
| 57013  | 3  | Database Administrator         |
| 57014  | 4  | Web Design And Developer       |
| 57015  | 4  | Web Design And Developer       |
| 57016  | 4  | Web Design And Developer       |
| 57017  | 4  | Web Design And Developer       |
| 57018  | 4  | Web Design And Developer       |
| 57019  | 4  | Web Design And Developer       |
| 57020  | 4  | Web Design And Developer       |
| 57021  | 4  | Web Design And Developer       |
| 57022  | 4  | Web Design And Developer       |
| 57023  | 4  | Web Design And Developer       |
| 57024  | 2  | System Analyst                 |
| 57025  | 2  | System Analyst                 |
| 57026  | 2  | System Analyst                 |
| 57027  | 2  | System Analyst                 |
| 57028  | 3  | Database Administrator         |
| 57029  | 3  | Database Administrator         |
| 57030  | 3  | Database Administrator         |

Source: Research Result (2021)

PIKSEL status is accredited by the Directorate General of Research Strengthening and Development No. 28/E/KPT.2019 with Indonesian Scientific Index (SINTA) journal-level of S5, starting from Volume 6 (1) 2018 to Volume 10 (1) 2022.
### Table 6. Professional Classification

| GRADUATES PROFILE | PROFESSION                                    |
|-------------------|-----------------------------------------------|
| INFORMATION       | = IT PLANNER                                  |
| SYSTEM MANAGER    | PROJECT MANAGEMENT                            |
|                   | COMPUTER NETWORK                              |
|                   | ARCHITECTURE                                  |
|                   | TEACHER                                       |
|                   | DIREKTOR OF INFORMATION                       |
|                   | SYSTEM                                        |
|                   | COMPUTER ENGINEERING                          |
|                   | SUPERVISOR INFORMATION SYSTEM                 |
| SYSTEM ANALYST    | = SOFTWARE ENGINEER                           |
|                   | BUILD STARUP                                  |
|                   | ERP                                           |
|                   | TEACHER                                       |
|                   | BUSINESS ANALYSIS                             |
|                   | COMPUTER SYSTEM ANALYSIS                     |
|                   | PROGRAMER ANALYSIS                            |
|                   | HR INFORMATION SYSTEM                         |
| DATABASE ADMINISTRATOR | = ADMIN DATABASE                               |
|                   | DATA ANALYST                                  |
|                   | STAFF ADMINISTRASI                            |
|                   | NETWORK SECURITY                              |
|                   | KOORDINATOR DATABASE                          |
| WEB DESIGN AND DEVELOPER | = UI/UX DEVELOPER                             |
|                   | GAME DEVELOPER                                |
|                   | GRAFPHIC DESIGNER                             |
|                   | WEB DEVELOPER                                 |
|                   | PROGRAMMER                                    |
|                   | APLICATION DEVELOPER                          |
|                   | CREATIVE DIRECTOR                             |
|                   | ILUSTRATOR                                    |
|                   | PROGRAMMER SERVICE                            |
|                   | TEACHER                                       |

Source: Research Result (2021)

Based on the calculation of the graduate profile, the appropriate profession can be determined. The student with SIDN 57001 showed the classification results of database administrator. Another example, student with
SIDN 57002 showed graduate profile classification of web design and developer, hence, the appropriate professions are UI/UX developer, game developer, web developer, programmer, application developer, and teacher. Below will be displayed the results of the suitability of the profession obtained from calculations with their current profession through a tracer study.

Table 7. The Conformity of the Current Profession with the Calculation Results

| SIDN  | C | PROFILE | PROFESSION         | COMPATIBILITY          |
|-------|---|---------|--------------------|------------------------|
| 57001 | 3 | DA      | STAFF IT           | NOT ACCORDANCE         |
| 57002 | 4 | WD      | DIGITAL PUBLISHER  | ACCORDANCE             |
| 57003 | 3 | DA      | TEACHER            | ACCORDANCE             |
| 57004 | 1 | IM      | PROGRAMMER         | NOT ACCORDANCE         |
| 57005 | 2 | SA      | ADMINISTRATION     | NOT ACCORDANCE         |
| 57006 | 3 | DA      | DATABASE ADMINISTRATION | ACCORDANCE         |
| 57007 | 3 | DA      | STAFF IT           | NOT ACCORDANCE         |
| 57008 | 2 | SA      | ADMINISTRATION     | NOT ACCORDANCE         |
| 57009 | 3 | DA      | DATABASE ADMINISTRATION | ACCORDANCE         |
| 57010 | 2 | SA      | ADMINISTRATION     | NOT ACCORDANCE         |
| 57011 | 2 | SA      | ADMINISTRATION     | NOT ACCORDANCE         |
| 57012 | 3 | DA      | ADMINISTRATION     | ACCORDANCE             |
| 57013 | 3 | DA      | ADMINISTRATION     | ACCORDANCE             |
| 57014 | 4 | WD      | ADMINISTRATION     | NOT ACCORDANCE         |
| 57015 | 4 | WD      | TEACHER            | ACCORDANCE             |
| 57016 | 4 | WD      | DATABASE ADMINISTRATION | ACCORDANCE         |
| 57017 | 4 | WD      | TEACHER            | ACCORDANCE             |
| 57018 | 4 | WD      | SALES              | NOT ACCORDANCE         |
| 57019 | 4 | WD      | ADMINISTRATION     | NOT ACCORDANCE         |
| 57020 | 4 | WD      | ADMINISTRATION     | NOT ACCORDANCE         |
| 57021 | 4 | WD      | ADMINISTRATION     | NOT ACCORDANCE         |
| 57022 | 4 | WD      | ADMINISTRATION     | NOT ACCORDANCE         |
| 57023 | 4 | WD      | TEACHER            | ACCORDANCE             |
| 57024 | 2 | SA      | HR INFORMATION SYSTEM | ACCORDANCE         |
| 57025 | 2 | SA      | SALESMAN           | NOT ACCORDANCE         |
| 57026 | 2 | SA      | SALESMAN           | NOT ACCORDANCE         |
| 57027 | 2 | SA      | ADMINISTRATION     | NOT ACCORDANCE         |
Table 7 shows the 43.3% of professional data in accordance with the calculation results.

