Contents lists available at Science-Gate

International Journal of Advanced and Applied Sciences

Journal homepage: http://www.science-gate.com/IJAAS.html

Identifying low-performing regions in Moroccan education: A deep learning approach using the PISA dataset

Ilyas Tammouch [1,*], Abdelamine Elouafi [1], Souad Eddarouich [2], Raja Touahni [1]

[1]*Telecommunications Systems and Decision Engineering Laboratory, Faculty of Science, Ibn Tofail University, Kenitra, Morocco*
[2]*Regional Educational Center, Rabat, Morocco*

ARTICLE INFO

Article history:
Received 22 January 2023
Received in revised form
17 May 2023
Accepted 22 May 2023

Keywords:
School reform movement
Academic performance
Regional performance inequalities
Clustering algorithms
Deep embedding clustering

ABSTRACT

This study highlights the ongoing nature of the school reform movement, emphasizing the need for continuous attention and action. Despite this effort, academic performance has exhibited relative stability in recent years, while significant regional performance disparities persist. Addressing these inequalities requires novel approaches to enhance educational quality. Past research has explored clustering algorithms in developed countries, providing insights into personalized teaching strategies based on students' learning style preferences. In response, our research aims to identify underperforming regions in Morocco, necessitating attention and intervention. We employ an unsupervised deep learning method called "deep embedding clustering" to group Moroccan students based on their performance. The results are subsequently visualized on a choropleth map, revealing intricate patterns and trends in educational performance that might not be immediately apparent. The analysis employs the comprehensive program for international student assessment (PISA) dataset, encompassing individual students' responses and plausible values reflecting cognitive abilities. The findings indicate that the "Guelmim-Oued Noun" region exhibits the highest performance level among all regions, while "Dakhla-Oued Eddahab," "Béni Mellal-Khénifra," and "Oriental" regions display lower performance levels. As a result, this study urges policymakers to incorporate tailored measures into regional policies to improve students' educational outcomes.

1. Introduction

The advent of the information and technology era has resulted in the accumulation of vast volumes of data. Converting this acquired data into actionable information is essential for making informed decisions. In the context of higher education, performance analysis holds particular significance, as student performance plays a crucial role in shaping their employment opportunities. Utilizing clustering algorithms provides an effective means to identify key attributes that may serve as barriers to accurate student assessment.

Shovon and Haque (2012) defined data clustering as an unsupervised statistical data processing technique aimed at grouping similar data into homogeneous clusters to unveil hidden patterns and relationships, facilitating faster and more efficient decision-making. Through cluster analysis, large datasets are partitioned into smaller groups, referred to as clusters, where each cluster comprises data objects that share similarities while differing from objects in other clusters. Employing clustering techniques aimed at identifying low-performing regions can assist educators in enhancing the quality of education.

Several studies have explored student performance and classification methods. Pasina et al. (2019) and Lailiyah et al. (2019) focused on developing models capable of predicting student performance based on various factors. The former employed clustering algorithms and decision trees to predict the academic performance of high school students using demographic and academic data, with the primary objective of providing institutions with a tool to identify at-risk students and offer early intervention strategies. The latter employed clustering and k-NN algorithms to predict student performance based on academic achievements, also

* Corresponding Author.
Email Address: ilyas.tammouch@uit.ac.ma (I. Tammouch)
https://doi.org/10.21833/ijaas.2023.07.015
Corresponding author's ORCID profile:
https://orcid.org/0000-0003-2752-4413

aiming to identify at-risk students and improve academic outcomes. Both studies reported experimental results demonstrating the effectiveness of their proposed approaches.

The present paper centers on implementing a deep learning approach known as "deep embedding clustering" to cluster low-performing students based on their achievements, measured through plausible values reflecting uncertainty related to test content and circumstances. Additionally, the paper seeks to map the performance level of each Moroccan region using a choropleth map, identifying regions in need of attention and intervention to empower educators in enhancing educational quality. The paper is structured as follows: Section two delves into the dataset and the methods employed to cluster students based on their performance. Section three encompasses the production of the choropleth map and a discussion of the results, while Section four provides concluding remarks.

