Characterization of high school students in the department of Risaralda using the Chi-Square metric

Caracterización de los estudiantes de educación media en el departamento de Risaralda usando la métrica Chi-Square

L. A. Serna-Cardona (iD) ; K. A. Hernández-Gómez (iD) ; A. A. Orozco-Gutiérrez (iD)

Abstract— The education is a fundamental law that promotes the social and economic development of a nation. Also, it is how a country can achieve its sustainable development objectives. Because of this, various international organizations (Unicef-UNESCO- Banco Mundial -OECD) are promoting the coverage of education in underdeveloped countries. On the other hand, projects such as PISA that aim to carry out annual evaluations in more than 30 countries that serve as a reference of the educational level and encourages nations to improve the quality of education. Therefore, the objective of this work is to implement a methodology to predict the results of ICFES tests SABER 11 °. that can correct the educational problems that Colombia has presented in secondary education. For the development of this methodology, the ICFES repository database was used. Data were pre-processed using MATLAB software. Also, combined tests were carried out with the new Chi-Square metric developed by the researchers of the automation group of the Technological University of Pereira, obtaining results that were 20% higher compared with conventional classification techniques. In this project, the most influential characteristics were found in the students, who are responsible for the loss of the ICFES SABER 11° exam in Risaralda.

Index Terms — Categorial Data, Chi-Square, Education, ICFES, Machine Learning, Relevance Analysis.

Resumen— La educación es un derecho fundamental que promueve el desarrollo social y económico de una nación, La Educación es un derecho fundamental que promueve el desarrollo social y económico de una nación, además, es el medio por el cual un país puede alcanzar sus objetivos de desarrollo sustentable. A causa de esto, diversos organismos internacionales (Unicef-UNESCO-Banco Mundial-OCDE) están promoviendo la cobertura de la educación en los países subdesarrollados. Por otro lado, proyectos como PISA cuyo objetivo es realizar evaluaciones anuales a más de 30 países que sirve como referencia del nivel educativo e impulsa a las naciones a mejorar la calidad de la educación. Por ende, El objetivo de este trabajo es implementar una metodología para predecir los resultados de las pruebas ICFES saber 11°. que pueda subsanar la problemática educacional que viene presentado Colombia en la educación media. Para la elaboración de esta metodología, se utilizó la base de datos del ICFES saber 11°. Se preprocesaron los datos utilizando el software MATLAB. Además, se hicieron pruebas combinadas con la nueva métrica Chi-Square desarrollada por los investigadores del grupo de automática de la Universidad tecnológica de Pereira, obteniendo resultados superiores en un 20% comparados con las técnicas convencionales de clasificación. En este proyecto, se encontraron las características más influyentes en los estudiantes, las cuales son causantes de la pérdida del examen ICFES saber 11° en Risaralda.

Palabras claves— Chi-Square, Datos categóricos, Educación, ICFES, Machine Learning, Análisis de Relevancia.

I. INTRODUCTION

THE last century, education has become the benchmark for countries to determine the economic, social, and cultural level of each nation. Well, the quality-of-life index is strongly linked to the educational level by country [1,2]. According to research, the countries that invested the most in education in the last three decades grew exponentially and today they are the richest countries in the world and with a better quality of life than the rest, for example, USA, CHINA, CANADA, etc. [3,4]. Demonstrating that the educational level of each country is one of the fundamental weapons for its development. Different studies have shown that the contribution of education to a country linked to any area of it, from the political, educational, economic, military, commercial, technological, etc. Well, through educational investment is

that students with a very high intellectual level emerge, including researchers in charge of technological or military advances, for example, USA, GERMANY, JAPAN, etc. [5] For this reason, Colombia has sought to improve the education of its country, increasing the budget for the Ministry of Education, among other ways to increase the educational development of students [6]. As part of the solution to this problem, the ministry has been conducting tests with the databases of the state exam ICFES SABER 11°, to find out which are the performance factors that most affect the performance of students and schools. Different investigations have been carried out on the subject, having good results but still not solving all the problems present in the country [7-10]. Well, there is still a long way to go in machine learning, which will be able to provide the country and the secondary education institutions with tools that will allow them to have better control over the educational level of their students and carry out the pertinent actions that the case merits.

