Articles

Data mining applied to the segmentation of students and the acquisition of new profiles by means of clustering techniques.

Minería de datos aplicada a la segmentación de estudiantes y captación de nuevos perfiles mediante técnicas de clustering

Carlos Rafael Guffantte Salazar [*]
Santiago Israel Logroño Naranjo [*]

REVISTA TECNOLÓGICA
ciencia y educación
Edwards Deming

ISSN: 2600-5867

Edited by: Tecnológico Superior Corporativo Edwards Deming

July – December Vol. 9 - 2 – 2025

https://revista-edwardsdeming.com/index.php/es

e-ISSN: 2576-0971

Received: March 28, 2025

Approved: April 13, 2025

Page 33-47

SUMMARY

This project proposes the application of data mining tools, in particular clustering techniques, to identify the most representative profiles that have passed through the CONDUESPOCH driving school located in the city of Riobamba.

The research is justified by the need to improve the recruitment of students before each new academic period, and through the processing of historical information available to the institution it is possible to segment these individuals into different groups. The development is based on the application of K-Means and DBSCAN models, each with its strengths and weaknesses depending on the nature and distribution of the data available; the DBSCAN model is the most appropriate for the case study, reaching validation metrics of 0.78 for the silhouette coefficient and 0.27 for the Davies-Bouldin index.

The study adopts a quantitative and exploratory approach, with an applied methodology. The main characteristics of the most significant groups will be analyzed in order to propose recruitment strategies that can be effective for the institution.

The program seeks to improve the integration of data mining in institutional management, simplifying student recruitment processes so that the school maintains a high profile in the region with its peers; in addition to ensuring efficient and informed decision making based on the knowledge acquired.

Keywords: Data Mining, Clustering, Profile Segmentation, Capture Strategies, Data Science.

Resumen

This project proposes the application of data mining tools, in particular clustering techniques, to identify the most representative profiles that have passed through the CONDUESPOCH driving school located in the city of Riobamba.

The research is justified by the need to improve the recruitment of students before each new academic period, and through the processing of historical information available to the institution it is possible to segment these individuals into different groups. The development is based on the application of K-Means and DBSCAN models, each with its strengths and weaknesses depending on the nature and distribution of the data available; the DBSCAN model is the most appropriate for the case study, reaching validation metrics of 0.78 for the silhouette coefficient and 0.27 for the Davies-Bouldin index.

The study adopts a quantitative and exploratory approach, with an applied methodology. The main characteristics of the most significant groups will be analyzed in order to propose recruitment strategies that can be effective for the institution.

The program seeks to improve the integration of data mining in institutional management, simplifying student recruitment processes so that the school maintains a high profile in the region with its peers; in addition to ensuring efficient and informed decision making based on the knowledge acquired.

Key words: Data Mining, Clustering, Profile Segmentation, Capture Strategies, Data Science.

INTRODUCTION

In the current educational landscape, institutions face the constant challenge of better understanding the characteristics and needs of their students. This understanding is not only key to design more effective educational programs, but also to attract new profiles that fit their objectives (Romero & Ventura, 2020). In the case of the CONDUESPOCH driving school, located in Riobamba, there is a valuable history of academic and demographic data for approximately 1000 students. However, the lack of in-depth analysis of this information limits its ability to identify useful patterns that could transform its educational planning and recruitment strategies.

A very useful tool that helps to overcome this type of challenge is related to the application of data mining techniques, specifically clustering techniques, which allow grouping individuals belonging to the institution into segments with similar characteristics, revealing patterns that under ordinary circumstances might go unnoticed. For this purpose, there are algorithms such as K-means and DBSCAN, which facilitate the management of large volumes of data, in addition to providing practical results for decision making (Liao et al., 2012). By applying these tools in CONDUESPOCH, it will be possible to effectively explore all the available data, and from this to perform the segmentation of students to finally design personalized strategies based on the groups formed.

