JOURNAL OF NEW APPROACHES IN EDUCATIONAL RESEARCH
2023, VOL. 12, NO. 1, 153-170, e-ISSN: 2254-7339
https://doi.org/10.7821/naer.2023.1.1231

ORIGINAL

Use of Generative Adversarial Networks (GANs) in Educational Technology Research

Anabel Bethencourt-Aguilar[1], Dagoberto Castellanos-Nieves[2], Juan-José Sosa-Alonso[1] and Manuel Area-Moreira[1]

[1]Department Didactics and Educational Research, University of La Laguna, Spain
[2]Department Computer and Systems Engineering, University of La Laguna, Spain

ABSTRACT

In the context of Artificial Intelligence, Generative Adversarial Nets (GANs) allow the creation and reproduction of artificial data from real datasets. The aims of this work are to seek to verify the equivalence of synthetic data with real data and to verify the possibilities of GAN in educational research. The research methodology begins with the creation of a survey that collects data related to the self-perceptions of university teachers regarding their digital competence and technological-pedagogical knowledge of the content (TPACK model). Once the original dataset is generated, twenty-nine different synthetic samples are created (with an increasing N) using the COPULA-GAN procedure. Finally, a two-stage cluster analysis is applied to verify the interchangeability of the synthetic samples with the original, in addition to extracting descriptive data of the distribution characteristics, thereby checking the similarity of the qualitative results. In the results, qualitatively very similar cluster structures have been obtained in the 150 tests carried out, with a clear tendency to identify three types of teaching profiles, based on their level of technical-pedagogical knowledge of the content. It is concluded that the use of synthetic samples is an interesting way of improving data quality, both for security and anonymization and for increasing sample sizes.

Keywords ARTIFICIAL INTELLIGENCE, SYNTHETIC DATA, EDUCATIONAL RESEARCH, DIGITAL COMPETENCE, TPACK MODEL

Received 2022-09-15
Revised 2022-10-14
Accepted 2022-11-23
Published 2023-01-15

Corresponding Author
Anabel Bethencourt-Aguilar,
abethenc@ull.edu.es

C/ Heraclio Sánchez, 43, Código postal 38204, La Laguna, Santa Cruz de Tenerife, Spain.

DOI https://doi.org/10.7821/naer.2023.1.1231
Pages: 153-170

Funding: Ministry of Science, Innovation and Universities, Spain (Award:FPU19/04821)

1 INTRODUCTION

1.1 Generative Adversarial Networks (GAN): Origin, Use and Applications

Governance in political matters related to the use of data systems coexists with a widespread concern about access to and legislative limitation on the treatment of information (Bonnéry et al., 2019). Digital technologies can collect, store and distribute large amounts of personal data, putting the right to privacy in question, while also favoring the development

of strategies that safeguard its protection. In this dilemma between the resolution and creation of problems by technologies (Mick and Fournier, 1998 cited in Bonami et al., 2020), Privacy Enhancing Technologies (PET) promote data protection, allowing the publication of confidential information without violating the privacy of the people involved, or extracting the data without using confidential information to protect the right to privacy (Kyritsi, Zorkadis, Stavropoulos, & Verykios, 2019). In this context, Generative Adversarial Networks (GAN) were born, introduced by Goodfellow et al. (2014), and based on deep neural networks (DNN).

Recent advances in machine learning have made a variety of deep learning models available that learn from a wide range of data types (Liu et al., 2019). GANs can be considered a type of machine learning framework composed of two competing neural networks. These networks have been applied to a variety of formats such as the generation of images or videos (Burlina, Joshi, Pacheco, Liu, & Bressler, 2019; Vallez, Mata, Cotorro, & Deniz, 2019), music, text conversion or data generation synthetics (Bautista & Inventado, 2021; Goodfellow et al., 2014) and in different fields of knowledge such as in the diagnosis of diseases such as cancer in preventive medicine (Burlina et al., 2019; Vilardell et al., 2020), dissemination of synthetic data from patients with diabetes (Kaur et al., 2020) and the provision of freely accessible geological information for educational and research purposes (Lishchuk, Haller, Martinsson, & Bauer, 2021), among many other uses that nourish specialized academic literature.

The multiplicity and flexibility of these types of techniques can favor the creation of any set of relevant data regardless of the field of research and in different formats of representation and materialization. Through prediction and correction, neural networks "learn" to reproduce the data and to generalize it, which makes them especially suitable for the creation of artificial data. Tabular synthetic data is artificially generated data that mimics real-life data stored in tables and can be defined from sources such as student databases, behavioral analytics information, financial records from a university, or surveys. Starting from a given training dataset (the ones actually obtained from the empirical study), GANs learn to generate new data with the same statistics as the training set, maintaining the same original joint distribution (data structure) and even allowing an expansion of the original data (sample size) (Creswell et al., 2017; Liu et al., 2019; Vilardell et al., 2020).

The advantages of these procedures, derived from artificial intelligence, are clear in the context of the scarcity of data and the inherent difficulties in availability and access, the reduction of costs in obtaining data, and the ability to increase data security and privacy. The very nature of their creation allows synthetic data to function as a direct replacement for any behavioral, predictive, or transactional analytics. Even so, it is important to check whether the resulting models using synthetic data provide good results in cases where substantial differences in the behaviour of the data are found (Vallez et al., 2019; Yoon, Drumright, Van Der, & Schaar, 2020).

1.2 Benefits of GANs for Educational Research

These types of procedural techniques have the potential to resolve or remove some of the limitations, difficulties and shortcomings still present in the field of educational research. Educational research, limited from the beginning by data collection, needs to make a considerable effort to obtain a substantial population sample on which to base its studies. This effort, both temporal and economic, generally leads to small samples with, on many occasions, missing data and incomplete records that delimit and condition the emerging interpretations of these investigations. The academic literature on educational research typically reflects the limitations and difficulties that our field has when selecting and accessing a sample that contains representativeness and significance of the study population (Colas-Bravo, 1985; Mayorga-Fernández & Ruiz-Baeza, 2014). Similarly, we must add to this problem in educational research the complexities of research that include sensitive information such as the perceptions of educational agents, or private data such as interaction on digital platforms. Educational research promotes the use of strategies that anonymize or prevent re-identification; however, this implies certain losses of implicit information in the original data (Kyritsi et al., 2019).