Previous research shows that the socioeconomic and sociodemographic variables of students are a factor of great importance in education and directly related to academic performance [7-10], variables such as occupation or level Parents' education [11], and socioeconomic status or possession of some capital, are usually linked to student performance [12]. For this reason, these social variables allow a better approach to the student in the educational field, thus allowing the implementation of some measure that can provide a solution to young people who will represent future problems with state tests, allowing the government an improvement in the educational level of the nation. This, because the state exam becomes an indicator of the intellectual capacities of those tested, also, it is a fundamental requirement for the continuity of the academic training of citizens; it becomes decisive to identify the sociodemographic variables that influence directly or indirectly on the final results of the SABER 11° tests since, in the long run, it will become a factor that intervenes both in the quality of life of Colombians and in the country's development.

For this reason, they have carried different investigations out about the aforementioned problem. As, for example, the ministry of education, that using the SABER 11° test database, performs statistical analyzes to determine relationships between sociodemographic variables and the test results [13][20]. However, the analyzes are very superficial and not very rigorous since it limits them to the use of descriptive statistics and do not allow finding correlations of underlying variables that can be expressed in terms of the original ones. On the other hand, the Francisco José de Caldas District University set out to find in its work the individual determinants of performance in state tests, and for this, it carried out a comparative analysis of two regression techniques, multilevel regression model (HLM) and learning Kernel-based computational (SVM), in which there was a notable superiority by the HLM model both in its expressiveness and in the processing time. However, the methods used did not allow the individual relationships generated to described, and the results got were not as

expected [14][21].

Data mining is not a new topic within the educational field, and its techniques can explain and predict phenomena within the educational system [15,16]. An example of this is the use of machine learning techniques to predict the probability of attrition with high reliability [17]. On the other hand, educational institutions make use of these techniques to perform characterizations of the academic profiles of their students [18], because of this it is convenient to use the computational tools that mining provides for data.

Therefore, the principal aim of this article was to develop a tool that can select the most relevant characteristics at the time of losing the ICFES status exam. To later classify the users who will lose or win the tests to know 11th. This, based on the performance patterns found in the sociodemographic and socioeconomic variables and using machine learning techniques. Specifically, an approach based on supervised and unsupervised learning proposed, to test the advantages and difficulties of each method, which will finally be able to contribute to the research that the government has been carrying out on student performance factors [13]. But it will also contribute to this alternative path in which Colombia should gradually take hold, which is machine learning, specifically in software that can predict or classify these students.

II. METHODS

A. Chi-Square Distance

The chi-square distance is similar to the Euclidean. However, it is a weighted distance and a suitable metric for the analysis of databases with qualitative, categorical or nominal variables. The Chi-square distance compares the counts of responses from categorical variables with two or more independent features:

$$d_{ij} = \sqrt{\sum_{n=1}^{D} \frac{1}{\tilde{x}_n} \left(\tilde{x}_{in} - \tilde{x}_{jn} \right)^2}$$

Where

$$\tilde{x}_{in} = \frac{x_{in}}{\sum_{n=1}^{D} x_{in}}$$

$$\tilde{x}_n = \frac{1}{D} \sum_{n=1}^{D} x_{in}$$

Here, D is the number of features or dimensions. The Chi-

Square distance uses a contingency table, with the frequency of each attribute. The weighted distance C-S with categorical features allows a better treatment of these data. This is explained because it improves the separability of the classes, and allows an easier grouping or discrimination. However, an important drawback is the augment of dimensionality due to the data mapping to a space of dissimilarity. Therefore, it is necessary to use the algorithm t-SNE for reducing the dimensionality to 2 or 3 attributes. To preserve the structure of the databases, it was implemented the C-S metric within the distance function of t-SNE for simultaneously enhancing separability of categorical data and reducing computational times in learning algorithms [19][20][21].

B. Support Vector Machines (SVMs)

Support vector machines (SVMs) are prevalent in applications such as natural language processing, speech, image recognition and artificial vision. The full theory of SVMs can be found in [22]. This approach can be divided as follows:

- *Separation of classes*: It is about finding the optimal separating hyperplane between the two classes by maximizing the margin between the closest points of the classes.

- *Overlapping classes:* The incorrect data points of the discriminating margin are weighted to reduce their influence (soft margin).

- *Non-linearity:* When a linear separator cannot be found, the points are mapped to another dimensional space where the data can be separated linearly (this projection is realized via kernel techniques).

