Outlook

RV: CI3 2023 notification for peper 5311

De: ci32023@easychair.org <ci32023@easychair.org> en nombre de CI3 2023 <ci32023@easychair.org>
Enviado: domingo, 6 de agosto de 2023 18:40
Para: Diego Maiquez <dmaiquez@tecnologicoismac.edu.ec>
Asunto: CI3 2023 notification for paper 5311

Dear Diego Maiquez:

We are pleased to inform you that your work "La ciencia de datos: Machine Learning y Análisis multivariante en los estilos de aprendizaje" has been accepted for oral presentation and publication at the IV International Conference on Research and Innovation - CI3 2023, to be held from August 30 to September 1 of this year.
All papers accepted at the conference will be published in the CI3 2023 Proceedings and indexed in the SCOPUS bibliographic database." To prepare the FINAL VERSION of your paper (version to be published), read carefully and follow the instructions below:

- Papers must be prepared according to the template taken as reference, without page numbering, written in English and with an extension of 12 to 15 pages, including figures and references. The template can be downloaded from the conference website https://ci3.tech/es/.
- All the suggestions made by the reviewers must be considered for the FINAL VERSION of your paper.
- It is suggested that a native English speaker review the paper.
- The document must contain the names of the authors, affiliations and correspondence email (according exactly to the template).
- Final versions will be subject to a similarities check using the TURNITIN software. The maximum percentage of similarities accepted for publication is 15%.

For your paper to be published, it is MANDATORY that the document meets the requirements established in the previous point and that at least one of the authors (speaker) register for the event. Registration must be made through the CI3 2023 website in the PUBLICATIONS section (https://ci3.tech/es/registration/) and with the payment of the corresponding fee.

You must confirm your participation in the conference and submit the final version of your paper by August 27, 2023, with the following attached documents:

- Complete article in MS-WORD or Latex format (1 file).
- Complete article in PDF format (1 file).
- Scan the proof of payment of the conference registration, including paper ID in the Easychair platform, 1st author name and affiliation (1 file).
- For the presentation of the speaker on the day of his presentation, a document (Word or pdf) is required with the paper title, full name, academic degree and mini biography (5 lines) of the speaker (1 file).

- The 4 documents must be sent to info@ci3.tech, with a copy to ci3@ister.edu.ec.

Finally, we ask that you consider papers published in earlier versions of CI3 for citation in your accepted publication. The papers published in past editions can be consulted at the following link: https://ister.edu.ec/congreso/technical-papers/.

Do not hesitate to contact us if you need more information (info@ci3.tech).

We look forward to hearing your presentation at CI3 2023!

Regards,

Ph.D. Marcelo Zambrano V.
GENERAL CHAIR
CI3 2023

SUBMISSION: 5311
TITLE: La ciencia de datos: Machine Learning y Análisis multivariante en los estilos de aprendizaje

REVIEW 1
SUBMISSION: 5311
TITLE: La ciencia de datos: Machine Learning y Análisis multivariante en los estilos de aprendizaje
AUTHORS: Diego Maiquez, Diego Pabon, Mariela Condor, Gonzalo Rodriguez, Ana Oyasa and Mauricio Farinango

Overall evaluation SCORE: -2 (reject)
TEXT:
Los autores presentan solo un estudio de las técnicas y herramientas en la ciencia de datos.
El manuscrito carece de contribución significativa al campo.
Se sugiere primeramente según lo expuesto en el artículo aplicar estos modelos analizados en la segunda etapa y así evaluar los rendimientos obtenidos.
Hay carencia de resultados, en esta fase solo se ha realizado un estudio de los modelos de ML. Aunque los autores proponen un comparación de los modelos de ML y esta no se evidencia.

REVIEW 2
SUBMISSION: 5311
TITLE: La ciencia de datos: Machine Learning y Análisis multivariante en los estilos de aprendizaje
AUTHORS: Diego Maiquez, Diego Pabon, Mariela Condor, Gonzalo Rodriguez, Ana Oyasa and Mauricio Farinango

Overall evaluation SCORE: 2 (accept)
TEXT:
I am pleased to accept this paper for publication. The study explores the application of data science in education, specifically focusing on the analysis and comparison of data science

models to identify behavioral patterns, similarities, and anomalies in student data. The objective is to generate new and previously unknown knowledge and characterize student profiles based on learning styles, such as those proposed by Felder and Silverman.

The paper successfully demonstrates the effectiveness of various data science techniques in achieving the research objectives. The clustering method, particularly the k-means algorithm, proves efficient in extracting group characteristics. The decision tree model, employing the ID3 algorithm, achieves improved classification through information gain. Additionally, the Principal Component Analysis (PCA) model provides valuable insights by analyzing variability and reducing dimensions, especially when dealing with noisy or outlier data.