The use of these types of network algorithms can be relevant to research on small scales, such as studies conducted in the educational field in a delimited and specific context (Bautista & Inventado, 2021), which can take advantage of the generation of artificial data that increase the sample size. The use of GANs does not imply the loss of information; on the contrary, adjusted synthetic data can be generated to improve the quality of their relationships, in addition to increasing or reducing anomalous values found in the data sets (Lin, Jain, Wang, Fanti, & Sekar, 2020). At present, along with research emerging from the field of educational technology such as studies of behaviors in virtual learning environments (Cheng et al., 2020; Dorodchi, Al-Hossami, Benedict, & Demeter, 2019) and new studies on academic data for the digital transformation of universities (Bethencourt-Aguilar, Area-Moreira, Sosa-Alonso, & Castellano-Nieves, 2021; Ndou, Ajoodha, & Jadhav, 2020), new information processing procedures and techniques are also needed to guarantee the protection of privacy rights and to hold information about the users, patterns and circumstances of educational agents. The use of Artificial Intelligence in the field of education can solve existing problems regarding samples, treatment and dissemination of data, and is an example of the new methodologies that are being incorporated into education (Bonami et al., 2020).

The integration of Generative Adversarial Networks into education also provides secondary benefits worthy of attention. In addition to protecting the anonymity of the information, if it is shown that the equivalence and degree of similarity in the extension of the real samples is adequate, it would be possible to carry out other analyses that are only possible when a considerable number of records is reached, therefore reaching a better sample size. Or, for example, the results obtained with different sample sizes can be compared, when it comes to techniques that are sensitive to sample size. In other words, the extension of records multiplies the possibilities for analysis and for obtaining emerging interpretations born out of the growth of the data. Furthermore, the artificial creation of data has

an interesting consequence for the current movements in favor of open science in the field of education. The guarantee of data security can encourage availability so that the data is released and analytical techniques can be performed without the constraints that data collection implies (Dorodchi et al., 2019), as is happening in other disciplines (Lin et al., 2020; Yale et al., 2020), encouraging the exchange of datasets or libraries that are useful both for the learning of new students and for the testing of analytical and novel methodological techniques by experienced researchers (Kyritsi et al., 2019; Lishchuk et al., 2021).

1.3 Digital Competence and Technological Pedagogical Content Knowledge (TPACK)

Digital competence is one of the main lines of attention in the specialized academic literature on educational technology. Considering that the integration of technologies into teaching processes is a complex, dynamic and contextually determined problem, Mishra and Koehler developed the TPACK model as a framework to understand and think, in a systematic way, about the complexities posed by the integration of technologies into teaching (Koehler, Mishra, & Yahya, 2008; Mishra & Koehler, 2006).

The fundamental idea of the TPACK model is based on the fact that good teaching with technologies is based on the integration of three types of knowledge available to teachers: disciplinary knowledge or knowledge related to the content to be taught, pedagogical knowledge and technological knowledge. In the TPACK framework, teaching competence is nurtured by these multiple interactions (see Figure 1).

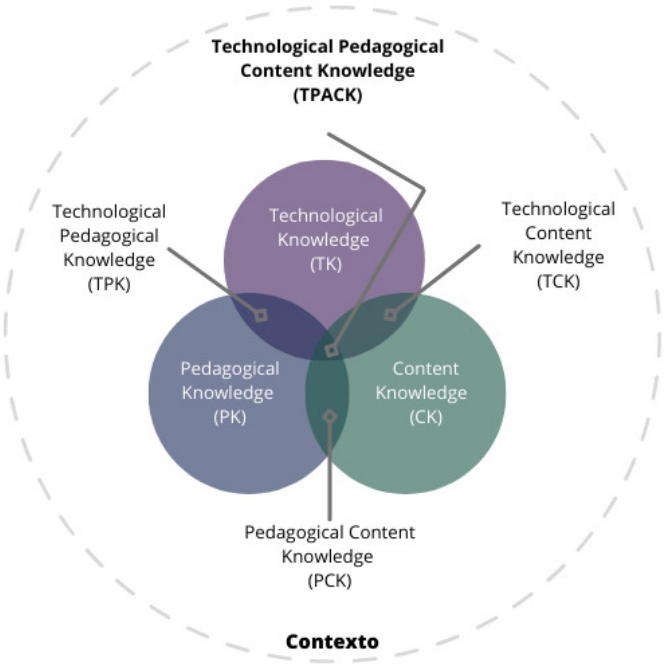

Figure 1 TPACK model (Koehler et al., 2008; Mishra & Koehler, 2006)

The TPACK model covers the understanding of the representations of concepts through the use of technologies; the pedagogical techniques that apply technologies in a constructive way to teach the content according to the learning needs of the students; knowledge of what makes concepts difficult or easy to learn and how technology can help address conceptual challenges; knowledge of the previous understanding of the students regarding the content and epistemological assumptions; and knowledge of how technologies can be used to build new epistemologies or reinforce old ones related to existing understanding.

Based on these contributions, a large amount of research has been generated that pays attention to the profiles and behaviour patterns of teachers and encourages continuous searching and reflection on the models of pedagogical use of technologies (Area-Moreira, Hernández-Rivero, & Sosa-Alonso, 2016) and on the very meaning of digital competence (Castañeda, Esteve, & Adell, 2018). In this sense, multiple studies have detected profiles according to digital competence such as "sporadic ICT user teacher, transmitter of knowledge" and "regular ICT user teacher, with a more diverse didactic approach" (Area-Moreira et al., 2016) or, more recently, another study on digital competence provided profiles that also show differential teaching styles, such as the "inspirational teacher", "creator" and "tutor" (Esteve-Mon, Llopis-Nebot, Viñoles-Cosentino, & Segura, 2020). This type of research on teaching practices is relevant in order to analyze, in the context of design and instruction, pedagogical improvements in the TPACK model (Koh, Chai, & and, 2014; Yeh, Chan, & Hsu, 2021), and, on the other hand, to favor studies on the perceptions of educational agents that extract conglomerate trends which can be used to draw interesting conclusions from the relationships between the emerging profiles (Huang & Lajoie, 2021; Koh & Chai, 2014; Reyes, Reading, Doyle, & Gregory, 2017) . Furthermore, studies on digital competence allow us to establish lines of improvement in university teacher training, such as for the detection of needs in the pedagogical use of ICT (Esteve-Mon, Llopis-Nebot, & Segura, 2020) or, specifically, for the evaluation of educational practice (Basilotta-Gómez-Pablos, Matarranz, Casado-Aranda, & Otto, 2022).

Following this theme, this study aims to perform an experiment with a small set of real data obtained from surveys by submitting it to different GANs that, treating it as synthetic tabular data, determine the creation of a new set of synthetic data on digital competence to analyze the usefulness of using these algorithms for educational research.

2 MATERIAL AND METHODS

The present study aims to explore the application of algorithms derived from Generative Adversarial Networks (GAN) in educational research. Specifically, an experiment will be carried out with a set of real educational data obtained from surveys that will determine the creation of a new set of tabular synthetic data on digital competence. The level of similarity in the conservation of the initial characteristics in the synthetic data will be verified as a relevant value for the protection of the data, as well as for the suitability and reliability of their expansion in the performance of the selected analyses.