The impact of this study focuses on turning CONDUESPOCH's historical data into a strategic advantage. By identifying clear student segments, the institution will be able to personalize its educational offerings, optimize resources and focus its recruitment efforts on the most representative and successful profiles. The main objective of this work is to use clustering techniques to segment students according to their demographic and academic characteristics, thus contributing to improve both institutional planning and the attraction of new profiles. To achieve this goal, three specific objectives are proposed: to identify student profiles from key data, to evaluate the performance of the K-means and DBSCAN algorithms in this context, and to design recruitment strategies aimed at the most prominent segments.

The purpose of this research, beyond the benefits that the CONDUESPOCH school will obtain, is that it can become a model that other educational institutions can replicate; offering an innovative alternative to face the challenges of the educational sector based on an approach based on data analysis and the subsequent segmentation and understanding of the students who belong to them. In addition, it is intended to answer the following question: How can clustering techniques improve the segmentation of students in CONDUESPOCH and optimize its recruitment process?

Theoretical Framework

The use of *data analytics* or data *analysis* in education has become a transcendental procedure to aspire to improve both teaching and the recruitment of new students in a given institution, not to mention the support it represents for strategic decision making. However, applying these techniques is usually not simple; on the contrary, it is certainly common to face obstacles related to the lack of information derived from inconsistencies or lack of registration in the institutions, in addition to the fact that some of them may not even have systems that allow them to manage the data efficiently, affecting the quality of the analyses. In addition, there is often a lack of technical training to use these tools properly. There is also the issue of the use of personal data, which requires careful consideration of ethical and privacy aspects (Castro R. et al., 2018).

Despite the complications associated with these challenges, the potential of data mining to change the way educational institutions are managed is enormous, establishing itself as a highly relevant resource for improving the student experience and making better use of the resources available to the institution. To mention an example, Dutt et al (2015) highlight the usefulness of algorithms such as K-means and hierarchical clustering to identify patterns in the data of students belonging to certain institutions, facilitating the design of effective and personalized strategies for each case. Similarly, Shrestha & Pokharel (2020) applied data mining techniques in platforms such as Moodle, through which they were able to identify groups of students with common characteristics; in addition to the areas where greater intervention was required due to the presence of students at risk of low performance.

In the case of new student recruitment, Estrada-Danell et al. (2016) developed a model that applies data mining with the intention of optimizing the enrollment process in higher education institutions. This study is responsible for identifying, through techniques such as decision trees, the individuals with the highest probability of academic success, in such a way that allows the institution to direct its efforts towards the most promising profiles, improving its educational management. These studies show that, apart from organizing the information to be managed much more efficiently, data analysis is capable of generating a significant impact on the personalization of education and, above all, on the improvement of institutional processes.

MATERIALS AND METHODS

The present research was developed in an exploratory and applied manner, since it is intended to analyze patterns in the data by means of grouping techniques, without a predefined structure and with a quantitative approach.

Research design and type of research

This study uses a quantitative approach to analyze the data, which will be numerical in nature, of students who have historically attended the CONDUESPOCH driving school, in order to detect the possible existence of significant patterns among them. This approach is imperative due to the need to resort to statistical techniques that facilitate obtaining concrete and verifiable results. This starts with the collection of historical data available in the institutional systems, followed by a detailed analysis to identify the key relationships between variables (Perreault, 2011).

The design is applied, as it seeks to use existing data to recruit new students, identifying relevant profiles. In addition, it is descriptive, focusing on categorizing the main characteristics of current students to extrapolate trends in the recruitment of new profiles. This analysis is performed at a single moment in time and without affecting the weight of the information represented by any variable is simply limited to the observation and analysis of the available data, classifying the study as non-experimental and cross-sectional (Baker & Inventado, 2014).

Population and sample

The population is composed of approximately 1000 student records of the CONDUESPOCH driving school, in which information academic and demographic information is available for each one of them. Taking into consideration that the totality of the available data will be analyzed, it will not be necessary to apply a sampling under any criteria, ensuring a general representativeness of the results.

Research methods and techniques

The following methods and techniques were used for data collection and analysis:

Collection of information

Data were extracted from multiple tables stored in the institution's central database. Variables considered include:

- **Demographics**: Age, gender and place of origin (province, canton, parish, etc.).
- **Academics**: Qualifications, Status.

Python software was used to develop the models, which has specialized libraries for data analysis and management such as Scikit-learn, Pandas and Matplotlib, among others, facilitating both the pre-processing of the data and the visualization of the results obtained.