2. Data and methods

2.1. Data

The program for international student assessment (PISA) is an evaluative framework aimed at gauging the proficiency of 15-year-old students across approximately 80 diverse countries and educational systems. Its primary objective is to assess students' capacity to apply their reading, arithmetic, and science knowledge and skills in tackling real-world challenges. In the specific context of PISA 2018, an additional assessment focusing on financial education was incorporated and implemented in the United States (Pholphirul and Teimtad, 2018; Brunello and Rocco, 2013; Schleicher, 2019).

The selection of the PISA dataset was driven by several compelling reasons. Firstly, PISA represents one of the few open-source empirical repositories encompassing data on Moroccan students' educational accomplishments. This characteristic facilitated our access to comprehensive information regarding students, including details about their parents' backgrounds, socioeconomic status, school conditions, class sizes, and other pertinent factors. The richness of the PISA dataset allowed for an in-depth exploration of the intricate relationships between educational achievement and its potential determinants.

Furthermore, a notable feature of PISA 2018 was the inclusion of a questionnaire that captured participants' attitudes toward education and emotional states, which was absent in previous iterations. This supplementary information provided valuable insights into the students' perspectives and emotional dispositions, contributing to a more holistic understanding of the educational landscape.

Regarding data computation, PISA 2018 employed a robust methodology to derive ten plausible values for each field and subfield, such as Mathematics, reading, science, global competency, Cognitive Process Subscale of Reading-Locate Information, Cognitive Process Subscale of Reading–Understand, Cognitive Process Subscale of Reading-Evaluate and Reflect, Text Structure Subscale of Reading–Single, and Text Structure Subscale of Reading-Multiple. Consequently, this resulted in a comprehensive set of 90 plausible values for each individual student, further enriching the dataset's analytical potential.

In light of these factors and with a view to contributing to education reform in Morocco, this paper draws upon the student-level survey data from PISA 2018, encompassing 6814 Moroccan students. By utilizing this robust dataset, our study aims to shed light on critical aspects of the education system and potentially inform policy initiatives aimed at enhancing educational outcomes in Morocco.

2.2. Plausible values

Wu (2004) stated that academic performance should reflect students' cognitive skills. But because these abilities are hidden and not readily visible, a set of plausible values is employed to depict the range of possible outcomes for a student's performance around the measured value. It represents the uncertainty associated with the test's content and conditions; this uncertainty in performance calculation is generated by the test's content. Given the difficulty of recognizing the concept to be tested, the test's settings, and the student's mental/physical state, the process for determining plausible values consists of mathematically constructing distributions (referred to as posterior distributions) and assigning a set of random values derived from the posterior distributions to each observation. The posterior distribution, h(θ|x), may be calculated as follows:

$$h(\theta|\mathbf{x}) = \frac{f(\theta|\mathbf{x})g(\theta)}{\int f(\theta|\mathbf{x})g(\theta)d\theta}. \tag{1}$$

That is, if a student's item response pattern is x, then h(θ|x) gives the student's posterior (θ) distribution. We also suppose that originates from a normal distribution g(θ), For a student with item response pattern x, plausible values are drawn at random from a probability distribution with density h(θ|x). As a result, plausible values not only convey information about a student's "ability estimate" but also about the uncertainty associated with this estimate.

2.3. Clustering student

2.3.1. Data-preprocessing

Due to the variance in the number of students in various Moroccan regions, we used a representative random sample (73 students in each region) with the closest mean to the population mean to compare findings across regions.

2.3.2. Determining the K value

The next step was to determine the number of clusters needed for efficient data clustering. We used the Elbow approach to achieve this goal. The deep embedding clustering method works by defining the clusters so that the total variation within a particular cluster is as small as possible. The total inside cluster sum of squares, represented by WSS, indicates how compact the cluster is, and our aim when using the clustering technique is to minimize the WSS. As a result, efficient data clusters are created.