- *Solution of the problem:* The whole task can be formulated as a quadratic optimization problem that can be solved by known methods.

SVMs belong to a class of machine learning algorithms called kernel methods. Common kernels used in SVMs include: RBG or Gaussian, linear, polynomial, sigmoidal, among others [23]. We choose the RBG function due to its flexibility for different type of data. We set the Gamma and C hyper-parameters of the RBF kernel through cross-validation.

III. DATASET AND EXPERIMENTAL SETUP

A. Dataset

We took the databases that are being worked on from the official ICFES website https://www.icfes.gov.co/investigadores-y-estudiantes-posgrado/acceso-a-bases-de-datos. Table 1 shows the pertinent information. It is important and pertinent to clarify that the information from the surveys that taken from the Ministry of Education and used during the execution of this project, not disclosed, it was completely confidential.

We carried a careful and refined review of the state-of-the art related to Colombian education out, as well as regarding machine learning applied to the recognition of student profiles based on information from the survey got by the Ministry of Education. Subsequently, We label and organize the information collected in the surveys carried out on Colombian students in the saber 11 tests.

B. Training and Specifications

The students are divided into those who won and those who lost the state test. Then, we a respective preprocessing implemented to eliminate missing values among other inconveniences, and the results got used for training the pattern recognition algorithms.

Also, we a module developed for the recognition of student profiles based on unsupervised learning. Specifically, we validated the methods applied in the state-of-the art, such as the k-means algorithm and Gaussian mixtures with Bayesian inference: Expectation-Maximization and variational inference. We validated these proposed unsupervised learning methods using performance metrics such as precision, dispersion, ROC curves, specificity, and sensitivity on student identification results. Finally, we a supervised learning algorithm developed to predict the outcome of the ICFES Exam. As future work, we intend it to develop an application in Matlab or Python that articulates the machine learning modules.

We calculate the accuracy (AC) and computational times for all classifiers in each setup, under the same conditions. We perform a hold-out validation scheme, with ten repetitions for each experiment, taking 70% of the data for training and 30% for validation. The simulations were performed with Matlab software on a server Intel (R) Xeon (R), CPU E5-2650 v2-2.60~GHz, two processors with eight cores, and 280 GB-RAM.

TABLE I
DATASETS ICFES

Exam	Samples	Feature	Total Samples
Mathematics	395870	42	1583480
Social Sciences	395870	42	1583480
Natural Sciences	395870	42	1187610
Spanish	395870	42	1583480
English	395870	42	1583480

Table 1: Shows the relevant information from the database. It has 42 categorical characteristics. In addition, the algorithm processed 7.5 million data in total.

Fig. 1. Graph of the discrete database in 3 dimensions.

IV. RESULT AND DISCUSSION

In this work, we used the ICFES SABER 11 ° database, a categorical database that coded using the Matlab software, transforming it into a numerical-discrete database as shown in Fig. 1. In this, we can observe the overlap of the data, and the null visual separation in 3 dimensions, the categorical data not only represents this problem but also causes a low level of success. The first stage to remedy this problem is proper database preprocessing. For this, the Matlab functions were used to eliminate missing values and outliers as seen in Fig. 2. Doing a correct preprocessing allows the data to be visually separated. However, the problem has not been solved, the characteristics that have a very high correlation with each other must eliminated, so it improves both visually and at the level of success. For this, the three conventional methods for dimensionality reduction (CORR, PCA, RELIEF) used.

