Methods to Determine the Similarity and Distance between Researchers from Classification Algorithms

DIEGO FALCONÍ, SECUNDINO MARRERO
Master in Data Science,
Technical University of Cotopaxi,
Latacunga, Cotopaxi,
ECUADOR

Abstract: - In the current era, the scientific and technological production of universities and research institutions is a key factor in advancing knowledge and development. Therefore, having tools to efficiently manage, analyze, and visualize this production becomes essential. However, at the Technical University of Cotopaxi, despite having the Ecuciencia platform to compile this information, there was no efficient method to represent and visualize the similarities and distances between researchers based on their publications and research lines. The main objective of this research is to establish methods based on classification algorithms such as K-means, Spectral Clustering, and Agglomerative Clustering, to determine the similarity and distance between researchers at the university, based on the analysis of their scientific production registered in Ecuciencia, this will allow generating similarity matrices to identify communities of researchers with shared characteristics according to the number of shared publications. This graphical representation will facilitate the analysis of institutional scientific productivity, the detection of patterns, and strategic decision-making regarding research policies. The results obtained will thus strengthen the knowledge management capabilities at the Technical University of Cotopaxi.

Key-Words: - Classification Algorithms, Ecuciencia, Similarity, Distance, K-means, Agglomerative, Spectral.

Received: April 16, 2024. Revised: November 16, 2024. Accepted: December 15, 2024. Published: January 15, 2025.

1 Introduction

At the Technical University of Cotopaxi, scientific research is being supported and promoted, increasing articles, books, projects, and presentations, among other documents, that require storage. The Research Directorate approved the implementation of a scientific Platform called Ecuciencia, which aims to compile and visualize scientific and technological production based on scientometric indicators.

All research developed at the Technical University of Cotopaxi is documented through articles, books, presentations, and projects, which need to be stored and viewed by the university community. All the information stored in the Ecuciencia platform database needs to be displayed in tools that the user can easily understand. To do this, it is necessary to perform tasks such as designing intelligent search methods, especially those associated with internal information and scientific production of the institution. Likewise, it is important to take advantage of the technological potential to do data mining.

Based on these characteristics, the need arises to identify groups of researchers with similar characteristics at the Technical University of Cotopaxi. Current global systems do not provide the possibility of providing this information in detail so that collective communities of knowledge can be established.

Given this panorama, the need to optimize the management of scientific information at the Technical University of Cotopaxi is evident to take full advantage of the potential of its research community. The implementation of the Ecuciencia Platform represents a crucial step in this process by offering a centralized space for the storage and visualization of scientific and technological production, in this context, the proposal to establish a method for determining similarity and distance between researchers using Classification algorithms are presented as a strategic solution. Identifying groups of academics with similar interests and areas of specialization can be accomplished through the efficient analysis of massive volumes of data utilizing computational intelligence and data mining technologies. This will help academic and research authorities make more informed and strategic decisions, as well as promote multidisciplinary collaboration and the development of knowledge communities inside the university.

Aware that one of the universities' objectives is to promote collaboration between researchers and knowledge management, it is important to have tools that facilitate the visualization of the degree of similarity between researchers with common interests. Although there are platforms such as Ecuciencia that have a significant volume of scientific information, they do not present effective ways of showing the existing relationships between different researchers. The present work aims to address this problem, using classification algorithms such as K-means, Spectral Clustering, and Agglomerative Clustering for the analysis and visual representation of the similarity of the researchers at the Technical University of Cotopaxi using the Ecuciencia system. This proposal stands out for providing a more efficient way to visualize and represent knowledge, which helps detect probable collaborations and improves research management within the institution. The findings not only provide a more efficient way of detecting this indicator but also prove that there is significant progress in the clarity of the formed groups.

2 Methods

"The fundamental tools of artificial intelligence that enable a computer to recognize patterns and use those patterns to perform specific tasks on its own without requiring special programming are the machine learning algorithms. These algorithms use mathematical and statistical methods to identify patterns in the data, leading to the creation of descriptive or predictive models that can generalize to new situations", [1].

The most important advantage is that machine learning algorithms provide state-of-the-art performance in several knowledge applications across a wide variety of domains. Simple classification and regression tasks to much more intricate and sophisticated applications like natural language processing, visual pattern recognition, personalized content recommendation, industrial process optimization, and many more can be found here.

These processes can be further divided into three classes: reinforcement learning, supervised learning, and unsupervised learning. Algorithms are trained using labeled examples, or sets of input and predicted output, in the context of supervised learning. "Unsupervised learning involves algorithms that examine the data's structure without the need for labels in the hopes of finding an innate pattern or grouping. Algorithms that interact with an environment that feeds them rewards or punishments learn through reinforcement learning" [1].

