World Journal of Information Systems 3(3) 1-12
DOI: https://doi.org/10.17013/wjis.v3i3.59

Article

Clustering Socioeconomic and Spatial Patterns of School Dropout

Vanessa Maribel Choque Soto [1*], Victor Dario Sosa Jauregui [1*]
[1] Department of Informatics, Universidad Nacional de San Antonio Abad del Cusco, Cusco, Perú
* Correspondence: vanessa.choque@unsaac.edu.pe, victor.sosa@unsaac.edu.pe

(Article history: Received June 20, 2026; Received in revised form August 01, 2026; Accepted August 19, 2026; Available online September 15, 2026)

Abstract: Academic dropout is a fundamental problem at all educational levels. Several studies have been conducted to identify the characteristics that motivate students to drop out. Few studies have found patterns related to location. The objective of this article is to show the geographic patterns of academic dropout of undergraduate university students and to analyze them to improve decision-making in Peruvian universities. The methodology for this purpose is KDD (Knowledge Data Discovery), commonly known as data mining, which includes the Clustering technique and the application of the K-means algorithm, one of the most popular and efficient algorithms. The results reveal notable geographical patterns, including the university site and the district to which dropout students belong. The results help to identify the university campuses with the highest dropout rates and the districts from which the dropouts come. The latter is analyzed from a socioeconomic perspective.

Keywords: Clustering, Data mining, Educational data mining, Geographic dropout patterns, Student dropout.

1. Introduction

Academic dropout affects universities in terms of student capacity and economically. This problem has been studied from different perspectives, identifying sociodemographic and academic factors, institution-specific variables, student trajectory, among others. [1]

To achieve new insights, it is important to involve different tools. Data mining helps to identify unknown information in large volumes of data over a period of time.[2] It then allows for processing information not limited to a specific population, but to data sets coming from large periods of execution of the organizations.

In the article "Classification models for determining types of academic risk and predicting dropout in university students", he suggests expanding the results of the research by considering new variables, such as institutional and socioeconomic variables, and working with a much larger data set.[3] This article presents the use of clustering techniques to find

© 2026 ITMA. All rights reserved.

patterns of geographic location, which makes it possible to expand variables around the same problem.

On the other hand, [4] in their research, Analysis of University Student Attrition using Data Mining Techniques, concludes that academic results, such as socioeconomic status, influence a student's decision to remain in their respective career, and they also state that by managing these variables it is possible to reduce attrition rates in the university system. The article makes a socioeconomic analysis with information from the 2018 poverty map around geographic patterns.

Clustering is the process of grouping data into clusters so that objects within a cluster have high similarity to each other but are very different from objects in other clusters. The differences are evaluated based on the values of the attributes that describe the objects.[5]

The objective of the article is to show the geographic distribution of academic dropout at the district level using the Clustering technique to guide decision-making in universities. [6].

2. Methodology

2.1 Research and information techniques

The research covered undergraduate programs in the city of Cusco. The universe of the research corresponds to the professional programs of Systems Engineering with face-to-face modality.
The research techniques used were:

- Interviews with managers, academic directors, principal, students, and teachers of the case study and other local universities.
- Observation of situations, phenomena, and behaviors.
- Documentary review, curriculum, and transparency documents published on university websites.

The academic data set was requested in .csv format. It was a dataset from 10 years of activity.

2.2 Methodology process

The methodology used was knowledge discovery from data KDD, consisting of the stages shown in Figure 1.

© 2026 ITMA. All rights reserved.