2.3.3. Deep embedding clustering (DEC)

Deep neural networks cannot be trained on labeled data, in contrast to supervised learning. Therefore, the idea is to iteratively improve clusters using an auxiliary target distribution that is generated from the present soft cluster assignment. The grouping and feature representation are gradually improved by this procedure. The first step of deep embedding clustering is to use a non-linear mapping (DNN) to first convert the data space X into a latent feature space Z, which is often much smaller than X (Xie et al., 2016). As a result, the DEC clusters the data by simultaneously learning the theta parameters of the DNN, which maps the data points into Z, and a set of k cluster centers in the feature space Z. In other words, the method's objective is to reduce the distance in an embedding space between similar embedding vectors. It uses Kullback-Leibler (KL) divergence and autoencoders (AE; also known as deep autoencoders) to reduce dimensionality and better characterize the embedding vector representations of data. Aiming to forecast a target value that is equivalent to the input X, AE is an unsupervised learning algorithm (Hinton and Salakhutdinov, 2006). The network in an AE that connects the input X to the hidden layer Z is referred to as the encoder, and the network that connects the hidden layer to the expected output is referred to as the decoder.

The fundamental idea behind an AE is to learn the weight values for encoders and decoders through both feedforward and backpropagation while reducing the dimensionality of the hidden layer Z to minimize the information in the supplied inputs. The stacked AE uses greedy layerwise training (Bengio et al., 2006), which comprises two steps: Pre-training and fine-tuning. The layers of the encoder have [90, 500, 500, 200, 2] hidden units, and each layer performs unsupervised learning on a layer-by-layer basis with a mean squared error objective function. The decoder has hidden units of [2, 200, 500, 500, 90] in the reverse order of the encoder. Following the pre-training process, the encoder and decoder are combined after pre-training to carry out fine-tuned learning. To avoid model overfitting, the dropout method (Srivastava et al., 2014) was also used. The initial layers of the encoder and decoder are composed according to Eqs. 2 and 3.

$$\boldsymbol{h}_j = \boldsymbol{b}_j + \textstyle\sum_i \boldsymbol{g}_i \boldsymbol{w}_{ij} \tag{2}$$

$$\mathrm{SeLU}(\boldsymbol{h}_j) = \lambda \begin{cases} \text{and} h_j, & \text{and if } \boldsymbol{h}_j > 0 \\ \text{and} \alpha e^{h_j}, & \text{and if } h_j \le 0 \end{cases}. \tag{3}$$

After fine-tuning, the latent space layer of the encoder zi constitutes the first Z space representation. We refine the cluster centroid $\{\mu_j\}_{j=1}^{k}$ by continually updating zi to enhance clustering efficiency. The objective function for clustering reduces the difference between the soft assignment qij and auxiliary target distribution pij by utilizing KL divergence, as shown in Eq. 4. The likelihood that an embedding point zi in the Z space is clustered into j is denoted by the symbol qij.

$$q_{ij} = \frac{\left(1+\|z_i-\mu_j\|^2\right)^{-1}}{\sum_k \left(1+\|z_i-\mu_k\|^2\right)^{-1}}. \tag{4}$$

Eq. 5 illustrates how pij can be used to boost clustering coupling, while fj specifies the soft cluster frequencies, which are represented by Eq. 6.

$$p_{ij} = \frac{q_{ij}^2/f_j}{\sum_k q_{ik}^2/f_k}. \tag{5}$$

$$f_i = \textstyle\sum_i q_{ij}. \tag{6}$$

Additionally, KL divergence-based minimization of data distribution and embedding space distribution are beneficial for visualizing data and space reduction (Van der Maaten and Hinton, 2008). Eq. 7 provides the objective of the goal function for KL divergence.

$$L = \mathrm{KL}(P \parallel Q) = \textstyle\sum_i \sum_j p_{ij} \log \frac{p_{ij}}{q_{ij}}. \tag{7}$$

Fig. 1 illustrates the DEC procedure as a whole. In this study, the utilization of specific parameters is outlined. The Gaussian distribution of the learning parameters employed for the greedy layer-wise training was initialized with a standard deviation of 0.01. Following the pretraining phase, the entire deep autoencoder was further fine-tuned for 500 iterations without dropout, subsequent to a 200-iteration pretraining phase where a dropout rate of 20% was applied to each layer.