The comparative analysis of the original sample with the generated synthetic samples will be carried out with a data analysis technique implemented in SPSS under the name of "two-stage cluster" (Bacher, Wenzig, & Vogler, 2004; Chiu, Fang, Chen, Wang, & Jeris, 2001; Hurtado & Baños, 2017) . In this way, the possibility of exchanging original samples for synthetic ones in the field of educational research will be tested and, in parallel, the analysis technique itself will be studied, through a total of 150 different tests (original sample plus 29 synthetic ones in five different orders each).

The hypotheses of this work, therefore, are:

- H1. The dataset generated with synthetic data can replace the original dataset, maintaining the distributions and internal characteristics of the initial data.
- H2. The patterns detected on digital competence in teachers are maintained in the original and synthetic datasets.

2.1 Sample and Instrument Identification

The original dataset pertains to data obtained in the application of a survey instrument carried out on the teaching staff of postgraduate degrees in the academic year 2021-22 at the University of La Laguna, which was designed, applied and analyzed by the research team that authored this article. The original survey contains about 150 variables, although only 30 variables were selected for the generation of the synthetic data and these were related to the teachers' perceptions of the level of use of digital teaching resources, the frequency of carrying out personal digital use activities, digital competence, and their technological and technological-pedagogical knowledge. Only 12 items, more focused on the comparability of results in the application of the cluster analysis, were selected for the qualitative comparison.

1. Self-perceived digital competence
2. Knowledge to solve technical problems with technology (TK)
3. Ability to easily assimilate new technological knowledge (TK)
4. Up-to-date knowledge of important new technologies (TK)
5. Frequency of playing and testing technology (TK)
6. Knowledge of many different technologies (TK)
7. Technical knowledge required to use the technology (TK)
8. Sufficient opportunities to work with different technologies (TCK)
9. Knowledge of the most common technological resources and developments in technological and professional performance in my field of knowledge (TCK)
10. Ability to adequately and satisfactorily develop the content of my subject, combining new technologies and the most appropriate methods (TPK-TPCK)
11. Ability to select technologies that improve student learning (TPK-TPCK)
12. Ability to use strategies that combine content, technologies and adequate didactic approaches in teaching materials for the classroom (TPK-TPCK)

These 12 variables are ordinal and categorical and have been designed with five response options. In the "Perceived Digital Competence" item, they range between low level (1), low-medium level (2), medium level (3), medium-high level (4) and high level (5), contrary to the following variables that alternate in degree of agreement.

The sample obtained in the application of the survey has a total of 239 records belonging to the teaching staff of the ULL postgraduate degrees, of which 112 are women and 127 are men. Among the responding faculty, 136 people are civil servants, 90 are contracted doctors and 13 have another contractual relationship. The average age of the responding teachers is between 51 and 65, of which 44 are in the age range between 51 and 55 years old, 56 between 56 and 60 years old and 41 people between 61 and 65 years old. The predominant branch of knowledge of the teaching staff belongs to Social and Legal Sciences (n=80), followed by Health Sciences (n=43), Sciences (n=43), Arts and Humanities (n=42) and Engineering and Architecture (n=31).

The main purpose of this article is to study the possibilities of artificial sample creation, so starting from a small sample is in itself relevant to this work. On the other hand, the research which this article is based on is framed within the theme of the digital transformation of educational institutions. This is an issue that needs further study in the contexts themselves, evaluating the idiosyncrasies of each institution and allowing action lines to be drawn up according to their needs, as well as fostering adjusted and coherent decision-making for each case. The research sample is, therefore, small and contextualized to the University of La Laguna due to the characteristics of the study.

2.2 Synthetic Sample Creation Procedure

The development of a specific approach to synthetic data based on GANs should be influenced by the intended purpose and domain. The proposed methodology gathers the best accepted recommendations in data science (Shafique & Qaiser, 2014; Shearer, 2000). The implementation of the algorithms was carried out in an exploratory and iterative way, ranging from simple techniques to more sophisticated ones. The main phases to be highlighted would be: (a) Data pre-processing (feature extraction, imputation, scaling, GANs); (b) Creation of the models; (c) Model training; (d) Evaluation of the models; (g) Generation of synthetic data; (h) Evaluation of the generated data.

The real dataset has multiple incomplete or null records (36.8%). The small sample collected in the application of the survey corresponds to common events in educational research. However, the architecture of the algorithms used does not facilitate the processing and interpretation of null data, so, after assessing various imputation techniques, it was decided to eliminate the incomplete records. The dataset to be processed by the GANs has 151 records and 30 ordinal categorical variables, of which only 12 variables are presented in this article.

The experiment was conducted with four models based on different techniques for the generation of synthetic data. The implementations of the models used are described by Patki (Patki, Wedge, & Veeramachaneni, 2016). The first of these models is the TVAE, which is based on a deep data variational autocoder introduced in Xu (Xu, Skoularidou, Cuesta-

Infante, & Veeramachaneni, 2019), the second model is the CTGAN (Xu et al., 2019), which is a GAN-based method for modelling the distribution of tabular data and displaying rows of the distribution. The third model, which is called Gaussian Copula (Sklar, 1973), is based on mathematical functions of copula that can be defined as a mathematical function that allows the description of the joint distribution of multiple random variables by analyzing the dependencies between their marginal distributions. Finally, the fourth model is called CopulaGAN, and is a variation of the CTGAN model that uses the transformation of the Gaussian copulas to facilitate the task of the underlying CTGAN model in learning the training data.

2.3 Data Analysis Procedure

The analysis of the quality of the generated data was performed by following different steps until synthetic data with a desired goodness of fit was obtained, making it ,therefore, similar to the original. Once the model had been obtained, 29 datasets were generated starting with the same number of records as the original and with extended records with an increasing N (see Table 3), ranging between N=151 and N=20,000. The analysis of the dataset was carried out with the algorithm inserted in the SPSS analysis software, called two-stage cluster or in two phases, both in the original set and in the synthetic ones (30 sets in total).

The reason for choosing the two-stage cluster is because it is an analysis "in the absence of theory", aimed at classifying subjects according to the emerging categories in the data itself. Given that the purpose of this work is, in addition to verifying the types of teachers based on the degree of digital and technological pedagogical competence (according to the TPACK model), to verify to what extent the original samples are interchangeable with the synthetic ones, the type of analysis seemed ideal. Other analyses, oriented to variables (and not to subjects), more "theoretically oriented" and based on statistics dependent on the characteristics of the marginal and joint distributions of the variables that intervene in each analysis, could have obscured the verification of the equivalence of the samples (replicating the statistical data, but not the subjects).