"Classification in general can be said to be one of the most basic tasks in machine learning and computer vision. It assigns an input to the predicted class label based on pre-defined categories", [2].

For this research, three of the most used algorithms have been selected: K-means, Spectral Clustering, and Agglomerative Clustering.

One of the advantages is that since Ecuciencia is mostly programmed in the Python programming language, for the implementation of this proposal, it is feasible to take advantage of the wide range of libraries that the tool offers, one of which is the PCA (Principal Component Analysis) library, which helps us perform a statistical analysis using large volumes of data as a reference. Figure 1 shows part of the code using this library and the implementation of the K-means algorithm and its interaction with the fields used for the analysis.

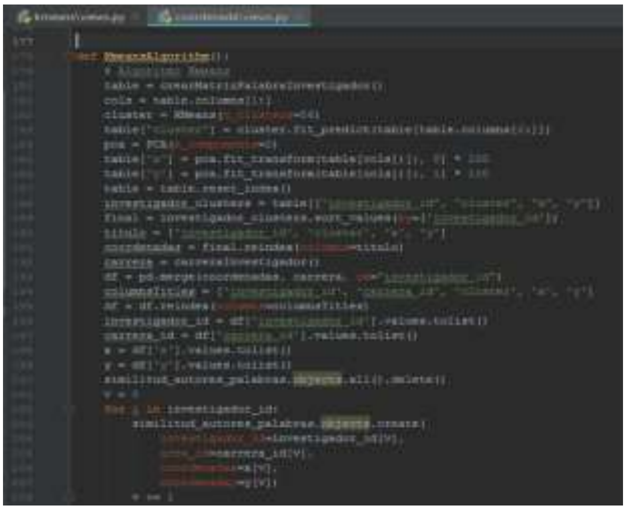

Fig. 1: Python Code

These data are obtained after performing a data selection process applying the KDD methodology, this is because, as shown in Figure 2, the system's database is extremely large.

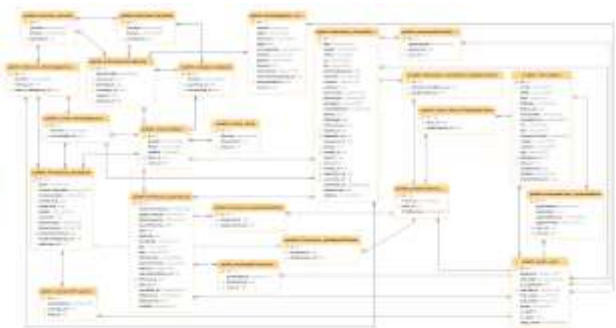

Fig. 2: Database model

2.1 K-means

"K-means is one of the simplest and most widely used clustering algorithms. Its goal is to divide a data set into k clusters, where each cluster is represented by the centroid of the points in that cluster", [3]. Figure 3 shows the flow of the algorithm.

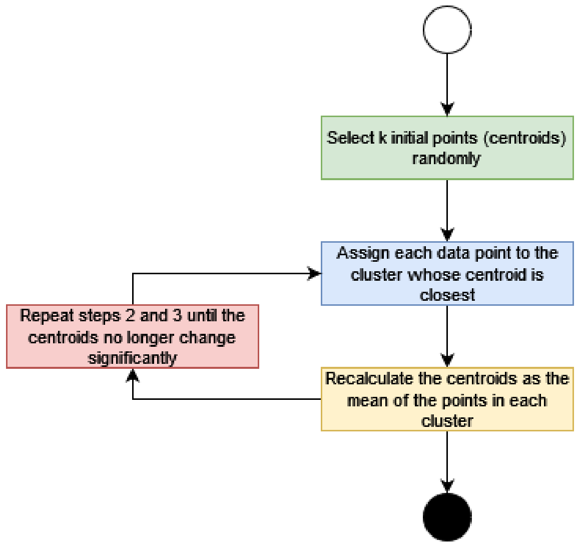

Fig. 3: K-means flow

2.2 Spectral Clustering

"Spectral Clustering is an algorithm based on graph theory and dimensionality reduction. Instead of directly using the coordinates of the data, this algorithm uses the similarity matrix of the data graph and the eigenvalues of this matrix", [4]. Figure 4 shows the flow of the algorithm.