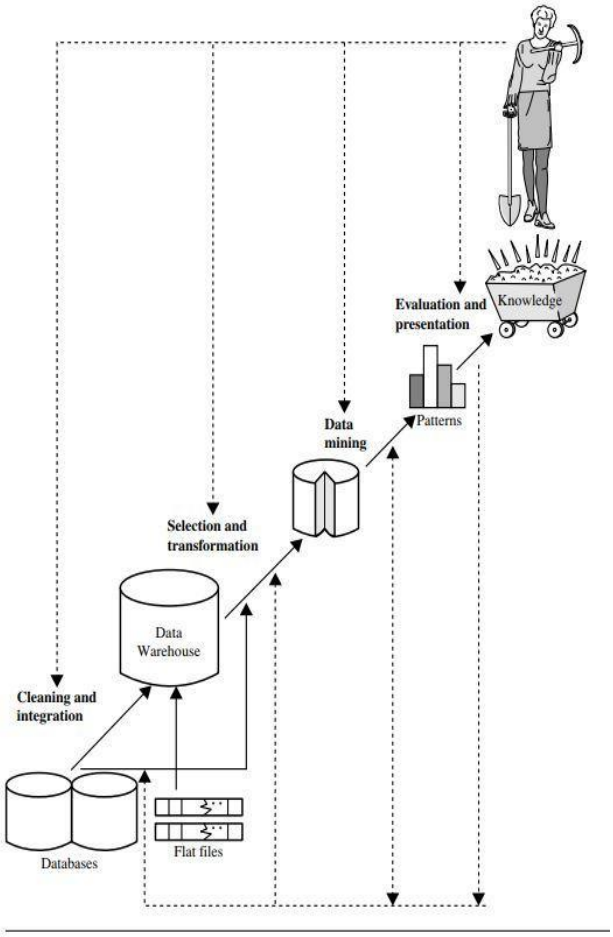


Fig. 1. KDD Process. Source Han Pei [5]

In the figure, based on KDD, the flow for the dropout model is proposed.

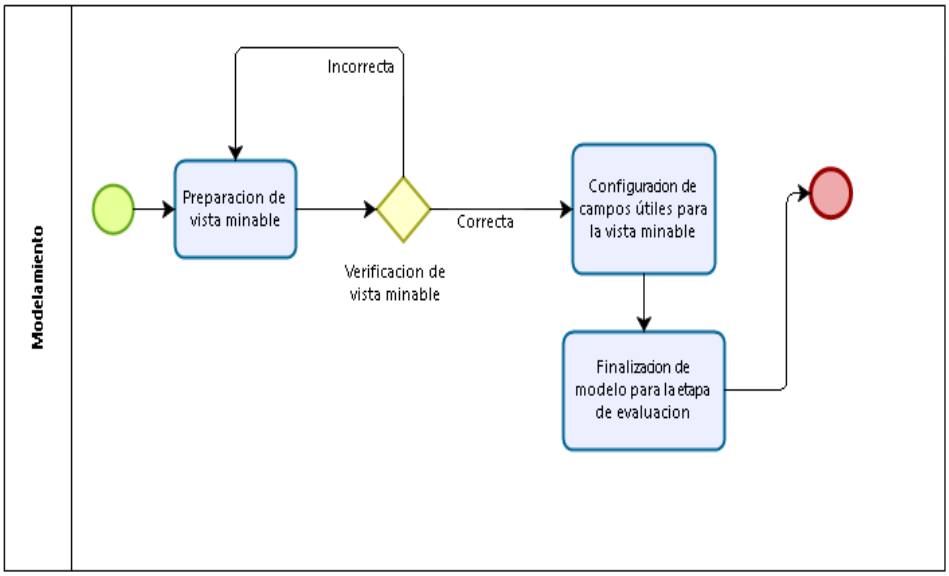


Fig. 2. Dropout model Flow

© 2026 ITMA. All rights reserved.

The clustering data mining technique was used. Within the Clustering technique K-means algorithm was used due to its popularity, simplicity, and lower computational complexity [7] For information processing, Open-Source Data Mining software packages were used (Open-Source Organization, 2020) for the extraction, transformation, and loading process that makes up the data preparation phase. [8]

The data mining software was Weka (Waikato Environment for Knowledge Analysis) [9]

3. Results

The study was conducted on information from 10 years of academic exercise of the Systems Engineering program. From which a model of attrition was constructed, resulting in 1451 records, based on which several academic, social, gender, and other factors emerged. This article focuses on the presentation of geographical patterns.

Figure 3 is the pre-processing view, which is loaded with the previously elaborated attrition model and shows the fields that the tool has identified, which can be selected according to the type of analysis required. Two geographic criteria have been selected: district and site name.