The paper also presents a data characterization approach, classifying the data into profile data, class data, and test data. This classification enables a comprehensive understanding of student profiles before and after a learning period, with class data representing competency-based student information.

Data science: Machine Learning and Multivariate Analysis in learning styles

Diego Máiquez[1[0009-0005-7211-6672]], Diego Pabón[1[0000-0002-0702-0763]], Mariela Cóndor[1[0009-0005-5101-4431]], Gonzalo Rodríguez[1[0009-0004-4047-0530]], Mauricio Farinango[1[0009-0000-8757-8216]] and Ana Oyasa[1[0009-0003-9538-0746]]

[1] Instituto Tecnológico Universitario ISMAC, Tumbaco, Belermo S2-02 Y, Quito 170184, Ecuador
dmaiquez@tecnologicoismac.edu.ec

Abstract. Data science is responsible for the analysis, interpretation and prediction of simple data to generate significant knowledge, it applies to any area that produces data, for example: sales, finance, production, health, education, etc. Data science in education uses machine learning techniques and multivariate analysis. Currently, Educational Data Mining or EDM is spoken, which unites the areas of education, Big data and data science to improve learning. In this study, data science models are analyzed and compared to identify patterns of behavior, similarity, and anomalous data from students to generate new unknown knowledge and characterize the profile of students according to learning styles such as those proposed by Felder and Silverman in addition to relating these styles to development by competencies. The results show that the most efficient methods are: Clustering and its k-means algorithm with which group characteristics are obtained, decision trees with its ID3 algorithm that through the gain of information a better classification is obtained and the PCA mathematical model or Principal Component Analysis that by its properties of variability analysis and dimension reduction allows to obtain more information from data with noise or outliers. A characterization of the data to be processed is also carried out, classifying it into profile data, class data and test data. These analyzed models will be implemented in a next phase in students of the Software Development career of the ISMAC Institute, which allows obtaining promising results to predict student learning styles and improve said learning process.

Keywords: data science, machine learning, multivariate analysis, data warehousing, learning style.

1 Introduction

The use of technology produces data, in recent years the growth of this data has been exponential. The question arises, what to do with so much data? Data science is responsible for the analysis, interpretation and prediction of a set of data to obtain significant knowledge from simple data, this is achieved with the help of information technology. Big Data and/or BI. Data science is present in areas such as: marketing, sales, finance, production, telecommunications, health, education, etc., and any area where data is produced [2,19,20].

In the area of education, there are traditional and modern data science techniques to evaluate psychometric and educational psychology data, among which are machine learning techniques and multivariate analysis techniques. Today there is talk of EDM or Educational Data Mining, which is composed of the areas of Big Data, data science and education whose objective is to improve learning, for this it uses computer mathematical methods and techniques such as those mentioned [1, 2,4,5].

Machine learning is a data science technique that is responsible for performing non-computationally programmed tasks from a training data set, it contains mainly supervised and unsupervised methods. Among the unsupervised are the clustering models that allow to associate data in groups of similar characteristics, one of the main clustering algorithms is K-means [1,2,3,7].

On the other hand, decision trees are supervised machine learning methods that allow discrete data classification tasks and continuous data regression tasks to be carried out, it is ideal for processes where there are several variables of different characteristics, among the most used algorithms is the ID3 that is useful in the analysis of the educational area as mentioned by some studies [1,9,10,11]

The multivariate analysis allows to analyze variables of different characteristics simultaneously, among the most used methods are the PCA or Principal Component Analysis that allows to reduce dimensions to obtain the main characteristics of a data set, a method closely related to the factorial analysis that is used in the area of education through psychometrics and educational psychology. [2,21]

On the other hand, there is the model for learning styles proposed by Felder and Silverman (1988) whose objective is to distinguish the learning style of students, it consists of four dimensions with their classifications of styles and they are: Comprehension with global/sequential styles, processing with active/reflective styles, perception with sensory/intuitive styles, and representation with visual/verbal styles [3,23,24].

On the other hand, competency-based learning refers to the approach to problems applied to real-world situations, a study proposes 8 elementary competencies that cover mathematics, native language, foreign language, social, etc. [25,26]. The development by competences is supported by e-learning technologies to improve the learning process [27].

The lack of knowledge of the student's profile causes problems for said student to learn efficiently, knowing the student's learning styles is useful to improve the aforementioned learning process [3,4,5,9]. In this study, the data science techniques and models applicable to analyze data from the area of education, particularly student processes, were analyzed and the following question was posed: Are predictive data science models adaptable to learning styles?