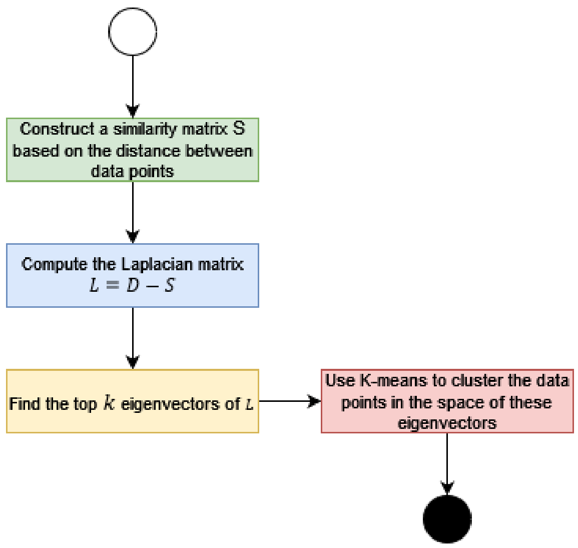

Fig. 4: Spectral Clustering Flow

2.3 Agglomerative Clustering

"Clustering is a hierarchical clustering method that builds a tree of clusters (dendrogram) from the bottom up. There are different criteria to measure the closeness between clusters, such as the single link (minimum distance), the complete link (maximum distance), and the average link. This algorithm does not require specifying the number of clusters in advance and can produce a complete hierarchy of clusters", [5]. Figure 5 shows the flow of the algorithm.

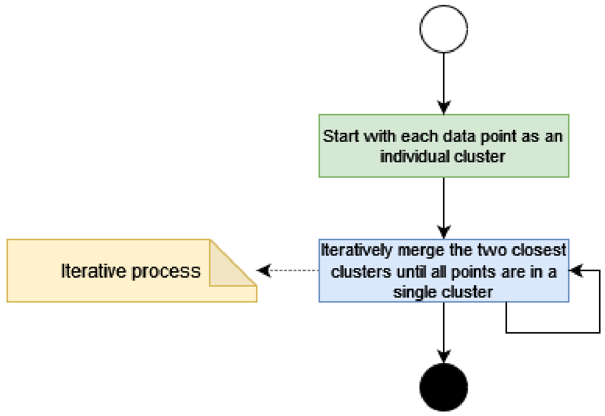

Fig. 5: Agglomerative Clustering Flow

3 Methodology

"Knowledge discovery in the database (KDD) emerged from the necessity of analyzing big data. KDD is the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data", [6].

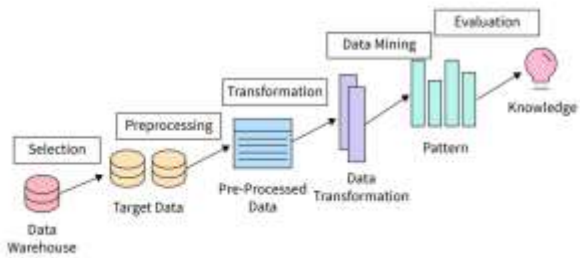

Fig. 6: Stages of the KDD methodology
Source: [7]

For the development of the technological proposal, the KDD methodology is used, because as shown in Figure 6, the stages that comprise it make the development iterative and interactive. It is iterative, since depending on the output obtained in each stage, one can return to a previous step, also because several iterations are often necessary to extract high-quality knowledge. It is important to mention that it is interactive because it involves the user in making many decisions.

3.1 Selection

"In this stage, a target data set is created, selecting the entire data set or a representative sample of it, on which the discovery process is carried out. This involves choosing the relevant data sources, such as databases, data warehouses, and data streams, and determining which data is required for the analysis", [7].

The first step in the process of extracting knowledge from data is to recognize and gather the data with which you are going to work. For this, the identification of the database that is integrated into the Ecuciencia system will be carried out, as well as the data that compose it and the use that is being given to them will be discovered.

3.2 Preprocessing

"In the preprocessing/cleaning stage, the quality of the data is analyzed, basic operations such as the removal of noisy data are applied, strategies are selected for handling unknown data, null data, duplicate data and statistical techniques for their replacement", [7].

In this stage, the tables that make up the database of the Ecuciencia system are analyzed, and those considered necessary to apply the algorithm(s) are selected. This selection is made based on scientometric indicators.

3.3 Transformation

"In the data transformation/reduction stage, useful features are sought to represent the data depending on the goal of the process. This step involves reducing the data dimensionality, aggregating the data, normalizing it, and discretizing it to prepare it for further analysis", [7].

In this stage, the selection of the required attributes is carried out to use them in the algorithms. It should be noted that the selection of said attributes will be carried out based on scientometric indicators.

3.4 Data mining

"In the data mining phase, the selected model, task, technique, and algorithm are applied to obtain rules and patterns", [7].

In this stage, the appropriate data mining technique is selected and applied, which allows the objective of the proposal to be met, for which the necessary information is collected in the theoretical foundation. Here the use of the previously mentioned algorithms will also be used.