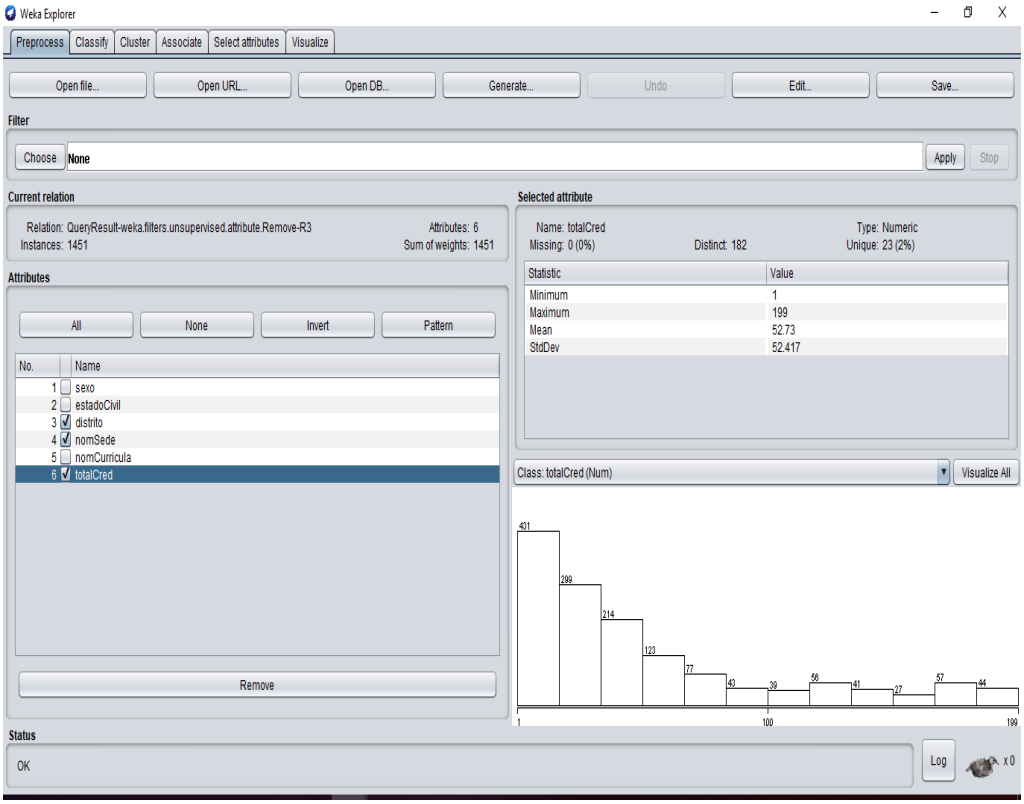


Fig. 3. Pre-processed database with geographic criteria selected

Figure 4 shows the bar chart of the site, which represents the place to which the program belongs. It shows two groups, one with 1317 data corresponding to the city of Cusco and the colors to its districts, and the other belonging to Puerto Maldonado with 134 data.

© 2026 ITMA. All rights reserved.

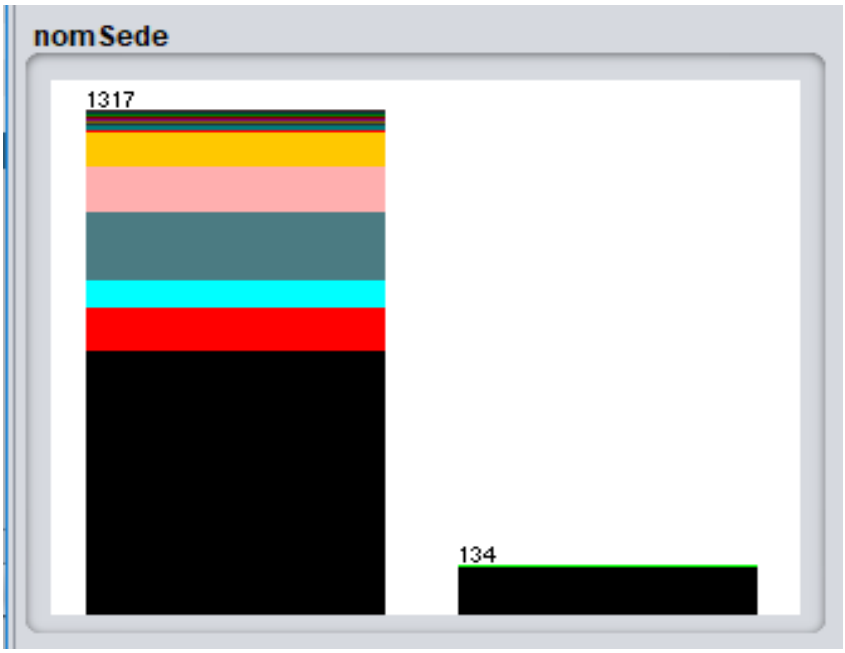


Fig. 4. Bar chart by site.

Figure 5 shows the bar chart of the district to which the students belong according to the attrition model, in each bar the number of students per district is displayed, which will be expanded in Table 1.