Each dataset was analyzed according to five different orders, including the original. The reason for this precaution arises from the sensitivity of the two-stage cluster algorithm, used for formulating the quality of the model and the number of clusters it generates, to the order in which the data is presented for the creation of the nodes (Hurtado & Baños, 2017), which also made it possible to control this effect on the results. In this way, between the original dataset and the 29 synthetic ones, 150 datasets were generated (five different orderings in 30 samples).

A clustering process was carried out on each of the datasets to establish the equivalence of the sample. As a result, a quality index of the generated cluster structure (silhouette) was obtained in each test (five tests per sample, taking into account the five different orderings in each dataset). The five values obtained per sample were summarized in their mean value in each case in order to make the comparison between the 30 datasets considered in this study (see Figure 2).

As a complement to the above information, the number of clusters that emerged from each test was considered in the comparison. The mode of the numerical value of the clusters obtained was used to integrate what was obtained from the five orderings of each dataset.

Noise treatment was not used in any of the operations in this process of creating the clusters, because there were no reasons to justify testing under these conditions.

Finally, a qualitative analysis (profiles) of these clusters was carried out as another indicator of the similarity and adequacy of the synthetic datasets with respect to the real data.

3 RESULTS

3.1 Degree of Similarity and Quality of the Generated Model

The main metrics and hyperparameters of the four trained models are presented in Table 1. The main hyperparameters (epochs and batch) are similar in most of the models in order to evaluate their performance with real training data. The best result (similarity) between the model obtained and the real data used for training was the CopulaGAN-WAN model. However, the training time was considerably longer than in the other three models (TVAE, CTGAN and GaussianCopula). This work valued the orchestration of technologies based on GANs for the analysis of surveys, so the models presented did not pursue the optimality of the solutions.

Table 1 Main metrics and hyperparameters of the obtained models

Models	Hyperparametrers	Epochs/ Batchs	Time (s)	Score
TVAE	discriminator steps = 5	300/ 5000	8	0.54
CTGAN	discriminator steps = 5	300/5000	62	0.68
CopulaGAN - WAN	discriminator steps = 5	300/5000	451	0.89
	generator dim = (256, 256, 256)			
	discriminator dim = (256, 256, 256)			
GaussianCopula	-	300/5000	8	0.59

The evaluation of the similarity of the original data (151 records) to the same amount of generated synthetic records was carried out with the help of statistical metrics. Metrics based on the two-sample Kolmogorov-Smirnov test are used, an extension of the Kolmogorov-Smirnov test that works with the numerical transformation of the values and the Kullback-Leibler divergence, which is a non-symmetric measure of the similarity or difference between two probability distribution functions. The results of these metrics were obtained from a comparison between the real dataset and the synthetic ones. The synthetic data generation process was performed five times, obtaining the metrics of each iteration. The final results of these metrics collect the average of these measurements. These results indicate a very good similarity between the set of real data obtained in the survey and the synthetic data generated by the model (see Table 2).

Table 2 Main evaluation metrics between real and synthetic data

Iterations	Kolmogorov-Smirnov	Kolmogorov-Smirnov Extended	Kullback–Leibler
1	0.9415	0.9415	0.7974
2	0.9395	0.9395	0.7842
3	0.9358	0.9358	0.7724
4	0.9427	0.9427	0.7864
5	0.9303	0.9303	0.8034
Mean average	0.9380	0.9380	0.7888

3.2 Clustering in the Original and Synthetic Set

The results of the comparison of the original data and the synthetic data in the combination of the variables used in our data sets and, specifically, those attributed to digital competence, technological knowledge and technological pedagogical knowledge, show that the generated set of synthetic data can replace the original set, as evidenced by the results of the clustering analysis.

In the grouping performed by the two-stage clustering algorithm, an average model quality of 0.26 was obtained from the five files in the original data set. In the synthetic data sets with the same number of records as the original dataset (151), an average value of 0.27 was reached, even higher than that obtained in the original dataset. After comparing the highest value produced in the five randomly ordered files in both data sets, a model quality of 0.28 was obtained in the original data; in the case of the synthetic data set, the value was 0.31. In other words, the results obtained in the synthetic table that replicates the original dataset, both in its distribution and in the number of records, generated a two-stage cluster model practically equivalent to the original, allowing the replacement of the real dataset by the synthetic ones

The datasets with extended records show average results in the quality of the cluster that oscillate around the value 0.20, with a minimum value of 0.17 and a maximum of 0.23 (not counting the one obtained in the synthetic set of 151 records). Results similar to those in the original dataset are generated in 19 of the 29 datasets. Among the results, there are also 10 synthetic datasets that include a model quality (on average) of less than 0.20 (criterion value at which the cluster structure is considered to be good enough) and a tendency to a gradual increase in the lability in the number of clusters, starting from N=10000. However, the graph of the maximum and minimum values of the model quality indicator shows that a maximum value above the value 0.20 was obtained in all the datasets (Figure 2). The following table shows the average quality of the clusters generated in the five random files.

The study of the quality of clustering in any dataset also focuses on the number of clusters and the qualitative quality of these divisions. The results provide evidence that the number of clusters generated in these analyses is the same in most of the tests. Eighty per cent of the analyzed datasets and their respective frequencies classify the total data into three groupings. Variants can be seen in some sets such as the extended synthetic set of 1900 records that created four groupings and, on the other hand, in some tests performed in five

Table 3 Quality indices of the generated clusters

Dataset	CI	Cluster no.	Dataset	CI	Cluster no.	Dataset	CI	Cluster no.
Original	0.26	3	2500	0.22	3	11000	0.20	3
151	0.27	3	3000	0.22	3	12000	0.19	2
400	0.22	3	3500	0.22	3	13000	0.19	3
650	0.21	3	4000	0.22	3	14000	0.19	2
900	0.23	3	5000	0.21	3	15000	0.20	3
1150	0.21	3	6000	0.20	3	16000	0.20	3
1400	0.20	3	7000	0.19	3	17000	0.20	2
1650	0.19	3	8000	0.20	3	18000	0.22	3
1900	0.19	4	9000	0.19	3	19000	0.17	2
2000	0.20	3	10000	0.19	2	20000	0.18	3

sets with an extension of more than 10000, the classification is in two groupings (never in all of them).

The quality of the generated datasets was analyzed in the present study, but results also emerged from the statistical analysis of clustering inserted in the SPSS software. The results show a latent variability in the two-stage cluster algorithm. The order of the records in the dataset had a marked influence on the quality of the generated cluster. The sensitivity of the algorithm can alter the quality indices and the cluster number, requiring the creation of files with different orders to test the data and to be able to extract a cluster with the highest relevant quality index.Figure 2 shows the silhouette of the quality index (means) of the analyzed sets and the highest frequency (mode) of the number of groupings that were generated.