The decision tree machine learning methods with its ID3 algorithm and clustering with the k-means algorithm and the PCA multivariate analysis method or Principal Component Analysis turned out to be the most efficient for analyzing student process data related to learning styles. These data were classified into: profile data, test data and class data according to the characteristics and resources available in the study group of the ISMAC Technological Institute in the Software Development career.

This study corresponds to the first phase of the project "Learning capacities of students in the Software Development career of the face-to-face modality, using Big Data and/or BI tools" (project in progress) from which the results have been obtained, mentioned modeling that will be implemented in the second phase of said project hoping to obtain promising results that provide us with new knowledge of the student data of the sample of the mentioned institution, for this the mathematical-computer model will allow predicting the learning style of each student to later personalize and improve the student's own learning.

2 Data Science in Education

2.1 Data Science

Data science is responsible for the study of data, its interpretation, analysis and inference or prediction, to generate information and then significant knowledge about an area of study from simple data. With the advancement of technology, the data has increased exponentially, this large amount of data today is treated with Big Data technology and analyzed with data science [2, 20].

Data science analyzes any area or sector that produces data, among the main ones are: *Finances:* Detection of credit card fraud, customer or product risk analysis, etc. *Sales:* Loss prevention, activity-based recommendations, dynamic pricing, etc. *Manufacture:* Product research, improvement of production processes and product quality, etc. *Cybersecurity and threat intelligence:* Intelligent detection of cyber attacks, etc. *Medicine*: disease prevention, pharmaceutical research, etc.; and other areas such as telecommunications, government, education etc. [2].

Among the most used techniques in data science are machine learning and multivariate analysis. *Machine Learning* or learning machines are models that take advantage of the data and resources of a computer to predict or decide and make a decision on a computationally unprogrammed task. On the other hand, multivariate analysis allows analyzing several variables of different characteristics simultaneously. [1, 2, 8, 11, 20]

2.2 Machine Learning in Education

Machine learning is a data science technique that handles supervised and unsupervised methods, the supervised ones contain methods and algorithms such as decision trees and the unsupervised ones contain methods and algorithms such as clustering which are related to data analysis in the area of education due to its properties and characteristics [1,2]. Machine learning, in turn, is related to Educational Data Mining EDM as per its acronym in Spanish, which is one of the main ways to analyze information from the educational area [1].

Educational data mining or EDM is born from the union of data science, Big Data and the field of education in order to improve educational learning. The EDM has a historical trajectory that goes back to the year 2000 and was extended in formalized EDM conferences that took off in 2008. In 2011, a group of researchers formed the EDM society, with which they proposed a series of more formal definitions of this area [1,5].

Some of the most important definitions mentioned in the EDM handbook are: *Learning analytics.* - It is the measurement, collection, analysis and reports of data about students/learners and their context, with the purpose of understanding and optimizing learning and the environment where it occurs. *Academic Analytics.* – It is the application of statistical techniques and institutional data mining to produce business intelligence and solutions to universities and administrators [1].

The objective of the EDM is to mine or find a new unknown knowledge, this process is called KDD or Knowledge Discovery of Databases, this process is carried out with the aforementioned machine learning models, which according to its properties and characteristics [1,3].

The generation of new knowledge is used to solve problems such as the prediction of academic performance, creating intelligent tutors, analyzing behavior patterns, similarities and anomalous ones, quantifying the effectiveness of the teaching-learning process, developing new EDM technological tools, developing algorithms, carry out replication studies in other domains of education, etc., which allows optimizing and improving the learning process [1,3,4,5].

To manage the EDM, different technological tools are used, among the most important are the Moodle platform, which allows you to manage tasks, evaluations, grades, attendance, platform use logs, interaction in forums, etc., this information is useful for working in processing models [14]. WEKA software written in java language provides many artificial intelligence functionalities and statistical methods with much support for experimentation [5]. There is also the KEEL software written in open source java (GPLv3) which solves KDD or knowledge discovery tasks using training techniques, feature selection, discretization, missing data analysis, etc., this software has a particularity, it is focused on research and education [17].

Clustering

Clustering is an unsupervised machine learning technique that allows grouping or associating data into groups that contain similar characteristics, in such a way that pattern recognition is performed when analyzing the information. Some of the methods for clustering are: by partition, by density, hierarchies, etc., and some of the most used algorithms are:

K-means. - This algorithm is fast, robust and simple, it allows the grouping of data based on the least squared distance of a centroid, it is sensitive to outliers and noise.