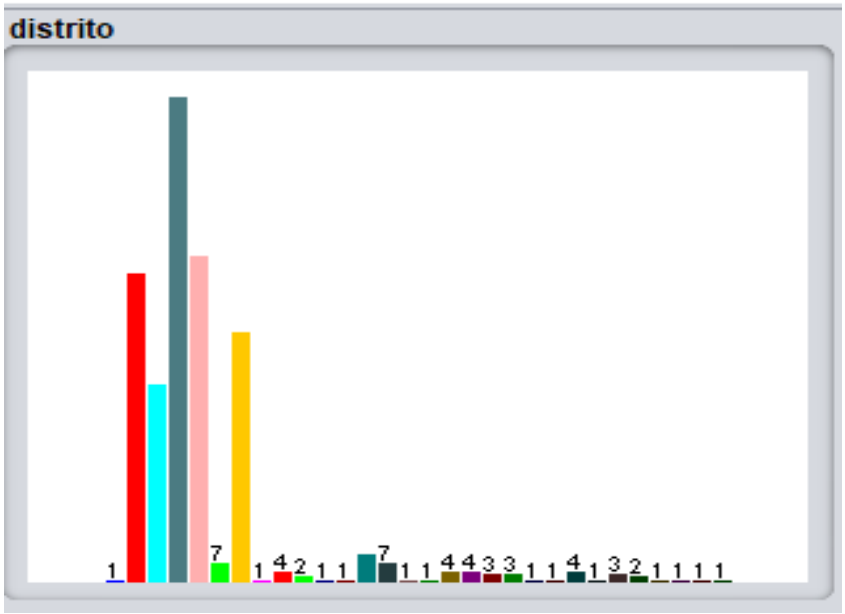


Fig. 5. Bar chart by district.

Figure 6 shows the results of the processing of the dropout model, applying K-means. It generates a cluster with 2 percentages for each site, which in this case confirms the data 1317 for Cusco and 134 for Puerto Maldonado.

© 2026 ITMA. All rights reserved.

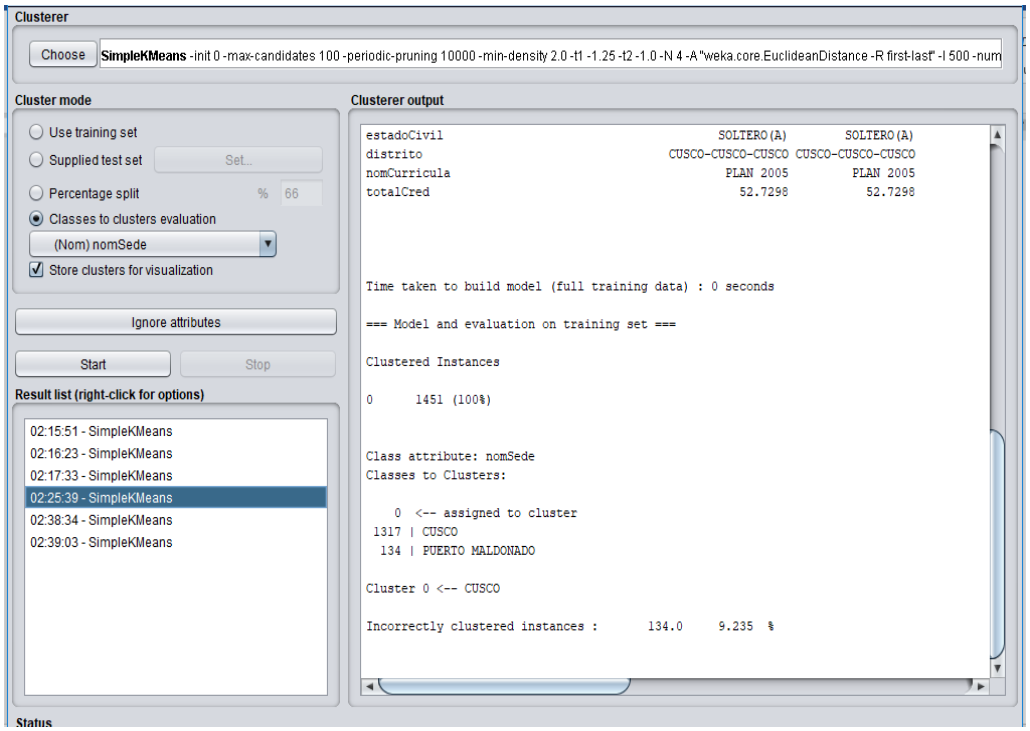


Fig. 6. Results of the K-means algorithm

Figure 7 shows the district field with a single cluster, showing the number of dropouts per district from which they come, information that will be expanded in Table 1.