Fig. 3 shows us the total correlation of the database where it can observe that the last characteristics have a very high correlation because they result from the ICFES exam in mathematics, social sciences, natural sciences, Spanish and English. Therefore, these characteristics eliminated, getting cleanliness of 42 characteristics to 27 as observed in Fig. 4. After thresholding with the first dimensionality reduction method, Fig. 5 shows us a decrease from 27 to 17 characteristics. On the other hand, Fig. 6 shows us one of the most used methods in machine learning to reduce dimensionality called PCA. With this method, we can reduce the dimensionality from 27 to 17 features.
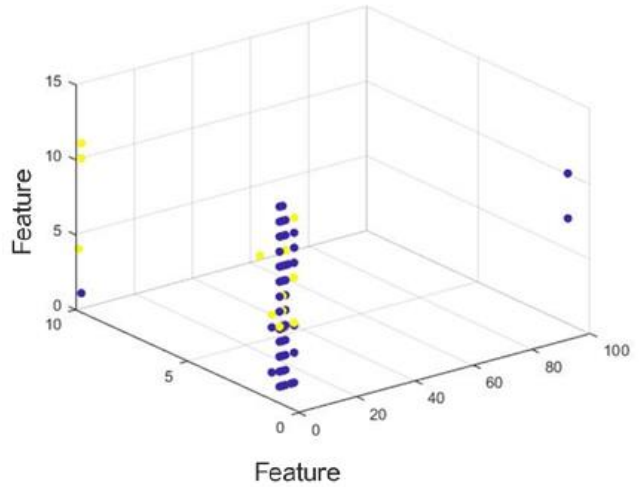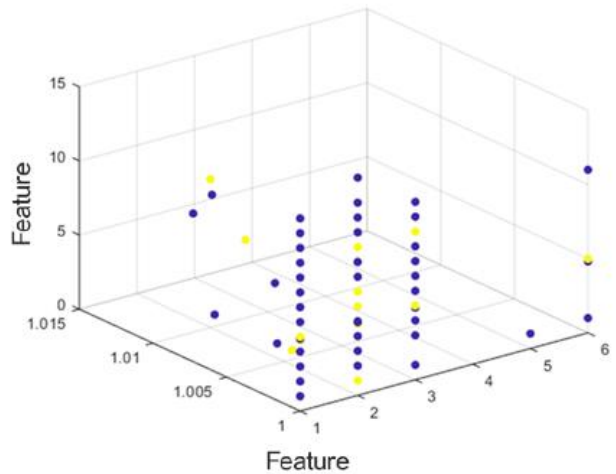

Fig. 2. Graph of the discrete database without Missing Values or Outliers in 3 dimensions.
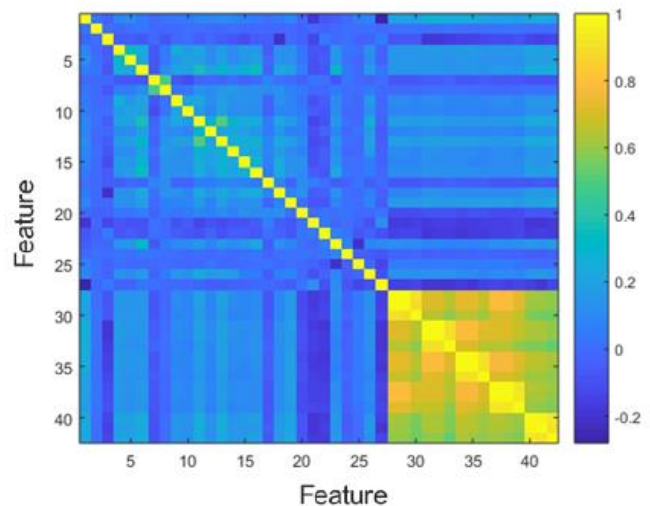

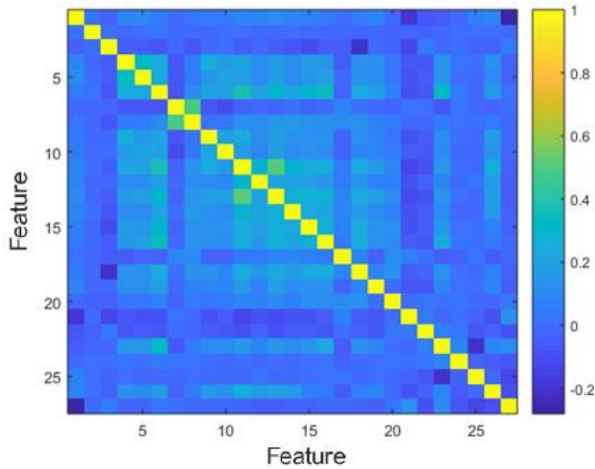Fig. 3. Correlation matrix of the preprocessed database.