Mean-shift. - This algorithm performs a grouping by updating centroids, it is computationally expensive, so it does not work properly with large amounts of data.

DBSCAN. - Data clustering based on non-parametric density to separate high-density and low-density groups into clusters, is resistant to outliers.

Hierarchical. - Technique that creates a hierarchy in the grouping of data according to their similarity with a bottom-up approach that allows to have a more informative organization of the data than other algorithms, information generally represented in a dendrogram graph [1,2, 3.7].

Decision trees

Supervised model used in machine learning that allows categorizing or classifying a data set according to established rules. It is composed of nodes, edges, in the nodes the variables to be studied are placed, this method consists of a training stage and a test stage where underfitting or overfitting are usually given in training [1,2,10]. This model solves classification tasks with discrete data and regression tasks with continuous data, they are efficient for analyzing user behavior and cybersecurity analysis. Decision trees use criteria for their operation, among the main ones are the Gini impurity or Gini index, given by

$$G = \left(\frac{1}{2n^2\mu}\right)\sum_{j=1}^{m}\sum_{k=1}^{m} n_j n_k \left|y_j - y_k\right| \tag{1}$$

and the information gain that arises from information theory and entropy [2,11]. Given by

$$I(p,n) = \left(\frac{-p}{p+n}\right) log_2\left(\frac{p}{p+n}\right) - \left(\frac{n}{n+p}\right) log_2\left(\frac{n}{p+n}\right) \tag{2}$$

Among the most used decision tree algorithms are: IBK, ID3, j48, C4.5, CART. *IBK*.- It is an algorithm that is called lazy type and does not create a single decision tree, each time it finds a new instance it generates a calculation of the relationship with other instances [6].

ID3.- It is based on the use of information gain through the entropy analysis of information theory, it consists of selecting an attribute as the root of the tree and creating a branch with each of the possible values of attribute, is focused for categorical classification, also related in the field of educational prediction [6, 9].

J48.- Induction algorithm that generates rules from subsets of the total data, an optimization is performed according to a calculation of goodness [6, 9].

C4.5.- Algorithm that works in a similar way to ID3, it works with continuous and discrete data sets, it also allows handling incomplete data sets.

CART. - Tree classification and regression algorithm, uses the Gini index and is applied to carry out splitting processes [1,6,11,14].

2.3 Multivariate Analysis in Education

Multivariate analysis is a technique that allows analyzing several variables with different characteristics at the same time. There are two types of multivariate analysis: explanatory or dependency models and descriptive or interdependence models. Descriptive models allow analyzing data from several variables that do not depend on a single variable or interdependence, some of the most used models are factor analysis and PCA or Principal Component Analysis, methods related to the area of education through psychometrics and educational psychology [2,21].

Principal Component Analysis or PCA

PCA or Principal Component Analysis is a data analysis technique that allows evaluating several variables of different kinds or that are not initially correlated, where patterns are established through the use of eigenvalues and eigenvectors in conjunction with correlation methods between the data, analyzing the variance, this method also allows a reduction of variables and works as a predictive method [2,12,13,21].

PCA is closely related to factor analysis. The PCA in education allows evaluating variables of various types such as: sociodemographic, socioeconomic, use of formal and informal resources for education, etc., and obtaining efficient results as shown by some studies [12,13,21].

Methods used in psychometrics.

Psychometrics allows us to evaluate or measure variables, aspects and characteristics in the area of education, in particular it is also applied in the evaluation of learning styles related to the learning abilities of students [15]. Among the most used methods are the Cronbach's Alpha coefficient to assess the reliability of a measurement scale and the Bartlett test to assess the variance of different population samples [16,21].

On the other hand, factor analysis allows analyzing correlation in unobserved variables from a set of observed variables, there are two types: Exploratory factor analysis or AFE and confirmatory factor analysis or AFC. On the other hand, the R-squared coefficient of determination allows evaluating the quality of fit of a linear regression model. If the coefficient is close to 1, it will have a greater linearity fit and therefore better quality; otherwise, the closer to 0 the model will be not so efficient [2, 8, 15, 16,21].

2.4 Datawarehousing

Big Data and/or BI allow the datawarehousing process to be carried out, with which different resources of different characteristics can be joined to analyze data, obtain information and be able to generate greater knowledge. These resources are usually internal to an organization or institution such as databases, Excel resources, pdf, etc., and external such as web links, social network forums, etc. [18].