Data analysis procedures

The data collected from the different tables stored in the institutional system were

grouped and integrated into a single table using Python. The procedure began with the preprocessing of the data, where missing and duplicate values were eliminated as part of the cleaning process, avoiding working with irregular information. Subsequently, the most relevant variables were identified using correlation and exploratory analysis techniques. Irregular distributions were identified in most of the data, so it was necessary to enrich the characteristics, obtaining new variables from the original ones, such as performance from grades; in addition, a column for age was generated from the dates of birth and beginning of the courses. The numerical variables were normalized using MinMaxScaler to ensure a uniform scale, which improves the performance of the algorithms; while for the categorical variables, one-hot coding was performed to ensure compatibility with the corresponding clustering techniques (Han et al., 2021).

Once the data were processed, two clustering algorithms were implemented:

- **K-Means**: The optimal number of clusters was determined using the elbow method.
- **DBSCAN**: An adjustment was made to the epsilon and min_samples parameters in such a way that significant clusters could be identified, in addition to detecting the existence of possible outliers. .

The quality of the clusters obtained for each case was evaluated using metrics such as the silhouette coefficient to estimate the cohesion and separation between groups, and the Davies-Bouldin index to assess their level of compactness and dispersion (Xu & Wunsch, 2018).

RESULTS

This section presents the findings obtained from the clustering of the register of students who have studied at CONDUESPOCH during the last five periods, in order to identify the most popular profiles and adapt strategies to improve future recruitment at the institution.

Evaluation Metrics

As shown inError! Reference source not found. , the DBSCAN model is the most appropriate to work with because it has better metrics than its similar K-Means model. This can be explained due to the quite particular distribution of the data, complicating the homogeneous grouping performed by K-Means and favoring the density analysis of DBSCAN (Bhuyan & Borah, 2023).

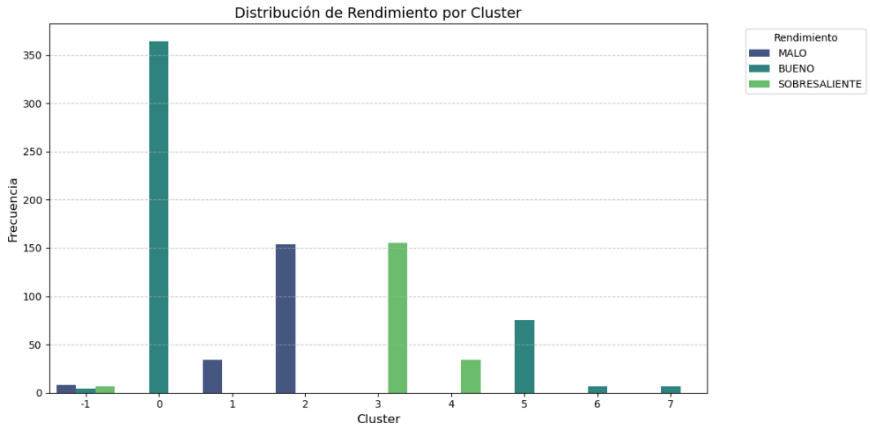
Table 1: Evaluation metrics for the applied models

MODEL	# CLUSTERS	COEF. SILUETA	D-B INDEX
DBSCAN	8	0,78	0,27
KMEANS	7	0,64	0,7

Distribution of Clusters by Performance

InError! Reference source not found. , the results of the clustering are presented. The most significant clusters were cluster 0, with more than 350 individuals, clusters 2 and 3 with more than 150 individuals and cluster 5 with 75 individuals; on the other hand, cluster 4 had only 34 individuals grouped together but with a high recorded performance. The large and well-performing clusters were used as a starting point for analyses leading to the development of recruitment strategies (Luan, 2002).