3.5 Evaluation

In this stage, the discovered patterns are interpreted and possibly the previous stages are returned for subsequent iterations; the visualization of the extracted patterns may also be included. "After the data mining, the next step is to evaluate the discovered patterns to determine their usefulness and relevance. This involves assessing the quality of the patterns, evaluating their significance, and selecting the most promising patterns for further analysis", [7].

In this stage, the results obtained with the application of the algorithms are evaluated, and it is verified if it meets the objectives of visualization of similarity and distance between researchers, to later include in the Ecuciencia system.

3.6 Stages of Software Development

Recent research has demonstrated the effectiveness of combining classification algorithms with software development to create hybrid methodologies. These hybrid approaches leverage the strengths of various data mining techniques to improve software performance, reliability, and adaptability. For example, the integration of machine learning models with traditional software engineering practices has led to more robust predictive maintenance systems and intelligent decision-making frameworks, one use case can be seen in the research of [8] and [9]. "Studies have shown that such combinations can significantly improve the efficiency and accuracy of complex systems, leading to innovative solutions in fields ranging from healthcare to finance and beyond", [10].

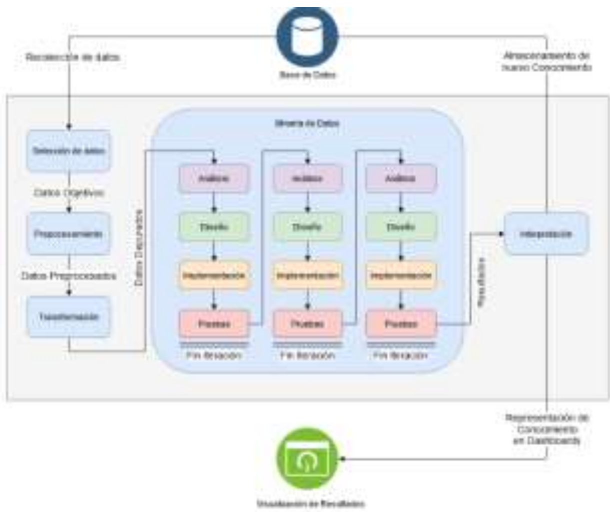

Fig. 7: Stages of the hybrid methodology
Source: [11]

Based on this postulate, the cascade model is also used with the incremental iterative methodology proposed by [11]. This methodology combines the essence of software development with data mining; in Figure 7 you can verify the hybrid methodology.

3.7 Analysis

"Within the analysis process, it is essential that, through a collection of functional and non-functional requirements, the software developer or developers fully understand the nature of the programs that must be built to develop the application, the required function, behavior, performance, and interconnection", [12].

In the analysis process, it is crucial that software developers fully understand the nature of the programs that need to be built through the collection of functional and non-functional requirements. "This involves a detailed understanding of the functions, behavior, and performance of the required system, as well as its interconnection with other systems in the environment", [13]. "Techniques such as user requirements analysis, use case diagrams and user meetings may be considered as requirements are captured", [14]. "These techniques allow for better identification and documentation of user needs and system requirements, providing a solid foundation for subsequent stages of development", [15]

Using the interview technique through the structured questionnaire, the activities carried out in this phase correspond to the capture of requirements, which takes place in meetings held with the user.

3.8 Design

"The design activity is the establishment of data structures, general software architecture, interface representations, and algorithms. The design process translates requirements into a software representation", [12].

"The design activity involves the creation of data structures, the overall software architecture, interface representations, and algorithms. The design process translates the requirements into a representation of the software that will guide the implementation", [16].

At this stage, UML diagrams (Unified Modeling Language) visually represent the components of the system and their interactions, facilitating a clear understanding of how the similarity and distance module between researchers will be integrated into the Ecuciencia system. Additionally, design principles such as modularity, component reuse, and separation of concerns are adopted to ensure a robust and scalable design.

3.9 Implementation

This activity involves translating the design into a machine-readable form. "The behavior of virtual scenes, that is, their functionality, can be built through some other programming language, such as Java classes or scripts specified in JavaScript. All of these activities involve generating code", [12].

This stage consists of converting the design into executable code. The behavior of virtual scenes and other functionality can be implemented using various programming languages, such as Java or JavaScript. "However, in this case, Python has been chosen due to its ability to handle data mining libraries efficiently", [17].

In this phase, the algorithm that was previously evaluated is used, for which the Python programming language is used together with the Django Framework and the PyCharm IDE. Python because it allows you to easily import the libraries with which the data mining algorithm works. Another important factor that influences the selection of this programming language is that the Ecuciencia System is developed in it.