One of the important threads is the ETL or Extraction, Transformation and Load and they are defined as follows: *Extraction.* - These are the tasks of extracting information from different data sources, web resources, etc. *Transformation*. - These are the tasks dedicated to the transformation, cleaning, elimination or treatment of null, ambiguous, incomplete, anomalous data, etc., according to the study area. For example, in the field of education, once the data has been correctly identified, it is not advisable to delete anomalous data, since precisely this type of data could give us important information about the students [1]. *Load.* - Data loading refers to the process of placing the purified or transformed data in a big datawarehouse or big database [6,7,18].

The type of data used for the Datawarehouse is structured and unstructured, the largest amount of information on the Internet is unstructured, all this information is processed in distributed type software architectures managed by software tools such as Hadoop, Spark, Power BI, etc. [18]. Power BI is a Microsoft tool that allows ETL processes to be carried out for data processing, it is connected to multiple sources of information such as databases, web resources, Excel files, etc.; on the other hand, it is connected to certain data science tools and to R and Python execution files, these languages being the most widely used for data science today [18,19].

3 Learning styles

3.1 Felder and Silverman model

As proposed by Felder and Silverman (1988), it classifies learning styles into four dimensions: *Understanding.* - Identifies the Global/Sequential concept, where the global student is governed neither by time nor by a calendar, he focuses on the objective that must be met, on the contrary, a sequential student is governed exclusively by time, his performance is presented using small logical steps to move quickly in the process. *Processing.* - Where the students are of the Active/Reflective type, the active ones are identified by manipulating things and working in a team, on the other hand, the reflective learns by thinking about things and likes to work alone. *Perception.* - Within this large group, the Sensing/Intuitive styles are identified, the sensitive student is the one who shows interest in the practical part while the intuitive student predominates in theory-oriented training. *Input.* - Here the Visual/Verbal styles are identified, where the visual style is related to the development of skills through the use of visual resources such as concept maps, statistical graphs, photos, videos, etc., on the other hand, a verbal student it is supported by the use of guided readings, active listening and note-taking, it learns from the discussion [3,6,22,23,24].

3.2 Development by competencies

The development by competences proposes a learning applied to real world situations, according to a study 8 elementary competences are proposed: Communication in mother tongue, communication in foreign languages, mathematical competence, science, technology and digital (CD), social (CS), learning to learn (CAAP), social and civic competences (CSC), sense of initiative and entrepreneurship (SIEE) and cultural awareness and expressions (CEC); as per their acronyms in Spanish. Competences are related to learning styles, an important relationship occurs between the competence of learning to learn and learning styles, it is important to know what the learning style of each person is to enhance their competence [25,26].

The development by competences is supported by e-learning technologies to improve the learning process. A competency-based assessment improves and personalizes the teacher's management while allowing a better characterization of the student's profile to be obtained [27].

4 Method

In this study, a methodology of qualitative approach, exploratory, documentary, bibliographic and bibliometric research was used to analyze and compare the main methods, techniques and computer mathematical models that allow processing data related to the area of education. To reach a study object, bibliometric analysis was used with the search equation given in Table 1.

Table 1. Search equation, bibliometric analysis.

Equation	Number of documents found in Scopus
"Learning capabilities "and "data science"	316

The bibliometric analysis was carried out in the Scopus database where the information was processed by the VosViewer software where after performing a concurrency analysis of the keywords in the articles referring to the defined search equation, the results shown were obtained. in Fig. 1.