The minibatch size was set to 100, while the learning rate was fixed at 0.01. Stochastic gradient descent served as the optimization method for both the layer-wise pretraining and the subsequent end-to-end finetuning of the autoencoder.

2.4. Model performance metrics

2.4.1. Silhouette coefficient (SC)

A good cluster has very little difference between samples from the same category and a very significant difference between samples from other categories. Rousseeuw (1987) presented the silhouette coefficient (SC), which may assess both properties simultaneously. Better clusters are provided by a model with a higher silhouette coefficient score.

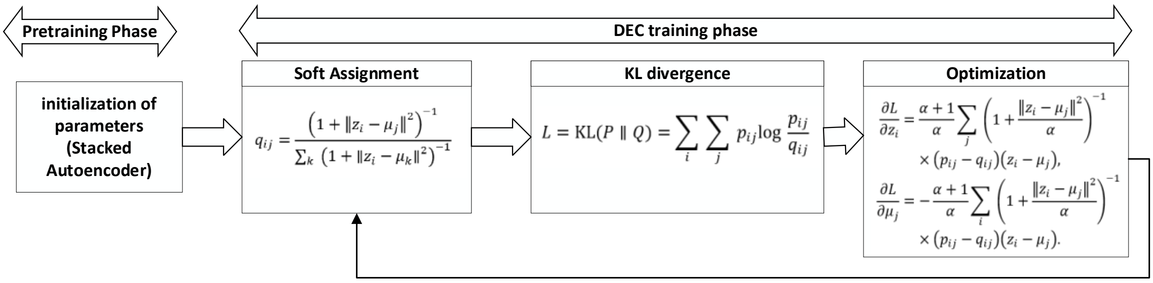

Fig. 1: The process of deep embedding clustering

For a single sample, the silhouette coefficient s is given as:

$$SC = \frac{b-a}{max(a,b)}. \quad (8)$$

where, a represents the average distance between a sample and all other points in the same cluster, and b represents the average distance between a sample and all other points in the next closest cluster. The silhouette coefficient for a group of samples is calculated by taking the mean of the silhouette coefficients for each sample.

2.4.2. Calinski–Harabasz index (CHI)

According to Kozak (2012), a model with a higher Calinski-Harabasz score has better clusters. Where s is the Calinski-Harabasz score for a collection of data E of size n_E that has been clustered into k clusters and is defined as the ratio of the between-cluster dispersion mean and the within-cluster dispersion mean.

$$CHI = \frac{tr(Bk)}{tr(Wk)} \times \frac{n_E - k}{k-1} \quad (9)$$

where, $tr(Bk)$ represents the trace of the dispersion matrix across clusters, and $tr(Wk)$ represents the trace of the dispersion matrix within a cluster defined by:

$$w_k = \sum_{q=1}^{k} \sum_{x \in C_q} (x - C_q)(x - C_q)^T \quad (10)$$

$$B_k = \sum_{q=1}^{k} n_q (C_q - C_E)(C_q - C_E)^T. \quad (11)$$

where, C_q stands for the collection of points in cluster q, C_q for the center of cluster q, C_E for the center of E, and n_q for the number of points in cluster q.