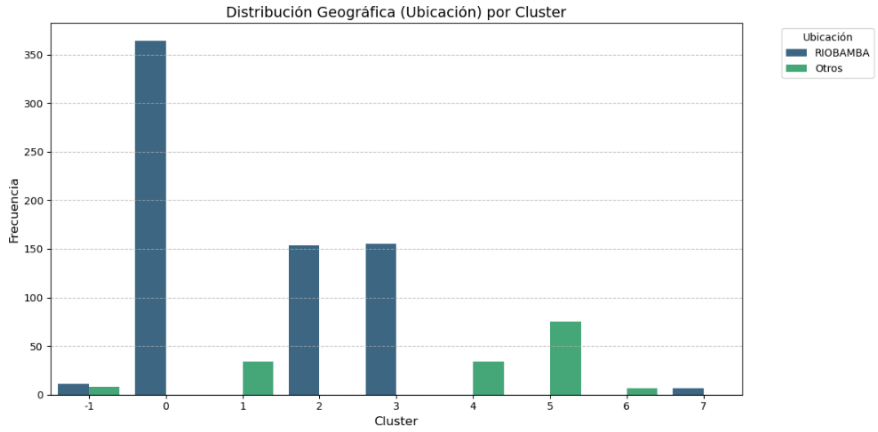
Figure 1: Performance by cluster



Distribution by location

According toError! Reference source not found. , most of the students are from the city of Riobamba, as indicated by the three largest clusters, while clusters 4 and 5, which were also of interest for the academic part, are distributed among other cantons of the country. This information allows for a more in-depth analysis of the regions and possible educational establishments in which the institution can be promoted.

Figure 2: Customer segmentation: Location by Cluster

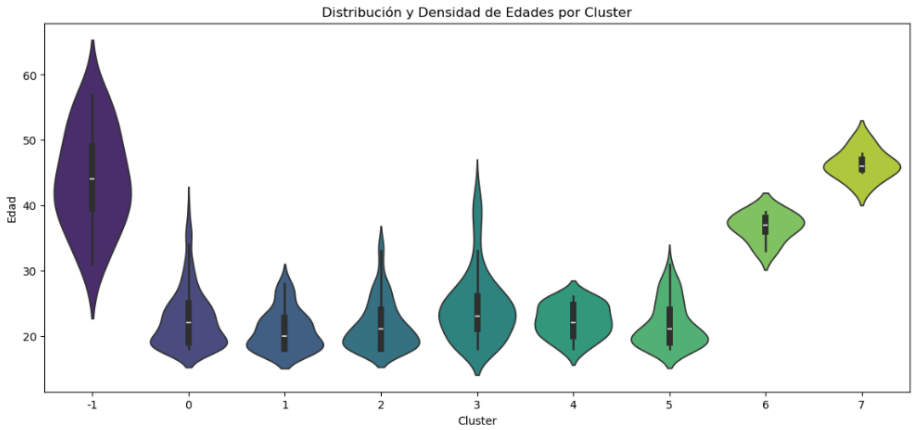


Within the groups focused on the analysis in Riobamba, there is a clear tendency of individuals who live in the urban parishes of the city, especially in Lizarzaburu; in addition, they are generally registered as members of public educational institutions. In the groups of interest outside of Riobamba, the majority of people indicate Guano as their city of residence or study at ESPOCH. It is worth noting the low presence of people from parishes close to the institution such as Licán or Yaruquíes, as well as from neighboring cantons such as Chambo on the other hand.

Age Distribution

As illustrated **Error! Reference source not found.** , clusters 6 and 7 are the most differentiated in terms of student age with respect to the others, with mean values of 36 and 46 years, respectively. The other clusters have a similar behavior, with a density concentrated between 20 and 25 years of age.