Fig. 4. Correlation matrix of the preprocessed database without the characteristics that cause a total correlation from 27 to 42.
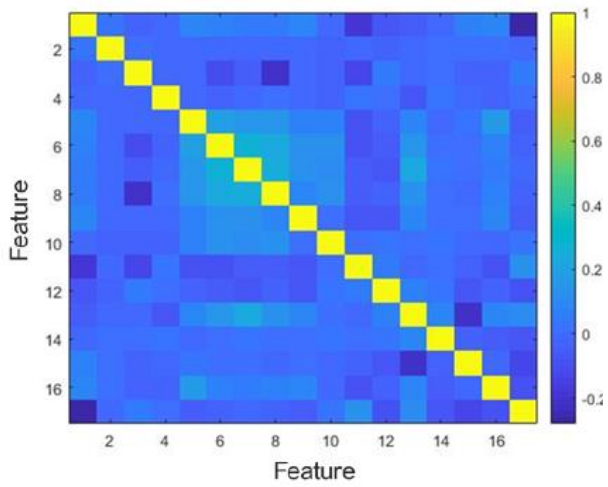

Fig. 5. Correlation matrix of the preprocessed database using the first dimensionality reduction method.
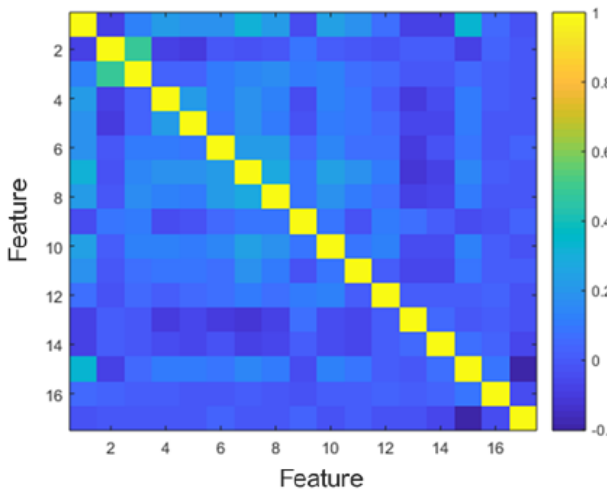

Fig. 6. Correlation matrix of the preprocessed database using the PCA dimensionality reduction method.

Therefore, the latest RELIEFF dimensionality reduction method gives us much better results than the previously exposed methods. The reduction goes from 27 to 11 features as shown in Fig. 7. Also, by reducing the dimensions that highly correlated, the database improved visually as expected as seen in Fig. 8. However, the separability of the database is still not optimal, this is because of the categorical data type. For this reason, a new metric developed by the Automatics group called Chi-Square distance (C-S) used.
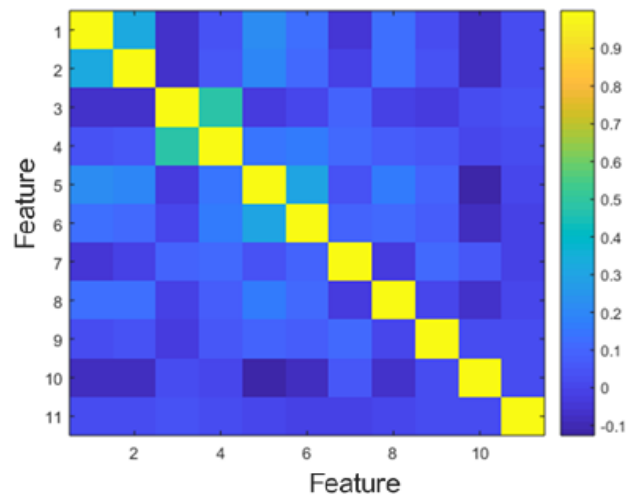

Fig. 7. Correlation matrix of the preprocessed database using the RELIEFF dimensionality reduction method.
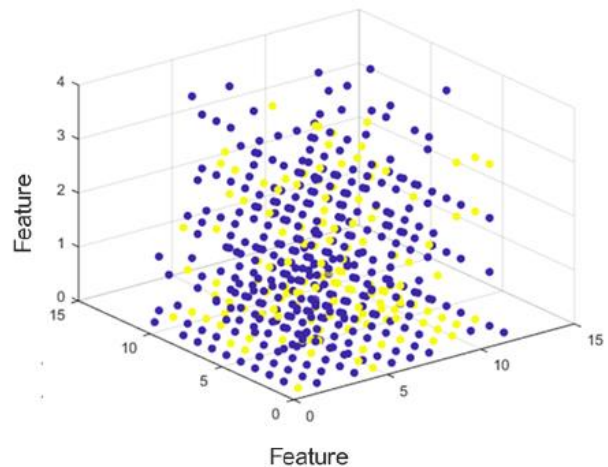

Fig. 8. Graph of the preprocessed database after reducing dimensionality with RELIEF.