2.4.3. Davies–Bouldin index (DBI)

According to Davies and Bouldin (1979), a model with a larger cluster separation correlate to a lower Davies-Bouldin index. Where the index is the average similarity between each cluster (Ci) and its nearest neighbor (Cj) for i=1, …., k Similarity is described in the context of this index as a Rij trade-off metric. The Davies–Bouldin index is calculated as follows:

$$DBI = \frac{1}{k} \sum_{i=1}^{k} \max_{i \neq j} Rij. \quad (12)$$

2.5. Performance map

After clustering students into two clusters, we group the clustered data by region, then we count the number of students assigned to cluster 0 (low-performing students) and the number of students assigned to cluster 1 (high-performing students) for each region. Then, using the following procedure, we determine the difference between the two counts and then scale it from -1 to 1:

1. Subtract the total number of students in cluster 1 by the total number of students in cluster 0.
2. Calculate the average between the two values.
3. Divide the difference obtained in step 1 by the average value computed in step 2.

We propose D, which stands for "the difference between the student assigned to cluster 1 compared to the student assigned to cluster 0." It is calculated using the following formula:

$$D = \frac{c1 - c0}{\left[\frac{c1+c0}{2}\right]} \quad (13)$$

where, C1 is the number of students assigned to cluster 1; C0 is the number of students assigned to cluster 0.

A region with a high-performance level has a difference greater than 0. (Most students were assigned to cluster 1). While a difference of less than 0 denotes a region that performs poorly (most students were assigned to cluster 0). Finally, we plot the data using choropleth maps, which are used to display statistical variance among map enumeration units, as explained by Stewart and Kennelly (2010). We then used it to illustrate all 12 Moroccan areas along with their degree of performance (the difference calculated between cluster 0 and cluster 1 for each region).

3. Results and discussion

3.1. Clustering

3.1.1. Estimating the number of clusters

After plotting the graph (Fig. 2), we chose 2 as the cut-off value because, while the WSS is continuing to decrease, it doesn't seem to be doing so at a significant enough rate to support the complexity increase brought on by more clusters.

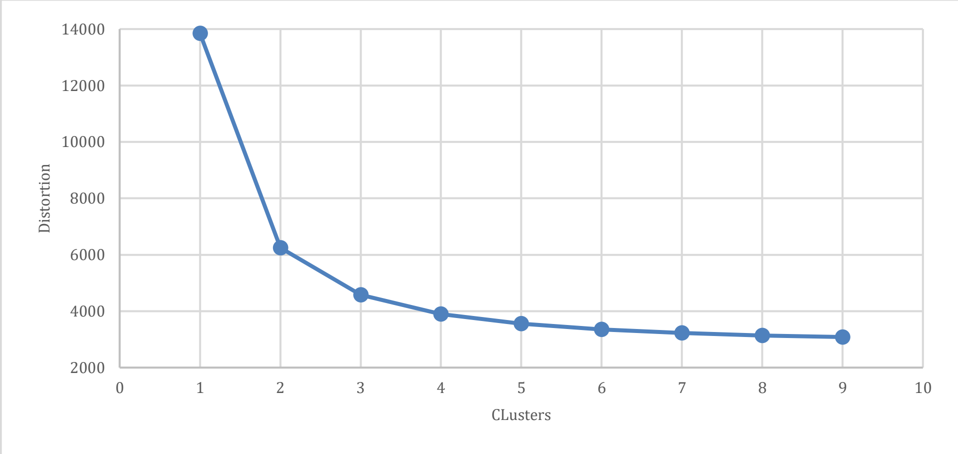

Fig. 2: Estimating the number of clusters using the Elbow method

3.1.2. Clustering students

The dataset was fed as an input to the deep embedding clustering algorithm, and according to the bar plot displayed in Fig. 3 comparing the two clusters, it can be seen that high-performing students lie in cluster 1, while low-performing students lie in cluster 0.