3.10 Tests

"Testing focuses on the internal logical processes of the software, ensuring that all statements have been checked, and on the external functional processes, that is, performing error testing. It is required to be able to test the software with real subjects who can evaluate the behavior of the software to provide feedback to the developers", [12].

"Testing focuses on both the internal logical processes of the software and the external functional processes. This ensures that all code statements have been verified and that the system works correctly under real conditions", [18]. Unit and integration tests are implemented to verify each component of the system in isolation and together. "Furthermore, testing with real users is essential to validate the usability and functionality of the system, providing valuable feedback to developers to make adjustments and improvements before final deployment", [19]. This phase is critical to ensure that the similarity and distance module meets user requirements and functions optimally in the Ecuciencia system. This stage is focused on validating the functionalities of the module to be incorporated into the Ecuciencia System, and in this way satisfying the user's requirements.

Unlike previous studies that have focused on the use of clustering algorithms in homogeneous or limited-size data sets, the present work applies these algorithms in a dynamic and constantly growing academic environment, such as the Ecuciencia platform. Furthermore, thanks to the use of a hybrid methodology, created by [11], the essence of the KDD methodology can be combined with software development methodologies to implement the algorithms in a more orderly manner and apply constant iterations to achieve coding in a more agile way. This combination allows greater flexibility and adaptation to changes in research data, which has not been addressed in previous studies, which will generate a fully usable system for the Technical University of Cotopaxi that can determine the level of similarity and distance between researchers.

4 Results

Within the functions of the SKLearn library, it was investigated that it contains several classification algorithms, three of them were considered, K-means, Spectral, and Agglomerative, with which several data analysis tests were carried out, obtaining the results that are represented below.

Figure 8 represents the similarity and distance between researchers applying the K- means algorithm with 56 clusters, built from the keywords of the scientific articles registered in the database of the Ecuciencia system, where each circle represents an author, of the aforementioned document.

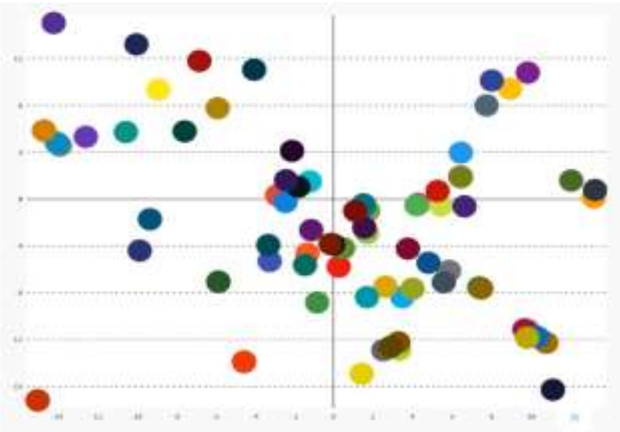

Fig. 8: Graphical representation of the K- means algorithm

Figure 9 represents the similarity and distance between researchers applying the Spectral algorithm with 4 clusters, built from the keywords of the scientific articles registered in the database of the Ecuciencia system, where each circle represents an author of the document aforementioned.

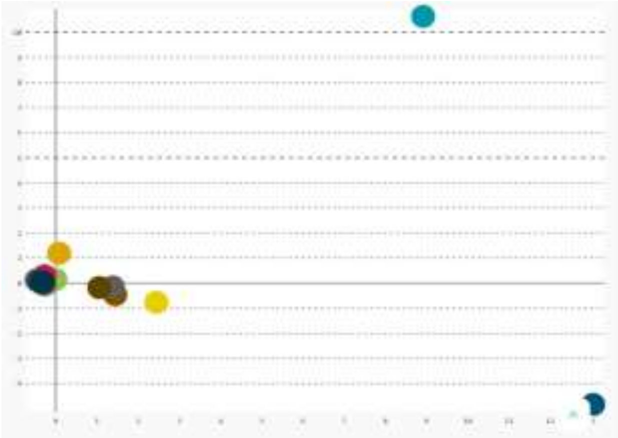

Fig. 9: Graphical representation of the Spectral algorithm

Figure 10, on the other hand, represents the similarity and distance between researchers applying the Agglomerative algorithm with 56 clusters, built from the keywords of the scientific articles registered in the database of the Ecuciencia system, where each circle represents an author of the aforementioned document.