This distance maps the qualitative data to a quantitative space, increasing its dimensionality and allowing a clear separability as shown in Fig. 9. With this distance not only is such a level of separability achieved, but it also increases the accuracy of the algorithms of classification as seen in Table 2. When mapping the data with CS and later using these data as input in a linear SVM classifier, it allows an increase of approximately 20% that using the data with no type of mapping as observed in Table 2.
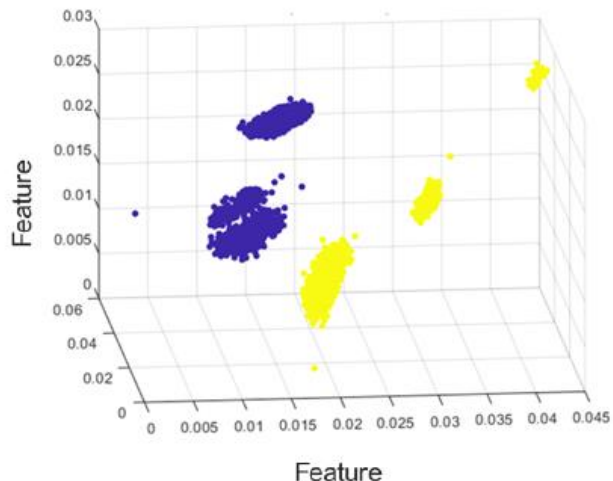

Fig. 9. Graph of the preprocessed database mapped with the Chi-square to another dissimilarity space.

TABLE II
DATASETS ICFES

Method	Classifier	Accuracy	Variance
Database Relieff	SVM Linear	69.59	1583480
Database C-S	SVM Linear	88.51	1583480
Database PCA	SVM Linear	67.15	1187610
Database CORR	SVM Linear	63.34	1583480

Table 2: Results of the accuracy in the prediction of the linear SVM algorithm, using the database with RELIEF, Chi-Square (C-S), PCA, CORR.

Finally, to determine which were the characteristics that most affected the students in the department of Risaralda, we used a relevance analysis between RELIEFF and C-S. As seen in Fig. 10, there are 11 characteristics that most harm Risaraldenses students and which explained and with their respective relevance in Table 3. The principal aim of this article was to determine with high precision the characteristics that most affected secondary education in Colombia.
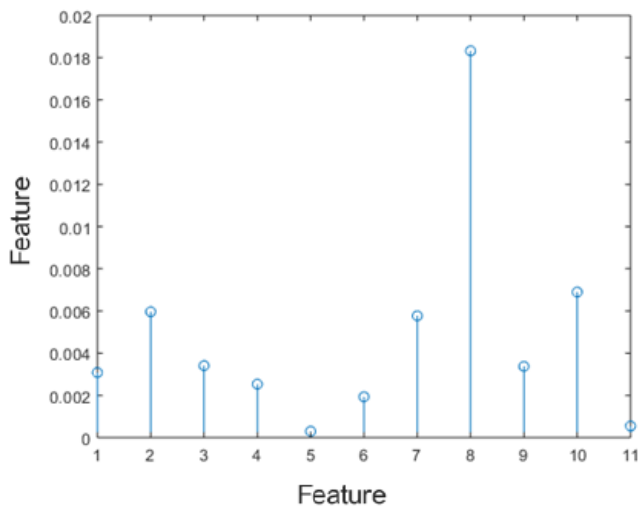

Fig. 10. Most significant values using the RELIEF C-S dimensionality reduction method.

TABLE III
DATASETS ICFES

Most influential features in Risaralda.	Relevance
Father's level of education.	3.1
Father's level of education.	6.0
How many people are in the home?	3.4
How many rooms does the house have?	3.0
They have internet at home.	1.9
They have a vehicle at home.	5.8
Book number in the home.	3.4
Stratum of the house.	2.5
Economic situation of the family.	18.3
How many hours does the student work?	6.9
Bilingual school.	0.5

Table 3: Relevance of the most influential characteristics in the department of Risaralda

Within the literature there are documents that deal with the problems of knowing 11 ° tests [7][8][13][14]. However, only descriptive statistics are used to provide solutions and advances in research related to Colombian education. An attempt to use data science was that of [15], however, when

replicating the tests, their results were more efficient using descriptive statistics, which showed an inefficient handling of kernels in SVMs. This makes our work a unique and stable approach in the Colombian country, with many possibilities of extrapolating it to all national departments to predict and improve the educational levels of Colombian students.