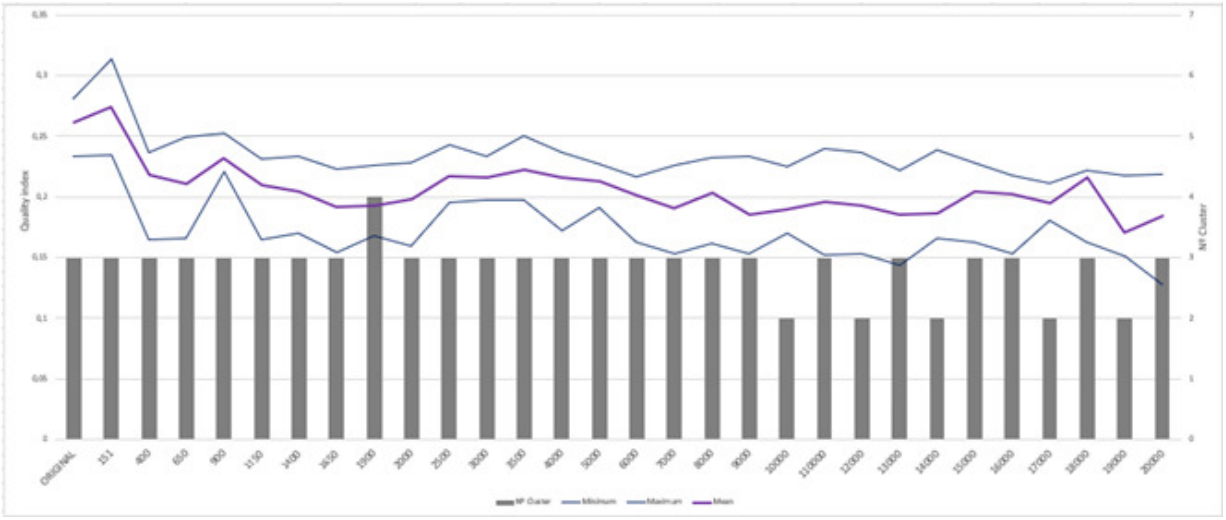

Figure 2 Quality index, variability and cluster number in the data sets

However, distinctions are not only seen in the mean value of the five random orders. The variability of the clustering process is also present in the specific results of each dataset.

The range of variability obtained between the maximum and minimum value referring to the quality of the cluster in the five random orders in each of the data sets, in the 30 data sets analyzed, ranges from values of less than 0.03 to higher variable ranges (0.09).

Despite the fact that the most frequent cluster number in the example was three, the existence of atypical values was also observed in the generation of the cluster number in the synthetic data, from the creation of two clusters to nine, which shows the sensitivity of the two-stage cluster algorithm and the dependence on the order of data entry.

3.3 Qualitative Analysis on Digital Competence Clusters

This last section of results describes the qualitative analysis carried out on the original datasets and on five synthetic sets (N=151, N=900, N=2500, N=8000 and N=18000) of the variables attributed to the digital competence of teachers. The analysis carried out allows a reading of the results in the form of their stability in the extended datasets and the description and interpretation of the clusters and the emerging profile in the teaching staff.

The clustering results show a grouping of the subjects into three clusters, with qualitatively very similar results, regardless of the dataset. In this same analysis, the results of three teaching profiles can be described in each of the samples, including the synthetic samples, with a similar behavior and definition between the different datasets. In terms of the present experiment, the analysis could have been performed with any of the samples (original or synthetic) and regardless of the number of records (151 or versions with a larger number of records), due to the stability found in obtaining the same result, in qualitative terms. The following graph (Figure 3) shows the abovementioned three grouping trends found in all the datasets, analyzed by calculating the weighted average between the variables.

The specific description of each cluster determines the characteristics by which the subjects were attributed to that group. The analysis determined the presence of a first cluster that can be described as teachers with low or medium self-perceived digital competence, who find it difficult to solve technical problems with technology, who are neither in favor of nor against having the ability to assimilate new technological knowledge easily, who do not usually test or play using technology, and who do not know about many different technologies. This grouping can be defined as "C1=techno-insecure teachers". The following cluster is characterized by having a medium-high level in self-perceived digital competence, having a certain ability to acquire technological knowledge, knowledge of the most common technological resources and developments in the technological and professional performance of their field of knowledge, ability to select technologies that improve student learning and to use in their teaching materials, and strategies that combine content, technologies and appropriate didactic approaches. This cluster can be renamed as "C2= techno-autonomous teaching staff". The last cluster presents a medium-high or high level in each of the variables; the teachers in this grouping have extensive knowledge in the use of technology and in how to solve technical problems, the ability to easily assimilate new technological knowledge, frequently use games and perform tests with technology, have fairly up-to-date knowledge of the current important technologies and are quite knowledgeable about the relevant technology in their professional field and area of knowledge, selecting technologies to favor