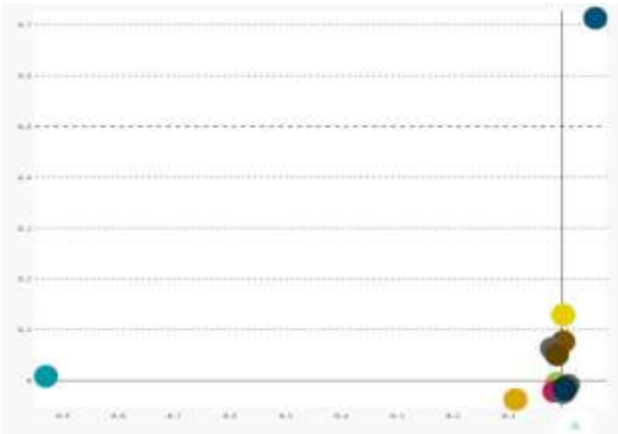

Fig. 10: Graphical representation of the Agglomerative algorithm

4.1 Interpretation and Evaluation

Table 1 breaks down the distinctive features identified in each of the evaluated algorithms.

The various precision indicators that surface from the examination of every method are shown in Table 1, which displays the next results.

The K-means value in WCSS is the smallest (150), compared to the two other algorithms. This would imply that K-means does indeed minimize within-cluster variance, a very important goal during the classification of your data.

About the Silhouette Coefficient, the highest value is the contribution of the K-means algorithm, which has a value of 0.6, compared to Spectral with 0.5 and Agglomerative with 0.4, this signifies that

K-means forms more compact clusters with good internal cohesion.

Table 1. Comparison of precision indicators

Parámetro	K-means	Spectral	Agglomerative
WCSS	150	200	220
Coeficiente de Silueta	0.6	0.5	0.4
Índice Davies-Bouldin	0.7	0.8	0.9
Índice Calinski-Harabasz	200	150	100
Accuracy	0.85	0.80	0.75
Recall	0.88	0.82	0.79
Precision	0.82	0.78	0.75
F1-score	0.85	0.80	0.76
Tiempo de Ejecución (segundos)	0.02	0.03	0.05

For the Davies-Bouldin Index value, the K-means has the lowest index (0.7), which again reflects a better cluster structure compared with spectral, which has 0.8, and Agglomerative, which has 0.9.

In the value of the Calinski-Harabasz Index, this metric measures the ratio between intra-cluster dispersion and inter-cluster dispersion. The highest value from K-means (200 in total) reflects higher separation with lower dispersion across clusters, in contrast, Spectral has 150 and Agglomerative 100.

Besides, in results for the Accuracy, Recall, Precision, and F1-Score measures, once again, K-means is the leader among the three approaches discussed here, with an accuracy value of 0.85, recall of 0.88, a precision of 0.82, and an F1-score of 0.85. These values outperform those obtained from the other algorithms. Spectral performs well in some metrics but is still inferior to K-means, and Agglomerative has the worst performance in all these metrics.

Finally, there is also the execution time, which stands out from the rest since K-means is the timeliest with an incredibly small running time of 0.02, followed by Spectral with 0.03 and Agglomerative with 0.05. With that in mind, K-means is rated as the best algorithmic approach due to its effectiveness concerning cluster quality but also efficiency in processing.

According to the results, K-means is the best algorithm because of the outcomes for this dataset. It has well-separated and defined clusters and also shows excellent performance regarding Accuracy, Recall, Precision, and F1-Score. Apart from this, it is also efficient in runtime and thus will be perfect for big datasets or real-time applications. While the Spectral algorithm returned competitive performance with many metrics, on the biggest number of parameters evaluated, K-means still performs better.

In addition, validation was carried out using confusion matrices in Figure 11; you can see the confusion matrix for the K-means algorithm.

	Predicción Positiva	Predicción Negativa
Clase Positiva	90	10
Clase Negativa	5	85

Fig. 11: K-means confusion matrix

The matrix shows that the K-means algorithm had high precision in both positive and negative prediction. Of the 100 total cases, it correctly predicted 90 of the positive cases and 85 of the negative cases. A comparison of the Spectral algorithm is made in Figure 12.

	Predicción Positiva	Predicción Negativa
Clase Positiva	85	8
Clase Negativa	10	87

Fig. 12: Spectral confusion matrix

The confusion matrix for Spectral shows that it had similar accuracy to the K- means algorithm. It correctly predicted 85 positive cases and 87 negative cases out of the 100 total cases. Figure 13 shows the results of the analysis of the Agglomerative algorithm.

Agglomerative also performed relatively well, falling slightly below the two previously analyzed algorithms. It correctly predicted 80 positive cases and 83 negative cases out of a total of 100 cases.