V. CONCLUSIONS AND FUTURE WORK

As partial conclusions, the efficiency of the Chi-square distance for categorical data stands out, as it shows great efficiency with grouping data, classification, etc., as we can see in Table II. Well, this dissimilarity measure allows mapping data from a qualitative space to a quantitative one, allowing optimal visual separability, see Figure 9. As part of the development of the work, a great deficiency also found in the department of la guajira in the tests to SABER 11 °, it should note that it is the department with the worst results and the most influential characteristics are the socioeconomic and demographic factor as expected. The results of the department of Risaralda show social and economic deficiencies, see Figure 10 and Table III, we can also show that it has deficiencies regarding Antioquia, Bogotá, etc. In the future, with this data it expected to use more robust supervised learning techniques to predict the students who will win or lose the state tests and later embed it in software that allows the government to identify these factors and provide an adequate solution.

REFERENCES

[1] "Support Sustainable Development and Climate Action | United Nations." https://www.un.org/en/our-work/support-sustainable-development-and-climate-action (accessed Jul. 03, 2021).
[2] A. Maldonado-maldonado, C. Rodríguez-sabiote, and I. Introducción, "Los organismos internacionales y la educación en México. El caso de la educación superior y el Banco Mundial," Perfiles Educ., vol. 22, no. 87, pp. 51–75, 2000.
[3] J. M. Redondo, "EL EXPERIMENTO CHILENO EN EDUCACIÓN: ¿Conduce a mayor equidad y calidad en la educación?," 2005. DOI: 10.4067/s0718-22362005000100005
[4] Banco Mundial, "Gasto público en educación, total (% del PIB) | Data," Instituto de Estadística de la Organización de las Naciones Unidas para la Educación, la Ciencia y la Cultura (UNESCO), 2018. https://datos.bancomundial.org/indicador/SE.XPD.TOTL.GD.ZS?end=2019&most_recent_value_desc=true&start=2019&view=map%0Ahttps://datos.bancomundial.org/indicator/SE.XPD.TOTL.GD.ZS?end=2019&start=1970%0Ahttps://datos.bancomundial.org/indicador/SE.XPD.TOTL.GD. (accessed Jul. 08, 2021).
[5] R. Turner, "El Programa Internacional para la Evaluación de los Alumnos (PISA). Una perspectiva general," Rev. Educ., vol. 2006, pp. 45–74, 2006.
[6] OECD, "Education in Colombia, Reviews of National Policies for Education." OECD Publishing Paris, 2016. DOI: https://doi.org/10.1787/9789264250604-en
[7] S. Acevedo and A. Jaramillo, "Perfil socioeconómico de los estudiantes de pregrado EAFIT." Recuperado el, 2007.
[8] N. G. Armenta, C. C. Pacheco, and E. D. Pineda, "Factores socioeconómicos que intervienen en el desempeño académico de los estudiantes universitarios de la Facultad de Ciencias Humanas de la Universidad Autónoma de Baja California," Rev. Investig. en Psicol., vol. 11, no. 1, p. 153, 2014. DOI: 10.15381/rinvp.v11i1.3888
[9] J. Caso Niebla and L. Hernández Guzman, "Available in: http://www.redalyc.org/articulo.oa?id=80539304," Rev. Latinoam. Psicol., pp. 487–501, 2007.
[10] K. Contreras, C. Caballero, J. Palacio, and A. M. Pérez, "Factores asociados al fracaso académico en estudiantes universitarios de Barranquilla (Colombia)," Psicol. desde el Caribe, vol. 0, no. 0, pp. 110–135, 2008.
[11] G. M. G. Vargas and F. R. Estrada, "Desafíos y problemáticas de la gestión universitaria en el sistema de la educación superior pública, aspectos a considerar en la región centroamericana," Rev. Gestão Univ. na América Lat. - GUAL, pp. 231–258, 2013, doi: 10.5007/1983-4535.2013v6n2p231.
[12] E. H. Padilla and M. J. G. Montesinos, "Modelo de ecuación estructural que evalúa las relaciones entre el estatus cultural y económico del estudiante y el logro educativo," Rev. Electron. Investig. Educ., vol. 13, no. 2, pp. 188–203, 2011.