Figure 3: Ages by Cluster



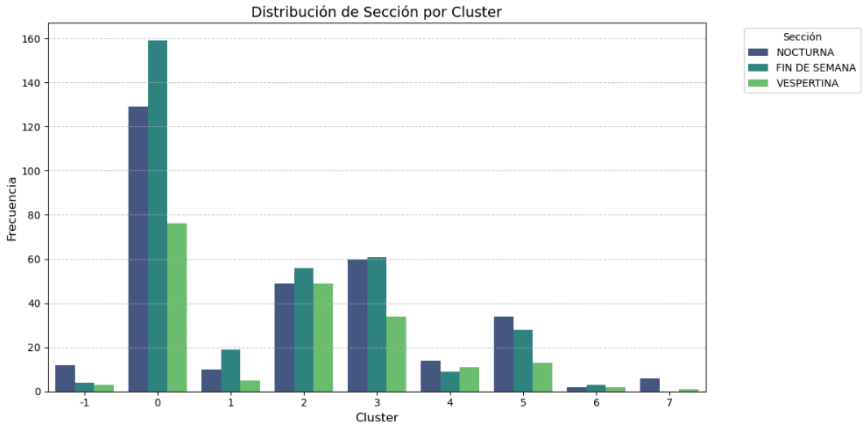
Age distribution for each group formed

This suggests that the vast majority of students who have historically been part of the institution did so tentatively while they were finishing their higher education studies or shortly after completing high school, and few have chosen to enter the institution after the age of 35.

Additional Observations

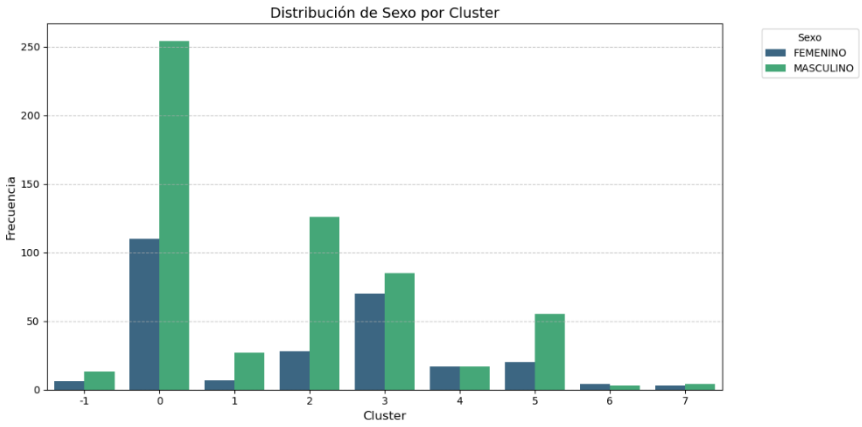
In**Error! Reference source not found.** , the distribution of people according to the section in which they took their courses is presented. There are no clear preferences between groups in this section beyond the apparent preference for the evening and weekend option over the afternoon, even so, this section has a fairly decent volume of students that keeps it as a good option to continue, especially for people based in Riobamba.

Figure 4: Sections per Cluster



In addition, the gender distribution for the clusters found is summarized in**Error! Reference source not found.** . Of the representative clusters, only in the outstanding ones is there an even distribution between male and female, while for the rest there is a clear majority tendency for the former.

Figure 5: Gender by Cluster



Proposed Strategies

Once the relevant information was known, and taking as a starting point what is mentioned by Kotler et al. (2012), which supports the use of segmentation and data for the development of recruitment strategies, some proposals were made, as described below:

- Focus campaigns in the urban sector of Riobamba, especially in the Lizarzaburu parish, which corresponds to the school's parish.
- Conduct talks in the educational institutions where most of the students registered in the interest groups are concentrated in order to maintain or improve the current affluence.
- Offer exclusive economic or academic benefits to students who are part of ESPOCH, considering the complications that may arise when dealing with third level students who usually come from other regions of the country. This can be taken advantage of given the existing link between the institutions, allowing to conveniently increase the number of potential applicants to join, emphasizing the evening or night shift depending on their time availability.
- To raise the possibility of reaching the students of the National University of Chimborazo, since there is very little presence of this segment within the interest groups, and they would contribute in an important way as people within the most popular age group.
- Suggest promoting campaigns in parishes such as Yaruquíes and Licán because they are relatively close to the school but have a notably lower proportion of people than other regions reviewed.
- Promote the institution's offerings in cantons near Riobamba such as Chambo or Penipe, given the success of Guano, which has had a high density of students.

- It is suggested to highlight the weekend offer for the people of the parishes and surrounding cantons previously mentioned, considering the frequent transportation issues required for the other sections.
- Given that the age segment is mostly between 20 and 26 years old, an active approach in social networks is recommended, through the generation of brief and dynamic content on platforms such as TikTok or Instagram, which are the most popular within this social group; complementing other alternatives such as Facebook and YouTube where the content can be oriented more to professional development and other opportunities. The possibility of directing the advertising campaign on these platforms through geolocalized ads, where the work would focus on the spaces already stipulated, would abundantly and efficiently facilitate the objective of getting in contact with potential new faces that can become part of the institution.