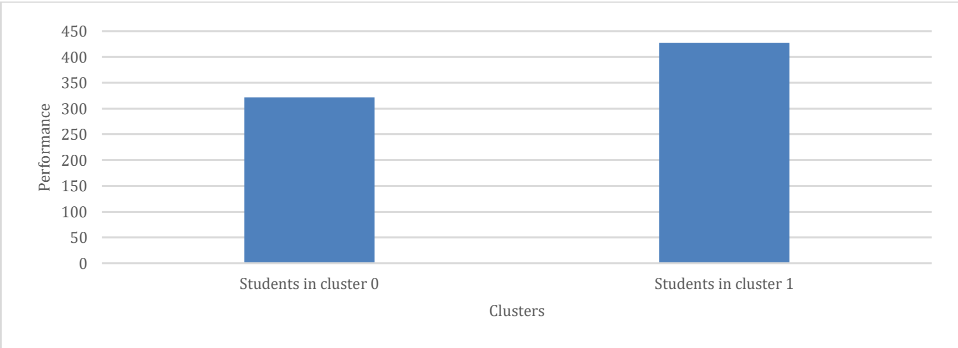

Fig. 3: Comparing clustered data using deep embedding clustering

3.2. Evaluation and comparison of three algorithms

Table 1 shows the comparative findings of each model under various evaluation criteria after measuring and comparing the impacts of the three algorithms on the behavior of clustering student performance using the calculations SC, CHI, and DBI. Showing that DEC (Deep embedding Clustering) could cluster the students more effectively based on their performance since its SC, CHI, and DBI scores were greater than those of k-means and autoencoder(encoder)+k-means.

Table 1: Evaluation and comparison of three algorithms using silhouette coefficient (SC), Calinski–Harabasz index (CHI), and Davies–Bouldin index (DBI)

Algorithm	SC	CHI	DBI
K-means	0.423	8264.12	0.8568
Autoencoder (encoder)+K-means	0.427	8276.32	0.8561
DEC (deep embedding clustering)	0.431	8289.41	0.8556

3.3. Performance choropleth map

The steps for creating a performance map are as follows:

- Step 1: For every region, we count the number of students assigned to cluster 0 (low-performing students) and the number of students affected to cluster 1 (high-performing students), as shown in Table 2.
- Step 2: We calculate D "the difference between the number of students in each cluster for every region on a scale from -1 to 1" as tabulated in Table 3 using the formula shown in Eq. 13.
- Step 3: The last step is to create a choropleth map displaying the performance level of all 12 Moroccan regions using the difference "D" calculated in step 2. The results displayed in Fig. 4, indicate that 'Guelmim-Oued Noun' has the highest academic achievement level, while ' Dakhla-Oued Eddahab' has the lowest academic achievement level among all regions. It should be noted that even regions closer to a performance level of 0 must be taken into account for further education reform.

Table 2: Number of students assigned to cluster 0 and students affected to cluster 1

Region	Number of low-performing students (cluster 0)	Number of high-performing students (cluster 1)
Tanger-Tetouan-Al Hoceima	49	24
Oriental	54	19
Fès-Meknès	41	32
Rabat-Salé-Kénitra	38	35
Béni Mellal-Khénifra	54	19
Casablanca-Settat	39	34
Marrakech-Safi	44	29
Drâa-Tafilalet	36	37
Souss-Massa	35	38
Guelmim-Oued Noun	33	40
Laayoune-Sakia El Hamra	38	35
Eddakhla-Oued Eddahab	58	15

Table 3: Difference between the number of students in each cluster

Region	D
Tanger-Tetouan-Al Hoceima	-0.27
Oriental	-0.67
Fès-Meknès	0.35
Rabat-Salé-Kénitra	0.60
Béni Mellal-Khénifra	-0.67
Casablanca-Settat	0.51
Marrakech-Safi	0.12
Drâa-Tafilalet	0.76
Souss-Massa	0.83
Guelmim-Oued Noun	1
Laayoune-Sakia El Hamra	0.60
Eddakhla-Oued Eddahab	-1