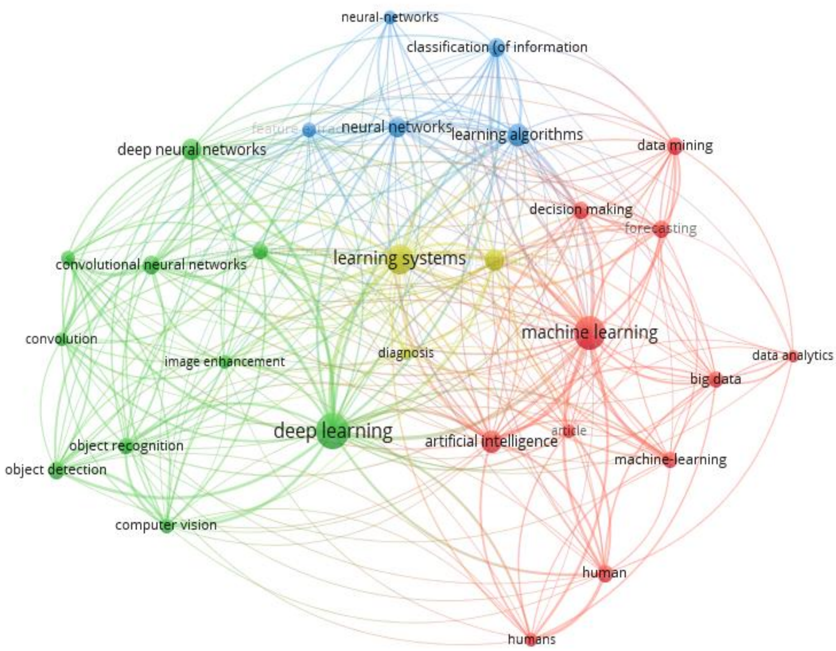

Fig. 1. Keyword Concurrency Analysis

In which the keywords are highlighted: “learning systems”, “machine learning”, “deep learning” and “artificial intelligence”. In addition to VOSviewer, the Bilbliometrix software was used to process the information from the documents found and shows that, in terms of the trends of the topics, it is relevant that they begin to develop since 2018, the concepts remaining constant over time. "machine learning" and "big data", in terms of the topics that have been most relevant during 2022 are "Deep learning", "object detection" and "computer vision", as shown in Fig. 2.

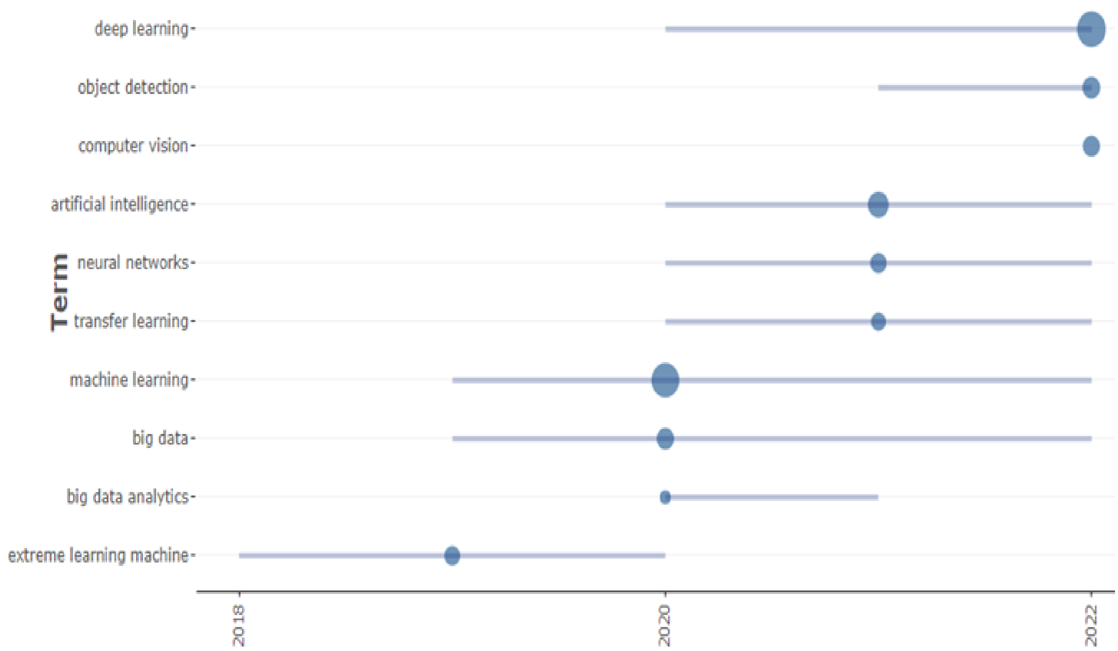

Fig. 2. Topic trends over time

The bibliographic analysis was carried out with searches for articles referring to the proposed theme, meetings and discussions between researchers, where the best methods to use both in the educational area and in the data science area were analyzed, taking into account that the subject of the educational data analysis with data science methods and Big data and/or BI technologies is relatively new [1].

5 Predictive models of learning styles and competencies

Table 2 shows the results of the analyzed models.

Table 2. Comparison of predictive models before and after completing an educational cycle based on the data to be processed to characterize the student profile.

Model or method	Before finishing an educational cycle	After completing an educational cycle
Clustering	Felder and Silverman test data.	Competency assessment data.
Decision tree	Felder and Silverman test data.	Competency assessment data.

Principal Component Analysis (PCA)	Felder and Silverman test data.	Competency assessment data.

To generate a new unknown knowledge or KDD Knowledge Discovery Databases in the area of education at the ISMAC Technological Institute in the students of the Software Development career, the machine learning and multivariate analysis models described in Table 2 are proposed, as follows:

Decision trees with the ID3 algorithm and its processing based on the gain of information, is ideal for characterizing the student profile from the point of view of learning styles and student competencies, since in each node there is an information maximization update or information gain through entropy minimization.