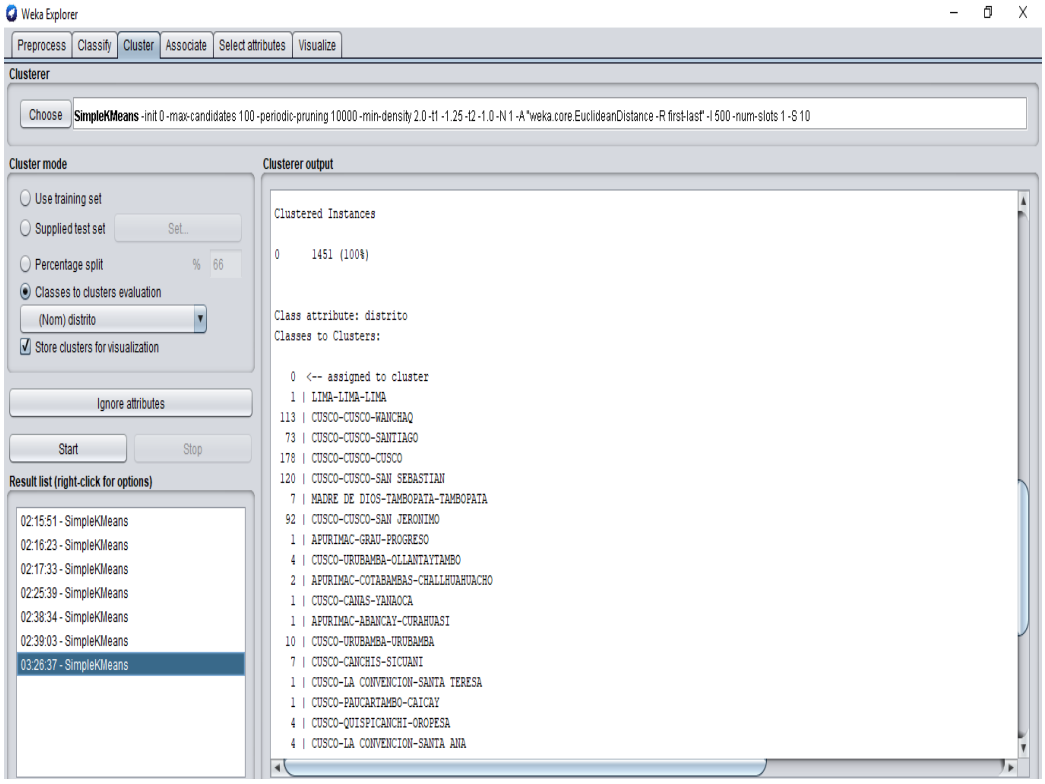


Fig. 7. District with one cluster

© 2026 ITMA. All rights reserved.

Table 1 shows detailed results by district, and Table 2 by site.

Table 1. Results by district

District	Dropouts
LIMA-LIMA-LIMA	1
CUSCO-CUSCO-WANCHAQ	113
CUSCO-CUSCO-SANTIAGO	73
CUSCO-CUSCO-CUSCO	178
CUSCO-CUSCO-SAN SEBASTIAN	120
MADRE DE DIOS-TAMBOPATA-TAMBOPATA	7
CUSCO-CUSCO-SAN JERONIMO	92
APURIMAC-GRAU-PROGRESO	1
CUSCO-URUBAMBA-OLLANTAYTAMBO	4
APURIMAC-COTABAMBAS-CHALLHUAHUACHO	2
CUSCO-CANAS-YANAOCA	1
APURIMAC-ABANCAY-CURAHUASI	1
CUSCO-URUBAMBA-URUBAMBA	10
CUSCO-CANCHIS-SICUANI	7
CUSCO-LA CONVENCION-SANTA TERESA	1
CUSCO-PAUCARTAMBO-CAICAY	1
CUSCO-QUISPICANCHI-OROPESA	4
CUSCO-LA CONVENCION-SANTA ANA	4
CUSCO-CALCA-PISAC	3
CUSCO-QUISPICANCHI-URCOS	3
AREQUIPA-AREQUIPA-JOSE LUIS BUSTAMANTE	1
CUSCO-ANTA-ANCAHUASI	1
CUSCO-CALCA-CALCA	4
CUSCO-ESPINAR-ESPINAR	1
CUSCO-QUISPICANCHI-ANDAHUAYLILLAS	3
CUSCO-LA CONVENCION-ECHARATE	2
CUSCO-QUISPICANCHI-HUARO	1
CUSCO-ANTA-PUCYURA	1
CUSCO-CUSCO-POROY	1
Total	1451

Table 2. Results by site.

Site	Dropouts
Cusco	1317
Puerto Maldonado	134
Total	**1451**

© 2026 ITMA. All rights reserved.