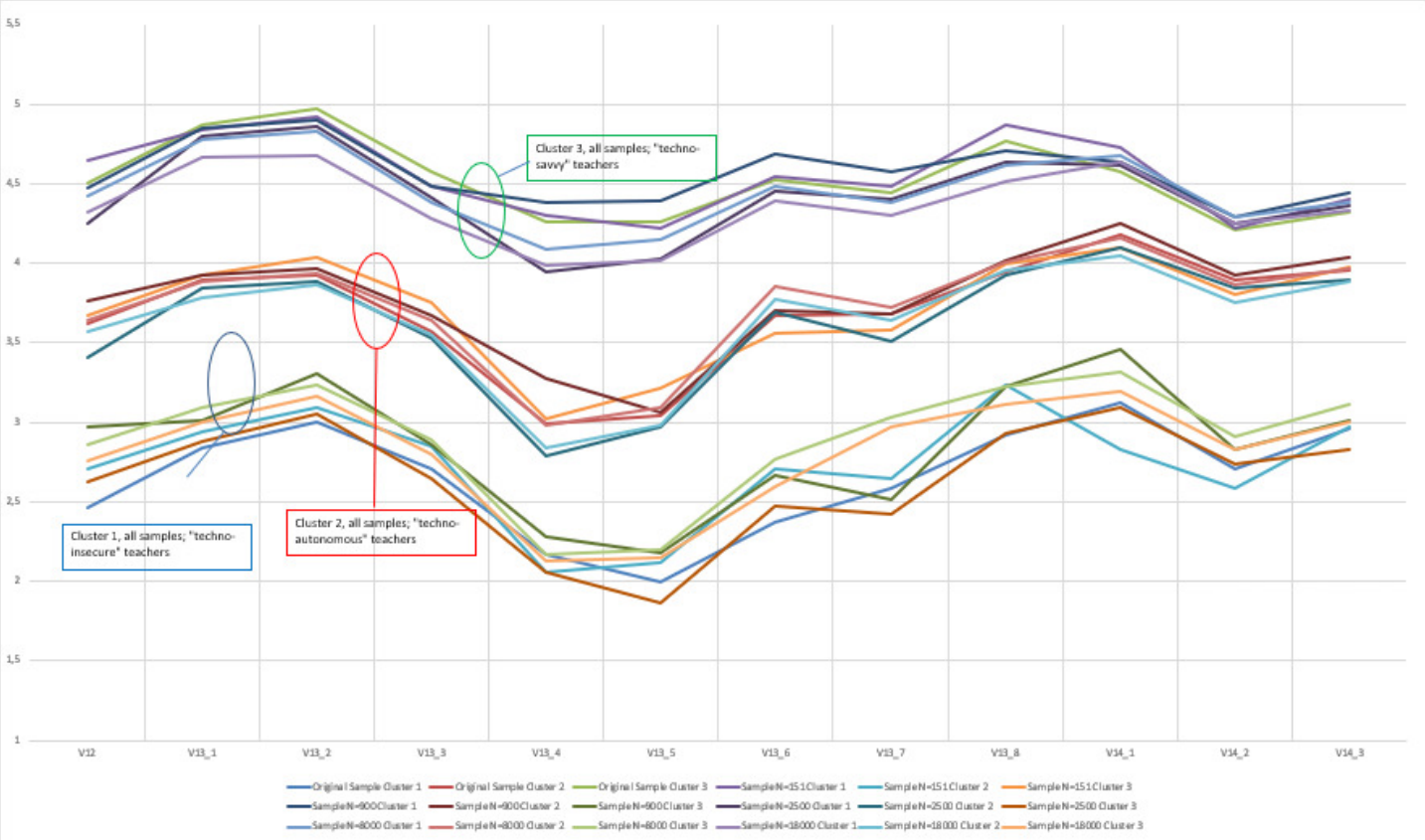

Figure 3 Emerging grouping trends in clustering

student learning and generating strategies that combine content, technologies and didactic methodologies. This third cluster is called, "C3=techno-expert teachers". These three emerging profiles of the teaching staff remain constant in all the analyzed datasets, reaffirming the conglomerate trend in the original dataset and in those created artificially using the model described in the present study. In this conglomerate analysis of teachers' self-perceived digital competence, no significant differences are observed considering variables such as gender, employment status or branch of knowledge; however, significant differences are observed in the age variable.

4 DISCUSSION AND CONCLUSIONS

The growing importance of data processing requires a citizenry that is literate, critical and aware of the responsible use of information. Without being overly optimistic in considering GANs and their possibilities for educational research (Romero, Morante, & López, 2022), they do provide techniques and procedures that can resolve some of the common and well-known problems in the educational field and offer new perspectives and possibilities. The verification in the present study on the potentialities of the use of Generative

Adversarial Networks (GAN), in the field of educational research, supports the conclusion that it is possible to obtain new synthetic datasets, with an adequate degree of similarity to the original, in terms of the characteristics of their joint distributions (data structure), from a typical original dataset in educational research situations: small N, missing data and requiring anonymization processes. It has also been verified that this type of procedure has significant potential to improve available real datasets, both in terms of anonymization and in terms of increasing the sample size and the quality of the initial data, without altering the distributional characteristics of the values of the variables involved (recreating the data structure). These advantages are relevant to current open science movements, as they ensure secure and transparent ways to share research and datasets (Kyritsi et al., 2019; Lishchuk et al., 2021).

It has been verified that, by applying an analysis "lacking theory" (Hurtado & Baños, 2017), aimed at simply detecting patterns in the replicated and increased data structures, in terms of the quality index of the clusters generated, the results obtained offer quantitatively and qualitatively very similar values in the different datasets with small oscillations that are justified more by the variability of the cluster analyses (Hurtado & Baños, 2017) than by differences or deviations between the original sample and the synthetic ones. In addition, the clustering operations were carried out without performing noise treatment; that is, without the records hindering the creation of the nodes having been set aside to avoid distortions in the grouping and, therefore, affecting the quality index of the generated model. If noise treatment had been applied, it would have been possible to ensure, perhaps, a higher quality in the generated cluster models, but this could have reduced transparency when checking the similarity between the original and synthetic datasets (one of the main objectives of this study).

On the other hand, the results found in the extended data set illustrate the sensitivity of the algorithm for generating the clusters to the order in which the data is entered for the creation of the cluster nodes. This lability of results is more a consequence of the characteristics of the clustering technique than of the differences in the datasets. For this reason, it is recommended that, when using this technique, the analyses are repeated, altering the order in which the data are arranged and verifying the grouping trend in the trials. For all of the above, the generation of synthetic data may be an interesting alternative.

In terms of the qualitative results of the clustering process, it can be seen that in both the original sample and in the five synthetic alternative samples (with increasing N) considered, three clusters emerged in all cases, identified as: C1=techno-insecure teachers; C2=techno-autonomous teaching staff; C3=techno-expert faculty. Regardless of the intrinsic value of the finding, the result confirms the replicability and interchangeability between the original and synthetic samples (the same three clusters with identical interpretation and meaning are identified in all the synthetic samples). The existence of different educational practices in the same context born from the perceptions of the university teaching staff themselves coincides with the findings of Area-Moreira et al. (2016), Reyes et al. (2017) and Koh et al. (2014), as well as those expressed in Esteve-Mon, Llopis-Nebot, Viñoles-Cosentino, and Segura (2020) and Basilotta-Gómez-Pablos et al. (2022) on the need to continue encourag-

ing teacher training in digital competence. In this regard, it is important to establish well-oriented training approaches to the different knowledge levels of the TPACK model that favor instructional designs with a well-developed theoretical foundation and that allows the differences to be alleviated and develops the perceptive or real capacities of the teaching staff (Koh et al., 2014; Yeh et al., 2021).

The ethical implications of the use of this technique have the same guarantees that are assumed for the original data set and those inherent to educational research regardless of the sampling, method or procedure. This technology should be understood in terms of the contributions it offers as a complement to new or existing methodologies, without replacing the relevant sampling, techniques and procedures in the field of educational research.

Referring to the hypotheses of this research, we can conclude that the generated dataset is comparable to the original set, maintaining the initial distributions and the same patterns detected on digital competence in teachers. In future studies, the authors will proceed to assess the possibility of carrying out other analyses that, due to their characteristics and requirements, cannot be conducted on the original dataset and which, after the extension of the sample and the improvement of the data set by the GAN, may provide new results and emerging interpretations.