4. Conclusion

The tests carried out in this study shows 3 times iterations. Based on the results of the closest distance cluster, if it shows the center 1 then it is directed to the information system manager profession, center 2 for system analysis, center 3 for database administrators, and center 4 for web design and developer. After processing the data, the proposed system successfully classifies the graduate profile, e.g. the database administrators with its corresponding profession (database administrator, data analyst, administrative staff, network security and database coordinator), web design developer and its corresponding professions (UI/UX developer, game developer, web developer, programmer, application developer, and teacher), etc. Professional data that has been obtained compared to their current profession using the tracer study method shows as much as 43.3% of professional data is in accordance with the calculation results. The cluster results are also influenced by the value of the initial centroid used and the amount of data used, and the difference in retrieval.

Author Contributions

Susi the topic; Susi and Alyauma Hajjah conceived models and designed the experiments; Susi and Alyauma Hajjah the algorithms; Susi and Alyauma Hajjah analysed the result.

Conflicts of Interest

The author declare no conflict of interest.
Reference
Agusta, Y. (2007). K-Means – Penerapan, Permasalahan dan Metode Terkait. *Jurnal Sistem Dan Informatika*, 3(Februari), 47–60.
Aranda, J., & Natasya, W. A. G. (2016). Penerapan Metode K-Means Cluster Analysis Pada Sistem Pendukung Keputusan Pemilihan Konsentrasi Untuk Mahasiswa International Class STMIK Amikom Yogyakarta. *Semnasteknomedia Online*, 4(1), 4.2-1-4.2-6.
Dzulhaq, M. I., & Imani, R. (2015). Sistem Pendukung Keputusan Pemilihan Konsentrasi Jurusan Menggunakan Fuzzy Inference Sistem Metode Mamdani. *Sisfotek Global*, 5(2), 75–80.
Faisal, & Permana, S. D. H. (2015). Sistem Penunjang Keputusan Pemilihan Sekolah Menengah Kejuruan Teknik Komputer dan Jaringan yang Terfavorit Dengan Menggunakan Multi-Criteria Decision Making. *Jurnal Teknologi Informasi Dan Ilmu Komputer*, 2(1), 11–19. https://doi.org/10.25126/jtiik.201521123
Fauzi, A., Marpaung, I. J. S., & Parde, A. M. H. (2018). Sistem Pendukung Pemilihan Pekerjaan Menggunakan Metode Apriori Berdasarkan Korelasi Jurusan Dengan IPK Untuk Mengetahui Pekerjaan yang Tepat. *Methodika: Jurnal Manajemen Informatika & Komputerisasi Akuntansi*, 2(2), 152–159.
Hastuti, N. F. (2013). *Pemanaatan Metode K-Means Clustering Dalam Penentuan Penerima Beasiswa*. Universitas Negeri Surakarta.
Kusumantara, P. M., Wibowo, N. C., & Novia, Y. (2019). Sistem Pendukung Keputusan Penentuan Profesi Linier Untuk Lulusan Program Studi Sistem Informasi Dengan Metode SAW. *Prosiding Nasional Rekayasa Teknologi Industri dan Informasi XIV Tahun 2019 (ReTII)*, 285–290.
Muningsih, E., & Kiswati, S. (2015). Penerapan Metode K-Means Untuk Clustering Produk Online Shop Dalam Penentuan Stok Barang. *Jurnal Bianglala Informatika*, 3(1), 10–17.
Murdani. (2013). Sistem Pendukung Keputusan Pemilihan Pekerjaan Bagi Graduates Terknik Informatika Dengan Metode Simple Additive Weighting (SAW). *Pelita Informatika*, 4(2), 40-44.
Sibuea, M. L., & Safta, A. (2017). Pemetaan Siswa Berprestasi Menggunakan Metode K-Means Clustring. *Jurteksi, 4*(1), 85–92. https://doi.org/10.33330/jurteksi.v4i1.28

Syafirizal. (2014). Sistem Pendukung Keputusan Menentukan Bidang Pekerjaan Bagi Lulusan LPP Penerbangan Menggunakan Metode ANP. *Seminar Nasional Informatika 2014*, 291–295.

Turban, E., Aronson, J. E., & Liang, T.-P. (2005). *Decision Support Systems and Intelligent Systems.* Andi Offset.