Clustering with the k-means algorithm to obtain group characteristics is efficient to characterize the student profile from the point of view of learning styles and student competencies to analyze the correlations between the different students and the classifying group they belong to, in the same way, verify the distinction between groups, this from a perspective that does not have predefined rules as it happens in decision trees.

The PCA or Principal Component Analysis is used to obtain a correlation between several variables of different kinds and reduce dimensions, which provides a focus on correlation and variability of the data to be treated, particularly the modeling of learning styles and student competencies with PCA to characterize the student profile is ideal since the model performs the correlation between variables based on variances and gives an approach that allows reducing variables to work with the most important ones. Note that the method with a variance approach does not rule out outliers of the sample or noise, which are important cases to analyze in the area of education.

Fig. 3. Keyword concurrency analysis (word cloud)

Fig. 2 and Fig. 3 show important results obtained in the bibliometric analysis; the use of machine learning is very present in data science studies. A study mentions the most used methods, among these very present clustering and decision trees [20], which are part of the proposed models, with a focus on education detailing specific algorithms such as ID3 for decision trees and k- means for clustering.

Table 3. Student data types.

Type of data	Description
Profile data	Sociodemographic and socioeconomic information on students.
Class data	Historical information of total and partial grades, subjects, attendance, etc., both historical and current.
Test data	Information on the student profile evaluated with psychometric tests.

Table 3 shows a classification of data to be processed with the machine learning and multivariate analysis models from different sources, both external and internal to the study group (Software Development students), as follows: Own or internal data of the study group. - Profile data and class data. External data of the study group. - Test data. It should be noted that class data is managed today by e-learning platforms such as Moodle [14] where there is very important stored information such as grades, student participation in discussion forums, access levels and test completion time. and tasks, all this information is stored in logs, database tables, web resources, etc., in e-learning platforms [27].

Table 2 proposes the machine learning and multivariate analysis methods to process the data according to the classification in Table 3, with the aim of carrying out a KDD process to obtain new unknown knowledge with the identification of behavior patterns, similarity and anomalous through a predictive process to identify the student learning styles of the students of the Software Development career of the ISMAC Institute before or during their student life within the institution, this will allow to personalize the student's learning and improve their skills and abilities.

In addition, it is proposed to implement this data infrastructure, processing with machine learning models and multivariate analysis and generation of knowledge in the second stage of the project "Learning capacities of students in the Software Development career of the face-to-face modality, using tools of Big Data and/or BI", hoping to obtain promising results that provide us with new knowledge of the student data from the sample of the aforementioned institution, which allows establishing decision-making to improve student learning.

6 Conclusions

The most efficient data science methods for data processing of student processes and the prediction of learning styles in Software Development technology careers are decision trees with the ID3 algorithm, clustering with the k-means algorithm and the analysis of principal components or ACP.

The EDM or Educational Data Mining is a relatively new area that unites data science, Big data technology and education to improve learning.

The bibliometric analysis shows us a notable presence of machine learning techniques within data analysis with Big Data technologies.

The classifications of the data of the study group of the Software Development technology careers of the ISMAC Institute, were defined as: *Profile data.* - That correspond to sociodemographic and socioeconomic information. *Class data.* - Which correspond to historical information from past and current academic periods managed by e-learning platforms such as Moodle and Test data. - Corresponding to the Felder and Silverman learning styles tests.

The definition of a software infrastructure for datawarehousing with Power BI tools and Python language has been chosen to implement the different computer mathematical models in current technologies and customize the KDD or knowledge discovery process.

References

1. Romero, C., Ventura, S., Pechenizkiy, M.: Handbook of Educational Data Mining. CRC Press, Boca Ratón (2011).
2. Sarker, I.H.: Machine Learning: Algorithms, Real-World Applications and Research Directions. SN COMPUT. SCI. 2(160), (2021).
3. Costaguta, R.: Identificación de estilos de aprendizaje dominantes en estudiantes de informática. XV Congreso Argentino de Ciencias de la Computación, pp. 331-340 (2009).
4. Porcel, E., Dapozo, G.: Modelos Predictivos y técnicas de minería de datos para la identificación de factores asociados al rendimiento académico de alumnos universitarios. XI Workshop de Investigadores en Ciencias de la Computación, pp. 635-639 (2009).
5. Ayala, E., López, R.: Modelos predictivos de riesgo académico en carreras de computación con minería de datos educativa. Revista de Educación a Distancia (RED). 21(66), (2021).
6. Caman, R., Salguero, A.: Herramientas para detección de estilos de aprendizaje en estudiantes de educación superior. Revista Tecnológica ESPOL – RTE 30(3), 106-121 (2017).
7. Ballesteros, A., Sánchez, D.: Minería de datos educativa: Una herramienta para la investigación de patrones de aprendizaje sobre un contexto educativo. Latin-Amercian Journal of Physics Education 7(4), (2013).