Error! Reference source not found. summarizes the recruitment strategies proposed based on the clustering results obtained.

Table 2: Summary of Strategies based on cluster distributions

Segmentation	Description	Remarks	Proposals
Location	Riobamba	High concentration of people in Lizarzaburu parish, followed by the other urban parishes.	Focus the campaign in the urban area of the city, especially in Lizarzaburu. Suggest promoting the offer in Licán or Yaruquíes, which are areas adjacent to the school.
		High tendency of people in fiscal institutions such as: Carlos Cisneros, Riobamba, Maldonado, Juan de Velasco.	Conduct lectures in these institutions to maintain or improve the influx of people.
		Within the higher education institutions, there is a majority of people belonging to ESPOCH, with almost no presence of people	Offer exclusive benefits to students who belong to ESPOCH. Promote the offer in other higher education institutions in the city.

		from other universities.	
	Others	The vast majority is concentrated in Guano, with no influence from the other surrounding cantons.	Consider this trend for future campaigns to make a presence in Guano. Evaluate the possibility of visiting nearby cantons such as Chambo to publicize the offer and try to expand the volume of people from these areas.
Age	20-27 years	Great majority of people under 27 years old, high concentration around 22 years old.	Have an active social media approach to reach this age group effectively.

DISCUSSION

The results reflect that data mining techniques, in the particular case of clustering algorithms, can serve as a tool of notable relevance in the educational field to identify segments of students with different characteristics, as pointed out by Baker & Inventado (2014), where these methods are considered among the keys to develop and move to a new level the management of institutions, in addition to assessing their applicability for other purposes such as the analysis of student behavior.

With regard to the selected models, we agree with Lazaro-Camasca & Nuñez-Medrano (2023), who applied, among others, K-Means and DBSCAN models to identify performance patterns in a group of students; obtaining varied results between both and validating them based on metrics such as the silhouette coefficient, in addition to the consideration that for this particular purpose it would be useful to have a larger number of groups to diversify the classes and focus attention on certain groups.

Particularly, the proposed strategies are aligned with the stipulations of Parra Armendariz et al. (2022), who in their work propose approaches based on geographic segmentation (to identify the locations with the greatest potential, as well as to focus recruitment campaigns), data mining and marketing models that suggest defining

strategic audiences and inter-institutional alliances, generating attractive digital content for the groups in question and anticipating the needs of potential students.

In the case of CONDUESPOCH, applying clustering techniques to segment students represents a unique opportunity to improve internal planning and develop more effective recruitment strategies. This approach would not only strengthen the institution's competitiveness, but can also serve as a replicable model for other educational organizations interested in leveraging their data to make informed decisions (Khan & Ghosh, 2021).

This study has allowed us to learn more about the main characteristics of the students who have been part of the CONDUESPOCH driving school, which was useful in order to plan recruitment strategies with targeted offers based on the sector with which the groups were identified.

The distribution and nature of the school data had a notable influence at the time of interpreting the results of both models applied, clearly evidencing in the evaluation metrics that the DBSCAN model formed a more solid grouping than the K-Means model, making it the ideal model to propose recruitment proposals with a high level of confidence.

The recruitment strategies were designed taking into account the most successful academic profiles within the total population, and for the most part, in accordance with the most numerous groups identified; it is important to adapt the way in which people will be reached based on factors such as their age and distance from the institution.

The study has achieved its main objective, which was to apply clustering techniques to segment the school's students into groups, based on their demographic and academic characteristics, allowing the development of proposals aimed at benefiting the recruitment of students in an informed and strategically planned manner.