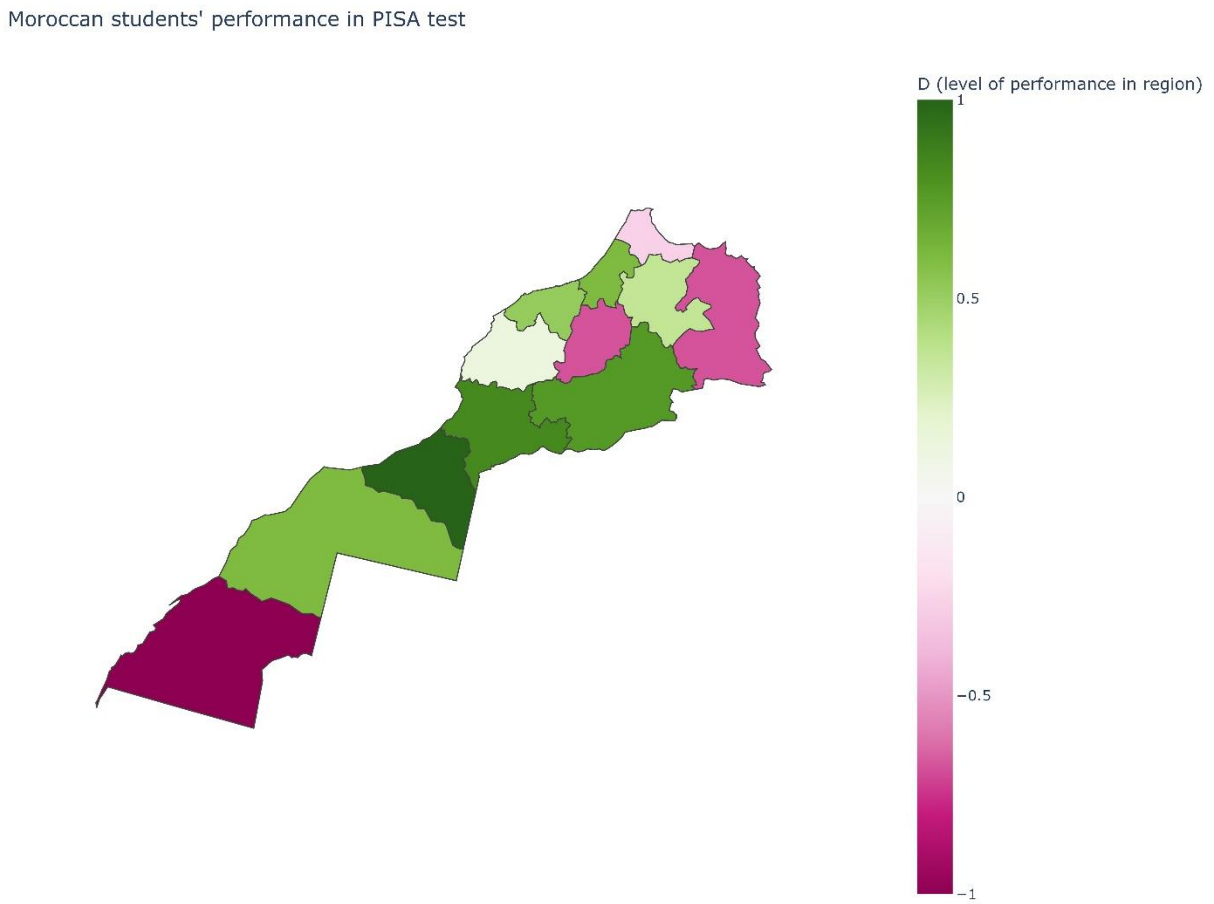

Fig. 4: Performance map

4. Conclusion

The primary objective of this research is to employ an unsupervised deep learning approach, "deep embedding clustering," to cluster Moroccan regions based on their students' academic achievements in the PISA test. The ultimate goal is to generate a map illustrating regions with lower academic performance among the 12 Moroccan regions, with the intention of facilitating targeted educational reforms. This approach involves learning a mapping from the data space to a lowerdimensional feature space and iteratively optimizing a clustering objective to identify regional patterns and disparities in the realm of Moroccan education.

The results of the study highlight that "Dakhla-Oued Eddahab," "Béni Mellal-Khénifra," and "Oriental" exhibit the lowest academic achievement levels among all 12 regions. This finding calls for urgent attention from academicians and policymakers to implement appropriate measures in these regions. Possible strategies may include offering targeted support to disadvantaged schools, empowering vulnerable families, combating gender stereotypes, providing assistance to single-parent households, and fostering supportive learning environments within schools. Implementing such measures is crucial for enhancing students' educational outcomes in these regions.