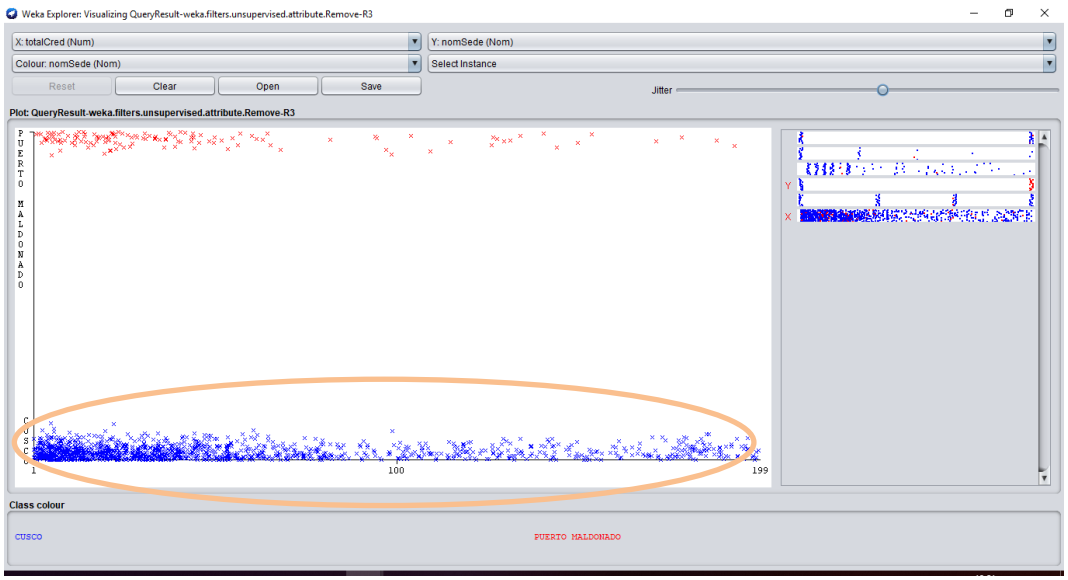


Fig. 8. Scatter plot by site.

Figure 9 shows the scatter plot by district, showing several districts grouped, with a greater predominance in the Cusco district.

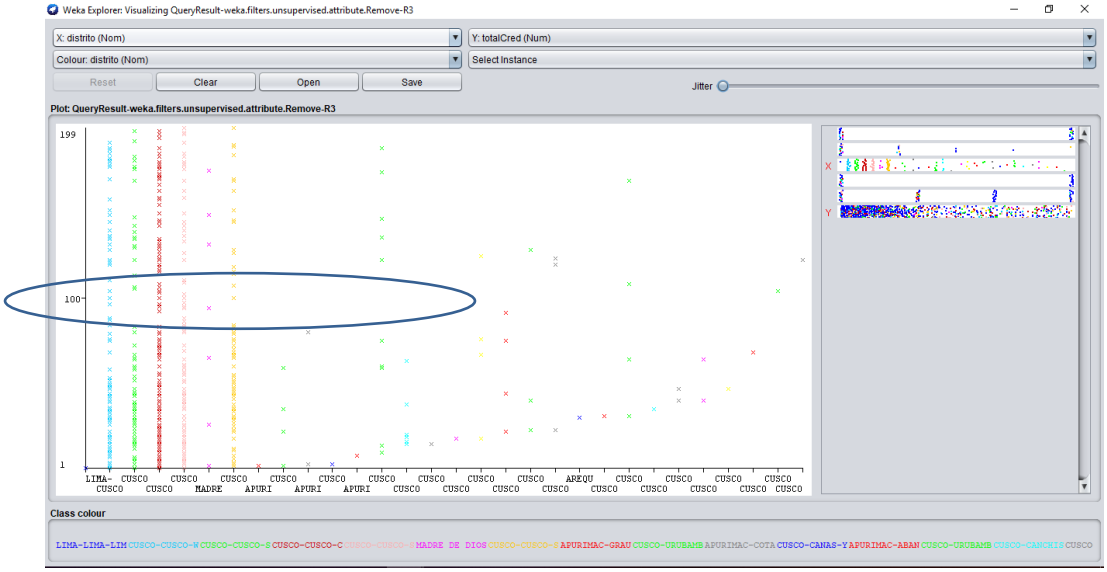


Fig. 9. Scatter plot by district.

Different tests were generated based on the K-means algorithm. In these tests, the number of clusters was varied. The test with 4 clusters identifies the main districts to which the dropouts belong, as shown in Figure 10.