5 AUTHORS' CONTRIBUTIONS

1. **Anabel Bethencourt Aguilar:** Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Resources, Software, Supervision, Validation, Visualization, Roles/Writing - original draft, Writing - review & editing.
2. **Dagoberto Castellanos Nieves:** Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Visualization, Roles/Writing - original draft, Writing - review & editing.
3. **Juan José Sosa Alonso:** Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Visualization, Roles/Writing - original draft, Writing - review & editing.
4. **Manuel Area Moreira:** Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Visualization, Roles/Writing - original draft, Writing - review & editing.

ACKNOWLEDGEMENTS

Funding: This work was supported by the Ministry of Science, Innovation and Universities through a contract in the University Teacher Training Program (FPU) in the Department of Didactics and Educational Research of the Faculty of Education of the University of La Laguna (FPU19/04821). The doctoral student is Anabel Bethencourt Aguilar, the director is Manuel Area Moreira, and the co-directors are Juan José Sosa Alonso and Dagoberto Castellanos Nieves.

This research was conceived in the research and innovation group EDULLAB: Laboratory of Education and New Technologies of the University of La Laguna. The research team would like to thank the School of Doctorate and Postgraduate Studies, the Office of Planning and Communication, the Office of the Vice-President for Digital Agenda, Modernization and the Central Campus, the Office of the Vice-President for Teaching Innovation, Quality and the Anchieta Campus and the Virtual Teaching Unit (UDV) of the University of La Laguna for their support and collaboration in the research that frames the doctoral thesis.

The research team wishes to thank Patrick Dennis for his translation services.

Funded by: Ministry of Science, Innovation and Universities, Spain

Funder Identifier: http://dx.doi.org/10.13039/100014440

Award: FPU19/04821

REFERENCES

Area-Moreira, M., Hernández-Rivero, V., & Sosa-Alonso, J. J. (2016). Modelos de integración didáctica de las TIC en el aula. *Comunicar: Revista Científica de Comunicación y Educación*, *24*(47), 79–87. https://doi.org/10.3916/C47-2016-08

Bacher, J., Wenzig, K., & Vogler, M. (2004). SPSS TwoStep Cluster - a first evaluation. *Arbeits- und Diskussionspapiere*, *2*(2).

Basilotta-Gómez-Pablos, V., Matarranz, M., Casado-Aranda, L., & Otto, A. (2022). Teachers' digital competencies in higher education: A systematic literature review. *International Journal of Educational Technology in Higher Education*, *19*(1), 1–16.

Bautista, P., & Inventado, P. S. (2021). Protecting Student Privacy with Synthetic Data from Generative Adversarial Networks. In I. Roll, D. McNamara, S. Sosnovsky, R. Luckin, & V. Dimitrova (Eds.), *Artificial Intelligence in Education* (pp. 66–70). Springer International Publishing. https://doi.org/10.1007/978-3-030-78270-2_11

Bethencourt-Aguilar, A., Area-Moreira, M., Sosa-Alonso, J. J., & Castellano-Nieves, D. (2021). The digital transformation of postgraduate degrees. A study on academic analytics at the University of La Laguna. *2021 XI International Conference on Virtual Campus (JICV)* (pp. 1–4). https://doi.org/10.1109/JICV53222.2021.9600311

Bonami, B., Piazentini, L., & Dala-Possa, A. (2020). Educación, Big Data e Inteligencia Artificial: Metodologías mixtas en plataformas digitales. *Comunicar: Revista Científica de Comunicación y Educación*, *28*(65), 43–52. https://doi.org/10.3916/C65-2020-04

Bonnéry, D., Feng, Y., Henneberger, A. K., Johnson, T. L., Lachowicz, M., Rose, B. A., … Zheng, Y. (2019). The Promise and Limitations of Synthetic Data as a Strategy to Expand Access to State-Level Multi-Agency Longitudinal Data. *Journal of Research on Educational Effectiveness*, *12*(4), 616–647. https://doi.org/10.1080/19345747.2019.1631421

Burlina, P. M., Joshi, N., Pacheco, K. D., Liu, T. Y. A., & Bressler, N. M. (2019). Assessment of Deep Generative Models for High-Resolution Synthetic Retinal Image Generation of Age-Related Macular Degeneration. *JAMA Ophthalmology*, *137*(3), 258–264. https://doi.org/10.1001/jamaophthalmol.2018.6156

Castañeda, L., Esteve, F., & Adell, J. (2018). ¿Por qué es necesario repensar la competencia docente para el mundo digital? *Revista de Educación a Distancia (RED)*, *56*. Retrieved from https://revistas.um.es/red/article/view/321581

Cheng, Y., Dai, Z., Ji, Y., Li, S., Jia, Z., Hirota, K., & Dai, Y. (2020). Student Action Recognition Based on Deep Convolutional Generative Adversarial Network. *Proceedings of the*

32nd 2020 Chinese Control and Decision Conference (pp. 128–133). Retrieved from http://www.webofscience.com/wos/alldb/full-record/WOS:000621616900023

Chiu, T., Fang, D., Chen, J., Wang, Y., & Jeris, C. (2001). A robust and scalable clustering algorithm for mixed type attributes in large database environment. *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 263–268). https://doi.org/10.1145/502512.502549

Colas-Bravo, M. P. (1985). Dificultades y errores metodológicos en la investigación educativa. *Enseñanza & Teaching: Revista interuniversitaria de didáctica*, *3*, 165–172.

Creswell, A., White, T., Dumoulin, V., Arulkumaran, K., Sengupta, B., & Bharath, A. (2017). Generative Adversarial Networks: An Overview. *IEEE Signal Processing Magazine*, *35*. https://doi.org/10.1109/MSP.2017.2765202

Dorodchi, M., Al-Hossami, E., Benedict, A., & Demeter, E. (2019). Using Synthetic Data Generators to Promote Open Science in Higher Education Learning Analytics. *IEEE International Conference on Big Data (Big Data)*, 4672–4675. https://doi.org/10.1109/BigData47090.2019.9006475

Esteve-Mon, F., Llopis-Nebot, M., & Segura, J. (2020). Digital Teaching Competence of University Teachers: A Systematic Review of the Literature. *IEEE-RITA*, *15*(4), 399–406.

Esteve-Mon, F., Llopis-Nebot, M. A., Viñoles-Cosentino, V., & Segura, J. (2020). Digital Teaching Competence of University Teachers: Levels and Teaching Typologies. *International Journal of Emerging Technologies in Learning*, *17*(13).