REFERENCES

Baker, R. S., & Inventado, P. S. (2014). Educational Data Mining and Learning Analytics. In J. A. Larusson & B. White (Eds.), *Learning Analytics: From Research to Practice* (pp. 61-75). Springer New York. https://doi.org/10.1007/978-1-4614-3305-7_4

Bhuyan, R., & Borah, S. (2023). A Survey of Some Density Based Clustering Techniques. *National Conference on Advancements in Information, Computer & Communication*, 1. https://doi.org/10.13140/2.1.4554.6887

Castro R., L. F., Espitia P., E., & Montilla, A. F. (2018). Applying CRISP-DM in a KDD Process for the Analysis of Student Attrition. In J. E. Serrano C. & J. C. Martínez-Santos (Eds.), *Advances in Computing* (pp. 386-401). Springer International Publishing.

Dutt, A., Aghabozrgi, S., Ismail, M. A., & Mahroeian, H. (2015). Clustering Algorithms Applied in Educational Data Mining. *International Journal of Information and Electronics Engineering (IJIEE)*, 5. https://doi.org/10.7763/IJIEE.2015.V5.513. https://doi.org/10.7763/IJIEE.2015.V5.513

Estrada-Danell, R. I., Zamarripa-Franco, R. A., Zúñiga-Garay, P. G., & Martínez-Trejo, I. (2016). Contributions from data mining to the enrollment capture process in private higher education institutions. *Educare Electronic Journal*, *20*(3), 1-21. https://www.redalyc.org/articulo.oa?id=194146862011. https://www.redalyc.org/articulo.oa?id=194146862011

Han, J., Kamber, M., & Pei, J. (2021). *Data Mining. Concepts and Techniques, 3rd Edition (The Morgan Kaufmann Series in Data Management Systems)*.

Khan, A., & Ghosh, S. K. (2021). Student performance analysis and prediction in classroom learning: A review of educational data mining studies. *Education and Information Technologies*, *26*(1), 205-240. https://doi.org/10.1007/s10639-020-10230-3. https://doi.org/10.1007/s10639-020-10230-3

Kotler, P., Keller, K. L., Edition, D., Maria, T., Mues, A., Monica, Z., Gay, M., De La, M., Eloisa, L., Rivera, A., Hernandez, M., Enrique, E., & Bianchi, C. (2012). *Marketing management ADAPTATION AND TECHNICAL REVIEW* (G. Domínguez, Ed.; 14th ed.). Pearson.

Lazaro-Camasca, E. N., & Nuñez-Medrano, Y. (2023). Segmentation of University Students using Clustering and considering a Virtual Cycle. *Proceedings of the LACCEI International Multi-Conference for Engineering, Education and Technology*, 2023-July. https://doi.org/10.18687/laccei2023.1.1.1355

Liao, S. H., Chu, P. H., & Hsiao, P. Y. (2012). Data mining techniques and applications - A decade review from 2000 to 2011. In *Expert Systems with Applications* (Vol. 39, Issue 12, pp. 11303-11311). https://doi.org/10.1016/j.eswa.2012.02.063

Luan, J. (2002). Data Mining and Its Applications in Higher Education. *New Directions for Institutional Research*, *2002*, 17-36. https://doi.org/10.1002/ir.35

Parra Armendariz, C., Ulloa Viteri, S., & Medina, P. (2022). Systematic literature review on educational marketing. *Religation. Journal of Social Sciences and Humanities*, *7*(33). https://doi.org/10.46652/rgn.v7i33.943. https://doi.org/10.46652/rgn.v7i33.943.

Perreault, K. (2011). Research Design: Qualitative, Quantitative, and Mixed Methods Approaches. *Manual Therapy*, *16*(1), 103. https://doi.org/https://doi.org/10.1016/j.math.2010.09.003

Romero, C., & Ventura, S. (2020). Educational data mining and learning analytics: An updated survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *10*(3). https://doi.org/10.1002/widm.1355. https://doi.org/10.1002/widm.1355.

Shrestha, S., & Pokharel, M. (2020). Data Mining Applications Used in Education Sector. *Journal of Education and Research*, *10*, 27-51. https://doi.org/10.3126/jer.v10i2.32721. https://doi.org/10.3126/jer.v10i2.32721

Xu, R., & Wunsch, D. (2018). Cluster Analysis. In *Clustering* (pp. 1-13). John Wiley & Sons. https://doi.org/https://doi.org/10.1002/9780470382776.ch1