Cluster#				
0	1	2	3	4
(212.0)	(585.0)	(353.0)	(121.0)	(180.0)
M	M	M	M	F
SOLTERO(A)	SOLTERO(A)	SOLTERO(A)	SOLTERO(A)	SOLTERO(A)
CUSCO-CUSCO-SAN JERONIMO	CUSCO-CUSCO-CUSCO	CUSCO-CUSCO-CUSCO	CUSCO-CUSCO-SAN SEBASTIAN	CUSCO-CUSCO-CUSCO
CUSCO	CUSCO	CUSCO	CUSCO	CUSCO
PLAN 2013	PLAN 2005	PLAN 1993	PLAN 2005	PLAN 1993
52.3302	58.3521	29.119	122.0992	34.6

Fig. 10. K-means results with 04 clusters

© 2026 ITMA. All rights reserved.

Table 3 summarizes the detailed results of the test with 4 clusters, showing the main districts with dropouts along with other fields.

Table 3. Results showing districts

Field	Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Sex	Male	Male	Male	Male	Female
Marital status	Single	Single	Single	Single	Single
District	Cusco, Cusco, San Jeronimo	Cusco, Cusco, Cusco	Cusco, Cusco, Cusco	Cusco, Cusco, Sn Sebastián	Cusco, Cusco, Cusco
Site	Cusco	Cusco	Cusco	Cusco	Cusco
Curricula	Plan 2013	Plan 2005	Plan 1993	Plan 2005	Plan 1993
Credits	52.3302	58.3521	29.119	122.0992	34.6
Total per field	212	585	353	121	180
Total	1451				

A socioeconomic analysis of the districts was made through indicators based on the Peru 2018 District Monetary Poverty Map. The districts listed with the highest dropouts have average percentages of monetary poverty above 7.7% and at their highest levels up to 18% as in the district of San Jeronimo. These data were taken from [10].

Figure 11 shows the bar chart of the average monetary poverty level of the 3 districts with the highest attrition. The figure also shows the bar chart of the dropout rate in each of these districts. Comparing both data Monetary poverty level and the dropout rate for each district.

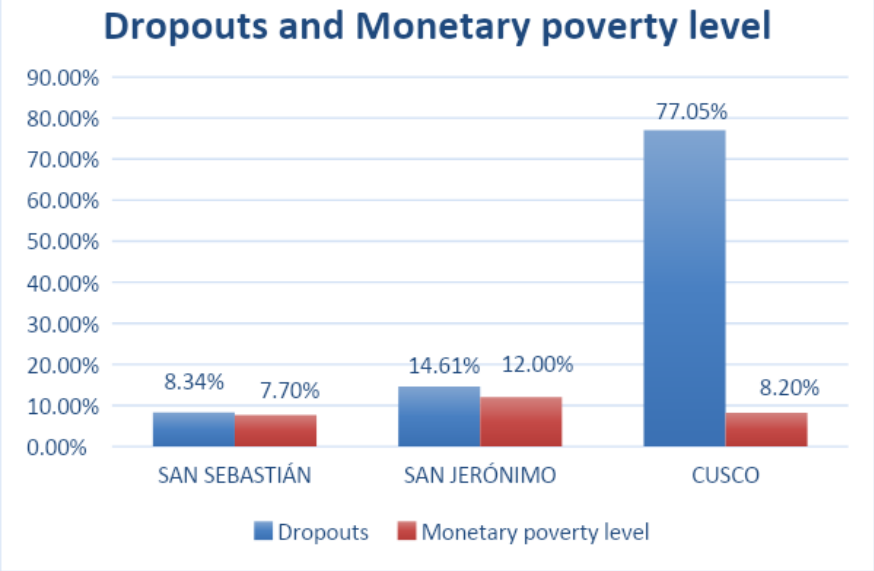


Fig. 11. Dropouts and Monetary poverty level

© 2026 ITMA. All rights reserved.

4. Discussion

Regarding geographic patterns, the first result found after data mining is that student dropouts occur in greater volume at the Cusco site compared to the Puerto Maldonado site.

About the districts, the city of Cusco has 8 districts. In the initial processing, there are several districts to which the dropouts belong, but in the K Means algorithm test with 4 groups, 3 districts, San Sebastian, San Jeronimo, and Cusco, appear as the districts with the highest dropouts.

In the socioeconomic analysis, according to the District Monetary Poverty Map. For the districts of San Sebastian and San Jeronimo, a similar relationship is identified between the levels of dropout and district monetary poverty. However, this does not occur in the district of Cusco, which, despite having high levels of dropout, has a low level of monetary poverty compared to the other districts. This analysis was carried out in order to identify the relationship between the indicators.

The results obtained will help to understand and recognize the importance of geographic location data in the problem of academic dropout in higher education. These patterns and trends will also help academic directors to improve institutional decision-making with respect to dropouts.