8. Dabhade, P., Agarwal, R.: Educational data mining for predicting student's academic performace using machine learning algorithms. Materials Today: Proceedings 47(15), 5260-5267 (2021).
9. Costa, E., Fonseca, B.: Evaluating the effectiveness of educational data mining techniques for early prediction of students' academic failure in introductory programming courses. Computers in Human Behavior 73(s.n.), 247-256 (2017).
10. Agarwal, S., Pandey, G. N., & Tiwari, M. D.: Data mining in education: data classification and decision tree approach. International Journal of e-Education, e-Business, e-Management and e-Learning, 2(2), 140 (2012).
11. Patel, H. H., & Prajapati, P.: Study and analysis of decision tree based classification algorithms. International Journal of Computer Sciences and Engineering, 6(10), 74-78 (2018).
12. Todhunter, F.: Using principal components analysis to explore competence and confidence in student nurses as users of information and communication technologies. Nursing Open, 2(2), 72–84 (2015).
13. Dumpit, D.Z., Fernandez, C.J.: Analysis of the use of social media in Higher Education Institutions (HEIs) using the Technology Acceptance Model. Int J Educ Technol High Educ 14, 5 (2017).
14. Romero, C., Ventura, S., Espejo, P. G., & Hervás, C.: Data mining algorithms to classify students. In Educational data mining 2008 (2008).
15. Dubois, J., Galdi, P., Paul, L. K., & Adolphs, R.: A distributed brain network predicts general intelligence from resting-state human neuroimaging data. Philosophical Transactions of the Royal Society B: Biological Sciences, 373(1756), 20170284 (2018).
16. Freiberg-Hoffmann, A., & Fernández-Liporace, M.: Learning Styles Type Indicator: una adaptación para estudiantes medios y universitarios argentinos. Academo (Asunción), 10(1), 1-12 (2023).
17. Alcalá-Fdez, J., Sanchez, L., Garcia, S., del Jesus, M. J., Ventura, S., Garrell, J. M., Herrera, F.: KEEL: a software tool to assess evolutionary algorithms for data mining problems. Soft Computing, 13, 307-318 (2009).
18. Santoso, L. W.: Data warehouse with big data technology for higher education. Procedia Computer Science, 124, 93-99 (2017).
19. Vásquez, R. A. D., Espinoza, J. L. A., & Cabrera, M. A. C.: Power bi como herramienta de apoyo a la toma de decisiones. Universidad y Sociedad, 14(S3), 195-207 (2022).
20. Liao, S.-H., Chu, P.-H., & Hsiao, P.-Y.: Data mining techniques and applications – A decade review from 2000 to 2011. Expert Systems with Applications, 39(12), 11303–11311 (2012).
21. Closas, A. H., Arriola, E. A., Kuc Zening, C. I., Amarilla, M. R., Jovanovich, E. C.: Análisis multivariante, conceptos y aplicaciones en Psicología Educativa y Psicometría. Enfoques, 25(1), 65-92 (2013).
22. Figueroa, N., Cataldi, Z., Méndez, P., Rendon Zander, J., Costa, G., Lage, F. J.: Los estilos de aprendizaje y las inteligencias múltiples en cursos iniciales de programación. In VII Workshop de Investigadores en Ciencias de la Computación (2005).
23. Nivela-Cornejo, M. A., Echeverría-Desiderio, S. V., Morillo, R.: Estilos de aprendizaje y rendimiento académico en el contexto universitario. Domino de las Ciencias, 5(1), 70-104 (2019).
24. Roa, K., Martínez, C.: Diseño de un ambiente virtual de aprendizaje soportado en los estilos de aprendizaje. Virtu@ lmente, 8(2), 68-86 (2020).
25. Marcos, B., Alarcón, V., Serrano, N., Cuetos, M., Manzanal, A.: Aplicación de los estilos de aprendizaje según el modelo de Felder y Silverman para el desarrollo de competencias clave en la práctica docente. Tendencias pedagógicas (2020).

26. González, B., Hernández, M., Castrejón, V.: Estilos de aprendizaje para el desarrollo de competencias en estudiantes de la Licenciatura en Enfermería. RIDE. Revista Iberoamericana para la Investigación y el Desarrollo Educativo, 8(16), 351-369 (2018).
27. Moodle, https://moodle.org/, last accessed 2023/06/05.