Furthermore, it is essential to acknowledge that the approach adopted in this study, aimed at mapping students' performance, demonstrates favorable generalizability. As a result, it holds the potential for broader applicability, extending its relevance to the 80 countries covered in the PISA 2018 assessment.

Compliance with ethical standards

Conflict of interest

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

References

Bengio Y, Lamblin P, Popovici D, and Larochelle H (2006). Greedy layer-wise training of deep networks. In the 19th International Conference on Neural Information Processing Systems, MIT Press, Vancouver, Canada, 19: 153-160. https://doi.org/10.7551/mitpress/7503.003.0024

Brunello G and Rocco L (2013). The effect of immigration on the school performance of natives: Cross country evidence using PISA test scores. Economics of Education Review, 32: 234-246. https://doi.org/10.1016/j.econedurev.2012.10.006

Davies DL and Bouldin DW (1979). A cluster separation measure. IEEE Transactions on Pattern Analysis and Machine Intelligence, PAMI-1(2): 224-227. https://doi.org/10.1109/TPAMI.1979.4766909

Hinton GE and Salakhutdinov RR (2006). Reducing the dimensionality of data with neural networks. Science, 313(5786): 504-507. https://doi.org/10.1126/science.1127647 **PMid:16873662**

Kozak M. (2012). "A dendrite method for cluster analysis" by Caliński and Harabasz: A classical work that is far too often incorrectly cited. Communications in Statistics-Theory and Methods, 41(12): 2279-2280. https://doi.org/10.1080/03610926.2011.560741

Lailiyah S, Yulsilviana E, and Andrea R (2019). Clustering analysis of learning style on Anggana high school student. TELKOMNIKA (Telecommunication Computing Electronics and Control), 17(3): 1409-1416. https://doi.org/10.12928/telkomnika.v17i3.9101

Pasina I, Bayram G, Labib W, Abdelhadi A, and Nurunnabi M (2019). Clustering students into groups according to their learning style. MethodsX, 6: 2189-2197. https://doi.org/10.1016/j.mex.2019.09.026 **PMid:31667119 PMCid:PMC6812368**

Pholphirul P and Teimtad S (2018). Living with parents and educational outcomes in developing countries: Empirical evidence from PISA Thailand. Journal of Population Research, 35: 87-105. https://doi.org/10.1007/s12546-017-9196-1

Rousseeuw PJ (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics, 20: 53-65. https://doi.org/10.1016/0377-0427(87)90125-7

Schleicher A (2019). PISA 2018: Insights and interpretations. Available online at: https://www.oecd.org/pisa/PISA%202018%20Insights%20and%20Interpretations%20FINAL%20PDF.pdf

Shovon HIM and Haque M (2012). An approach of improving students academic performance by using k means clustering algorithm and decision tree. International Journal of Advanced Computer Science and Applications, 3(8): 146-149. https://doi.org/10.14569/IJACSA.2012.030824

Srivastava N, Hinton G, Krizhevsky A, Sutskever I, and Salakhutdinov R (2014). Dropout: A simple way to prevent neural networks from overfitting. The Journal of Machine Learning Research, 15(1): 1929-1958.

Stewart J and Kennelly PJ (2010). Illuminated choropleth maps. Annals of the Association of American Geographers, 100(3): 513-534. https://doi.org/10.1080/00045608.2010.485449

Van der Maaten L and Hinton G (2008). Visualizing data using t-SNE. Journal of Machine Learning Research, 9: 2579-2605.

Wu M (2004). Plausible values. Rasch Measurement Transactions, 18(2): 976-978.

Xie J, Girshick R, and Farhadi A (2016). Unsupervised deep embedding for clustering analysis. In the 33rd International Conference on Machine Learning, PMLR, New York, USA: 478-487.