[13] ICFES, "Acceso a Bases de datos y diccionarios - Portal Icfes," 2020. https://www.icfes.gov.co/investigadores-y-estudiantes-posgrado/acceso-a-bases-de-datos (accessed Jul. 08, 2021).
[14] L. Ariosto et al., "Classification of Categorical Data Based on the Chi-Square Dissimilarity and t-SNE," Computation, 2020. doi: 10.3390/computation8040104
[15] A. Martínez, Un análisis del efecto de pares sobre el rendimiento académico para Colombia. 2012.
[16] R. Aguilar, H. D. Vargas-Cardona, A. M. Álvarez, Á. A. Orozco, and P. Navarro, "Automatic Identification of DBS Parameters from the Volume of Tissue Activated (VTA) Using Support Vector Machines," in Iberoamerican Congress on Pattern Recognition, 2018, pp. 740–747.
[17] R. T. PEREIRA2, - ANDRÉS CALDERÓN ROMERO3, and J. J. TOLEDO4, "Aplicación de la minería de datos HQ OD H [WUDFFLyQ GH SHUÀOHV de deserción estudiantil * 1 [Application of data mining in Resumen Introducción," Vent. Informática, no. 28, pp. 31–47, 2013, [Online].
[18] R. T. Pereira, A. C. Romero, and J. J. Toledo, "La Minería De Datos Como Un Método Innovador Para La Detección De Patrones De Deserción Estudiantil En Programas De Pregrado En Instituciones De Educación Superior," Asoc. Colomb. Fac. Ing. (ACOFI), Int. Fed. Eng. Educ. Soc. (IFEES), p. 9, 2013, [Online]. Available: https://www.acofipapers.org/index.php/acofipapers/2013/paper/viewFile/211/112.
[19] A. Kamagate, "Análisis de la deserción estudiantil en la Universidad Simón Bolívar, facultad Ingeniería de Sistemas, con técnicas de minería de datos," Rev. Pensam. Am., vol. 6, no. 10, pp. 41–51, 2013.
[20] S. Valero, A. Salvador, and M. García, "Minería de datos: predicción de la deserción escolar mediante el algoritmo de árboles de decisión y el algoritmo de los k vecinos más cercanos," Recur. Digit. para la Educ. y la Cult., pp. 33–39, 2010.
[21] L. A. Serna, K. A. Hernández, and P. N. González, "A K-Means Clustering Algorithm: Using the Chi-Square as a Distance," in International Conference on Human Centered Computing, pp. 464–470, 2018. Doi: https://doi.org/10.1007/978-3-030-15127-0_46
[22] C. Cortes and V. Vapnik, "Support-vector networks," Mach. Learn., vol. 20, no. 3, pp. 273–297, 1995. DOI: https://doi.org/10.1007/BF00994018
[23] M. Hu, Y. Chen, and J. T. Y. Kwok, "Building sparse multiple-kernel SVM classifiers," IEEE Trans. Neural Networks, vol. 20, no. 5, pp. 827–839, 2009. DOI: 10.1109/TNN.2009.2014229

Luis Ariosto Serna Cardona received his undergraduate degree in physical engineering (2017) and his M.Sc. degree in engineering (2021) from the Universidad Tecnológica de Pereira. He is director of research at CIAF educación superior and researcher in the department of engineering. Research interests: machine learning and deep learning.
ORCID: https://orcid.org/0000-0003-3985-4014

Kevin Alejandro Hernández Gómez received his undergraduate degree in physical engineering (2017) and his M.Sc. degree in engineering (2021) from the Universidad Tecnológica de Pereira. He is a research professorat CIAF educación superior and researcher in the department of engineering. Research interests: machine learning and signal processing.
ORCID: https://orcid.org/0000-0003-4506-9004

Álvaro Orozco-Gutierrez received his undergraduate degree in electrical engineering (1985) and his M.Sc. degree in engineering (2004) from the Universidad Tecnológica de Pereira, and his Ph.D. in bioengineering (2009) from the Universidad Politécnica de Valencia (Spain). He received his undergraduate degree in law (1996) from Universidad Libre de Colombia. Currently, he is a Professor in the Department of Electrical Engineering at the Universidad Tecnológica de Pereira. Research interests: machine learning.
ORCID: http://orcid.org/0000-0002-1167-1446