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., … Bengio, Y. (2014). Generative Adversarial Nets. *Advances in Neural Information Processing Systems*, *27*. https://doi.org/10.48550/arXiv.1406.2661

Huang, L., & Lajoie, S. P. (2021). Process analysis of teachers' self-regulated learning patterns in technological pedagogical content knowledge development. *Computers & Education*, *166*, 104169. https://doi.org/10.1016/j.compedu.2021.104169

Hurtado, M. J. R., & Baños, R. V. (2017). El análisis de conglomerados bietápico o en dos fases con SPSS. *REIRE: revista d'innovació i recerca en educació*, *10*(1), 118–126.

Kaur, D., Sobiesk, M., Patil, S., Liu, J., Bhagat, P., Gupta, A., & Markuzon, N. (2020). Application of Bayesian networks to generate synthetic health data. *Journal of the American Medical Informatics Association : JAMIA*, *28*(4), 801–811. https://doi.org/10.1093/jamia/ocaa303

Koehler, M. J., Mishra, P., & Yahya, K. (2008). Tracing the development of teacher knowledge in a design seminar: Integrating content, pedagogy, and technology. *Computers & Education*, *49*, 740–762.

Koh, J. H. L., & Chai, C. S. (2014). Teacher clusters and their perceptions of technological pedagogical content knowledge (TPACK) development through ICT lesson design. *Computers & Education*, *70*, 222–232. https://doi.org/10.1016/j.compedu.2013.08.017

Koh, J. H. L., Chai, C. S., & and, L. Y. T. (2014). TPACK-in-Action: Unpacking the contextual influences of teachers' construction of technological pedagogical content knowledge (TPACK). *Computers & Education*, *78*, 20–29. https://doi.org/10.1016/j.compedu.2014.04.022

Kyritsi, K. H., Zorkadis, V., Stavropoulos, E. C., & Verykios, V. S. (2019). The Pursuit of Patterns in Educational Data Mining as a Threat to Student Privacy. *Journal of Interactive Media in Education*, *1*.

Lin, Z., Jain, A., Wang, C., Fanti, G., & Sekar, V. (2020). Using GANs for Sharing Networked Time Series Data: Challenges, Initial Promise, and Open Questions. *Proceedings of the ACM Internet Measurement Conference*, 464–483. https://doi.org/10.1145/3419394.3423643

Lishchuk, V., Haller, E., Martinsson, O., & Bauer, T. E. (2021). Analytical Modeling of a Synthetic VMS Deposit Data: A Proxy Tool for Education and Initial Research. *Mining, Metallurgy and Exploration*, *38*(2), 863–874. https://doi.org/10.1007/s42461-020-00377-5

Liu, Y., Zhou, Y., Liu, X., Dong, F., Wang, C., & Wang, Z. (2019). Wasserstein GAN-Based Small-Sample Augmentation for New-Generation Artificial Intelligence: A Case Study of Cancer-Staging Data in. *Biology. Engineering*, *5*(1), 156–163. https://doi.org/10.1016/j.eng.2018.11.018

Mayorga-Fernández, M. J., & Ruiz-Baeza, V. M. (2014). Muestreos utilizados en investigación educativa en España. *RELIEVE - Revista Electrónica de Investigación y Evaluación Educativa*, *8*(2). Retrieved from https://doi.org/10.7203/relieve.8.2.4364 https://doi.org/10.7203/relieve.8.2.4364

Mishra, P., & Koehler, M. J. (2006). Technological Pedagogical Content Knowledge: A new framework for teacher knowledge. *Teachers College Record*, *108*(6), 1017–1054.

Ndou, N., Ajoodha, R., & Jadhav, A. (2020). Educational Data-mining to Determine Student Success at Higher Education Institutions. *2020 2nd International Multidisciplinary Information Technology and Engineering Conference, IMITEC 2020*. Retrieved from https://doi.org/10.1109/IMITEC50163.2020.9334139 https://doi.org/10.1109/IMITEC50163.2020.9334139

Patki, N., Wedge, R., & Veeramachaneni, K. (2016). The synthetic data vault. *2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA)* (pp. 399–410). IEEE.

Reyes, V. C., Reading, C., Doyle, H., & Gregory, S. (2017). Integrating ICT into teacher education programs from a TPACK perspective: Exploring perceptions of university lecturers. *Computers & Education*, *115*, 1–19. https://doi.org/10.1016/j.compedu.2017.07.009

Romero, W. A. M., Morante, M. C. F., & López, B. C. (2022). Alfabetización mediática crítica para mejorar la competencia del alumnado. *Comunicar: Revista científica iberoamericana de comunicación y educación*, *70*, 47–57.

Shafique, U., & Qaiser, H. (2014). A comparative study of data mining process models (KDD, CRISP-DM and SEMMA). *International Journal of Innovation and Scientific Research*, *12*, 217–222.

Shearer, C. (2000). The CRISP-DM model: the new blueprint for data mining. *Journal of data warehousing*, *5*, 13–22.

Sklar, A. (1973). Random variables, joint distribution functions, and copulas. *Kybernetika*, *9*(6), 449–495.

Vallez, N., Mata, A. V., Cotorro, J. J., & Deniz, Ó. (2019). ¿Es posible entrenar modelos de aprendizaje profundo con datos sintéticos? *XL Jornadas de Automática: libro de actas, Ferrol, 4-6 de septiembre de 2019* (pp. 859–865). https://doi.org/10.17979/spudc.9788497497169.859

Vilardell, M., Buxó, M., Clèries, R., Martínez, J. M., Garcia, G., Ameijide, A., … Borràs, J. M. (2020). Missing data imputation and synthetic data simulation through modeling graphical probabilistic dependencies between variables (ModGraProDep): An application to breast cancer survival. *Artificial Intelligence in Medicine*, *107*, 101875–101875. https://doi.org/10.1016/j.artmed.2020.101875

Xu, L., Skoularidou, M., Cuesta-Infante, A., & Veeramachaneni, K. (2019). Modeling tabular data using conditional gan. *Advances in Neural Information Processing Systems*, *32*. Retrieved from http://arxiv.org/abs/1907.00503

Yale, A., Dash, S., Bhanot, K., Guyon, I., Erickson, J. S., & Bennett, K. P. (2020). Synthesizing Quality Open Data Assets from Private Health Research Studies. *Lecture Notes in Business Information Processing*, *394*, 324–335. https://doi.org/10.1007/978-3-030-61146-0_26

Yeh, Y. F., Chan, K. K. H., & Hsu, Y. S. (2021). Toward a framework that connects individual TPACK and collective TPACK: A systematic review of TPACK studies investigating teacher collaborative discourse in the learning by design process. *Computers & Education*, *171*.

Yoon, J., Drumright, L. N., Van Der, & Schaar, M. (2020). Anonymization Through Data Synthesis Using Generative Adversarial Networks (ADS-GAN). *IEEE Journal of Biomedical and Health Informatics*, *24*(8), 2378–2388. https://doi.org/10.1109/JBHI.2020.2980262