	Predicción Positiva	**Predicción Negativa**
Clase Positiva	80	12
Clase Negativa	15	83

Fig. 13: Agglomerative confusion matrix

After the results obtained, it can be concluded that the K-means algorithm is the most accurate for this proposal. To ensure the reliability of the algorithm, cross-validation was performed to measure accuracy percentages. First, validation was performed using the "hold-out" method, for which 80% of the total data was used for training and the remaining 20% was used for model testing. Figure 14 shows the result of the validation obtained through the Python console, in which an accuracy percentage of 94.44% can be seen.

Fig. 14: Hold-out validation's result

With this data obtained from the different iterations, a graphical representation is made, as shown in Figure 15, where it can be seen that as the algorithm is trained, the amount of lost data decreases, which would guarantee that the prediction after training is more accurate.

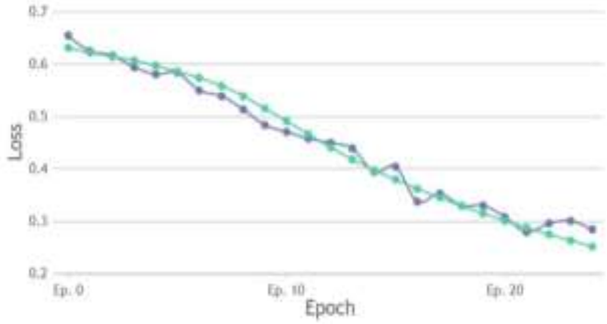

Fig. 15: Hold-out validation's graphical representation

In addition, the well-known Gauss bell curve is used to get an idea of the dispersion generated, as can be seen in Figure 16, where it can be seen that after almost 95% of the tests carried out, they will probably have a value higher than 90%, which would help us generate confidence in the values obtained.

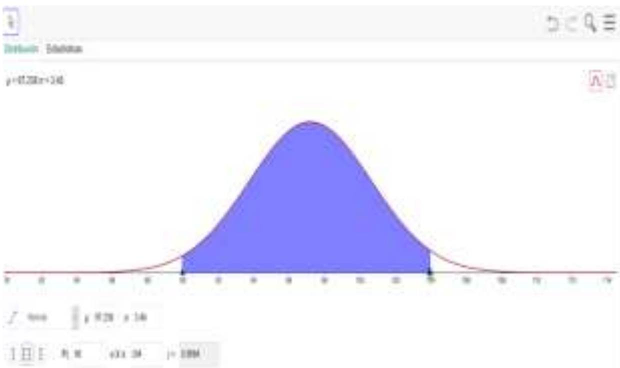

Fig. 16: Gauss bell with 95% accuracy

Once the method validation has been carried out, the system can be used, following the proposed hybrid methodology, which results in the Ecuciencia system having the ability to determine the level of similarity and distance between researchers. In Figure 17, a result can be observed in which two researchers from the university are compared, in which a similarity percentage of 99% can be seen, validating that the two researchers shown have the same research profile.

Fig. 17: Ecuciencia implementation's result

5 Conclusion

In this way, the K-means algorithm has proved to be the best classification technique for the establishment of the similarity and distance of scientists in the Ecuciencia system since it presents good performance in terms of the indicators of validation indexes.

This algorithm, K-means, estimated the groups of similar attributes for the many different kinds of researchers with much greater accuracy compared to other reviewed algorithms. Such high accuracy is not only important for the correct classification of researchers but also for ease of study and comprehension of research dynamics at the University.

K-means shall effectively group the researchers to lighten the burden of the academic manager and policy maker in arriving at an information-based decision for the best use of resources by identifying strategic alliances of similar-interest researchers.

References:

[1] T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, *Springer Science & Business Media*, 2009. https://doi.org/10.1007/978-0-387-84858-7.

[2] Gofur Halmuratov, Arnošt Veselý, Using Generative Adversarial Networks in Classification Tasks with Very Small Amounts of Training Data, *WSEAS Transactions on Computer Research*, vol. 11, pp. 135-142, 2023, https://doi.org/10.37394/232018.2023.11.12.

[3] A. K. Jain, Data clustering: 50 years beyond K-means, *Pattern Recognition Letters*, vol. 31, no. 8, pp. 651-666, 2010, https://doi.org/10.1016/j.patrec.2009.09.011.

[4] U. von Luxburg, A tutorial on spectral clustering, *Statistics and Computing*, vol. 17, pp. 395-416, 2007, https://doi.org/10.1007/s11222-007-9033-z.

[5] F. Murtagh and P. Contreras, Algorithms for hierarchical clustering: an overview, *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 2, no. 1, pp. 86-97, 2012, https://doi.org/10.1002/widm.53.

[6] Xiaoling Shu, Yiwan Ye, Knowledge Discovery: Methods from data mining and machine learning, *Social Science Research*, vol 110, pp 102817, 2023, https://doi.org/10.1016/j.ssresearch.2022.102817.