The objective of showing geographic location patterns at the district level is to identify and prioritize the areas with the highest dropout rates.

From the discovery of this knowledge and the traditional strategies that universities apply to their students, some measures can be improved, such as:

- Designing attention policies for students coming from dropout sites and districts.
- Articulating preventive tutoring strategies with dropout district municipalities.
- Focusing on information campaigns aimed at the areas with the highest dropout rates.
- Articulate and direct vocational orientation campaigns in the last years of secondary education in the main schools of each dropout district.

Consider the level of monetary poverty as a criterion for directing social campaigns such as scholarships, agreements, and discounts in dropout districts with low socioeconomic status.

The study has certain limitations, such as the dataset used as a resource having fields of academic origin only. Another limitation is that the dataset belongs to only one academic program. The visualization software does not associate maps of the districts of the city of Cusco, which did not allow for viewing the geographic report.

5. Conclusion

- Using the academic dropout model through the Clustering technique, it was possible to discover geographic patterns in the dropout behavior.

© 2026 ITMA. All rights reserved.

- From a data set of 10 years of operation of a professional program, a dropout model was constructed, resulting in 1451 dropout records.

- A careful work of data cleaning and a methodical process of Extraction, Transformation, and Loading will allow building an adequate dropout model to discover knowledge.

- The K-Means algorithm doesn't establish a defined number of groups; therefore, the researcher needs to review the results carefully until outstanding patterns are found.

- The geographic patterns identified are the location of the professional program and the district to which the dropouts belong.

- The pattern identified as a site shows a greater predominance of dropouts in Cusco, and the second site with the highest number of dropouts is Puerto Maldonado.

- Using the K Means algorithm with 4 clusters, 3 districts with predominant dropouts were identified: Cusco, San Sebastian, and San Jeronimo.

- The application of clustering allowed the district data to be grouped by maximizing intraclass similarity, which generated scatter plots of the districts with the highest concentration of dropouts.

- The socioeconomic analysis around the geographic data would be improved by having students' economic data, which would confirm the relationship between the dropout district and the level of monetary poverty.

References

[1] Alvarado-Uribe J., Mejía-Almada P., Masetto Herrera A.L., Molontay R., Hilliger I., Hegde V., Montemayor Gallegos J.E., Ramírez Díaz R.A., Ceballos H.G. Student Dataset from Tecnológico de Monterrey in Mexico to Predict Dropout in Higher Education (2022) Data, 7 (9), art. no. 119, Cited 0 times. DOI: 10.3390/data7090119. https://www.scopus.com/inward/record.uri?eid=2-s2.0-85138644437&doi=10.3390%2fdata7090119&partnerID=40&md5=0cfe5b56adb51128c3e0cf3305dce401

[2] Bedregal-Alpaca, N., Cornejo-Aparicio, V., Zárate-Valderrama, J., & Yanque-Churo, P. (2020). Classification models for determining types of academic risk and predicting dropout in university students. International Journal of Advanced Computer Science and Applications, 11(1).

[3] Sosa Jauregui, V. D. (2019). Modelo para la evaluación del perfil del ingresante universitario utilizando técnicas de Data Mining. REPOSITORIO UNSA, 1(1), 133. http://repositorio.unsa.edu.pe/bitstream/handle/UNSA/11023/UPsojavd.pdf?sequence=1&isAllowed=y

© 2026 ITMA. All rights reserved.

[4] Miranda, M. A., & Guzmán, J. (2017). Análisis de la deserción de estudiantes universitarios usando técnicas de minería de datos. Formación universitaria, 10(3), 61-68.

[5] Han, J., Pei, J., & Tong, H. (2022). Data mining: concepts and techniques. Morgan kaufmann.

[6] Choque Soto, V. M. (2019). Minería de datos aplicada a la identificación de factores de deserción universitaria en programas de pre grado.

[7] Nanda, S., & Panda, G. (2014). A survey on nature inspired metaheuristic algorithms for partitional clustering. Swarm and Evolutionary computation, 1-18.

[8] Talend Open Source. (20 de 10 de 2022). Talend. Obtenido de Talend: https://www.talend.com/products/talend-open-studio/

[9] Weka Project (20 de 10 de 2022). https://waikato.github.io/weka-site/index.html

[10] Instituto Nacional de Estadistica e Informatica. (2018). Mapa de pobreza monetaria distrital (1st ed.). INEI. https://www.inei.gob.pe/media/MenuRecursivo/publicaciones_digitales/Est/Lib1718/Libro.pdf

© 2026 ITMA. All rights reserved.