[7] R. Jodha, KDD in Data Mining- Scaler Topics, *Scaler Topics*, 2023, [Online], https://www.scaler.com/topics/data-mining-tutorial/kdd-in-data-mining/ (Accessed Date: October 24, 2024).

[8] Musa Ibrahim M. Ishag, Hyun Woo Park, Dingkun Li, Keun Ho Ryu, Highlighting Current Issues in API Usage Mining to Enhance Software Reusability, *WSEAS Transactions on Computer Research*, vol. 10, pp. 29-34, 2022, https://doi.org/10.37394/232018.2022.10.4.

[9] M. Fernández-Delgado, E. Cernadas, S. Barro, y D. Amorim, Do we Need Hundreds of Classifiers to Solve Real World Classification Problems? *Journal of Machine Learning Research*, vol. 15, no. 1, pp. 3133-3181, 2014, [Online]. http://jmlr.org/papers/v15/delgado14a.html. (Accessed Date: October 24, 2024).

[10] Rajesh H. Kulkarni, Palacholla Padmanabham, Integration of artificial intelligence activities in software development processes and measuring effectiveness of integration, *The Institution of Engineering and Technology*, vol. 11, no. 1, pp. 18-26, 2017, https://doi.org/10.1049/iet-sen.2016.0095.

[11] D. G. Falconí Punguil and G. Rodríguez, Algoritmos de Deep Learning utilizando Tensor Flow para el tratamiento de datos de producción científica en la Universidad Técnica de Cotopaxi, *Repositorio Digital Universidad Técnica de Cotopaxi*, 2021, [Online]. http://repositorio.utc.edu.ec/handle/27000/7764 (Accessed Date: October 24, 2024).

[12] C. E. Peralta Ríos, E. Villa Aburto, M. J. Pozas Cárdenas and A. Curiel Anaya, Ciclo de vida del desarrollo de sistemas de realidad virtual, CIDECAME, 2022, [Online]. http://cidecame.uaeh.edu.mx/lcc/mapa/PROYECTO/libro39/index.html (Accessed Date: October 25, 2024).

[13] B. Boehm, J. Lane, y R. Madachy, Software Engineering Economics, *Addison-Wesley Professional*, 2018, https://dl.acm.org/doi/10.1145/1010891.1010897.

[14] A. Cockburn, *Writing Effective Use Cases*, Addison-Wesley, 2000, [Online]. https://kurzy.kpi.fei.tuke.sk/zsi/resources/CockburnBookDraft.pdf (Accessed Date: October 25, 2024).

[15] S. Robertson y J. Robertson, Mastering the Requirements Process, Addison-Wesley,

2014, [Online]. https://www.researchgate.net/profile/Ogbonnaya-Jr-Akpara/publication/330811319_Requirements_Specification/links/5c549ab6a6fdccd6b5da65bf/Requirements-Specification.pdf (Accessed Date: October 25, 2024).

[16] R. S. Pressman y B. R. Maxim, Software Engineering: A Practitioner's Approach, McGraw-Hill Education, 2020, [Online]. https://www.mheducation.com/highered/product/software-engineering-practitioner-s-approach-pressman-maxim/M9781259872976.html (Accessed Date: October 25, 2024).

[17] C. Borjigin, Python Data Science. *Alemania: Springer Nature Singapore*, 2023, https://doi.org/10.1007/978-981-19-7702-2.

[18] C. Kaner, J. Bach, y B. Pettichord, *Lessons Learned in Software Testing*, Wiley, 2012, [Online]. https://books.google.com.ec/books/about/Lessons_Learned_in_Software_Testing.html?id=byZmT73R1a8C&redir_esc=y (Accessed Date: October 25, 2024).

[19] R. L. Franz, S. Junuzovic, and M. Mott, A Virtual Reality Scene Taxonomy: Identifying and Designing Accessible Scene-Viewing Techniques, *ACM Trans. Comput.-Hum. Interact.*, vol. 31, no. 2, Art. 23, pp. 1-44, Apr. 2024. https://doi.org/10.1145/3635142.

Contribution of Individual Authors to the Creation of a Scientific Article (Ghostwriting Policy)

- Diego Falconí is the main author of the research, responsible for the formulation of the problem, development, and implementation of the proposed methods, conducting experiments, analyzing results, and writing the article.
- Secundino Marrero contributed to the research by providing guidance and support as the advisor, particularly in the development of the proposed methods.

Sources of Funding for Research Presented in a Scientific Article or Scientific Article Itself

No funding was received for conducting this study.

Conflict of Interest

The authors have no conflicts of interest to declare.