International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

The Impact of Data Preprocessing on the Quality and Effectiveness of E-Learning

Mounia RAHHALI*[1], Lahcen Oughdir[2], Youssef Lahmadi[3], Mohammed Zakariae El Khattabi[4]

Submitted: 12/12/2023 **Revised**: 16/01/2024 **Accepted**: 01/02/2024

Abstract: This article provides a mini review of pre-processing techniques for educational big data in data mining. With the increasing availability of educational data, there is a need for efficient pre-processing techniques that can handle the volume, variety, and velocity of data. The article discusses various pre-processing techniques, including data cleaning, data transformation, and data reduction. The review concludes that pre-processing is a critical step in data mining, and the selection of appropriate techniques depends on the characteristics of the data and the research objectives.

Keywords: *Big Data, Pre-processing, Data Mining, Educational data mining.*

1. Introduction

The field of education is rapidly evolving, and so is the volume of educational data generated. With the increasing amount of data being generated [1], it has become increasingly important to process and analyze this data effectively to extract valuable insights that can improve educational outcomes. Data mining techniques have been used extensively in education to extract insights from educational big data [2]. However, data preprocessing is an essential step in data mining that involves cleaning, transforming, and reducing the data to make it suitable for analysis.

In recent years, the application of data mining in education has gained significant attention, and various studies have been conducted to explore the potential of educational big data analysis. However, there is a lack of literature that comprehensively reviews the pre-processing techniques used in educational big data analysis. This mini-review aims to address this gap by providing an overview of the pre-processing techniques used in educational big data analysis.

The mini-review discusses the various pre-processing techniques used in educational big data analysis, including data cleaning, data transformation, and data reduction. Data cleaning involves removing irrelevant data, identifying and correcting errors, and handling missing data. Data transformation involves converting data into a suitable format for analysis, while data reduction involves reducing the size of the data without losing valuable information.

The mini-review also highlights the importance of pre-processing in educational big data analysis, as it can significantly impact the accuracy and reliability of the results. Furthermore, the review provides insights into the challenges and limitations of pre-processing educational big data, such as the complexity of the data and the need for domain-specific knowledge.

Overall, this mini-review provides a comprehensive overview of the pre-processing techniques used in educational big data analysis and highlights the importance of pre-processing in data mining.

The remaining sections of this paper are structured as follows: Section 2 outlines the relevant prior work; Section 3 introduces the concept of Big Data; Section 4 provides an introduction to Knowledge Discovery and Data Mining; Section 5 describes the various data pre-processing steps and tools; Section 6 discusses the significance of data pre-processing in E-learning. Finally, the conclusion and future directions are presented in Section 7.

2. Related works

Data mining methods have been used in certain research recently to aid administrators and educators in enhancing e-learning environments. In [3] the authors propose a new student performance model with a new feature group defined as behavioral attributes. This model uses some data mining techniques on such a data set to measure the impact of student behavioral characteristics on educational outcomes.

The authors in [4] focus on EDM tools and tools commonly used in EDM analyses rather than the broader universe of tools used in more traditional and modern statistical analyses. While in [5], the authors compared the effectiveness of four data cleaning methods on two actual data sets, and suggests a guideline for selecting data cleaning tools. In [6] the authors investigated several data pre-processing-related concerns to develop a manual or tutorial for teachers and EDM practitioners. In [7] a review on data mining for academic decision support in education field is presented. [8] Present big data applications in education and explore how it improves the education process.

To conclude, several studies have explored the adoption of data mining approaches to tackle educational issues. However, only a few of them have delved into the significance of educational data preprocessing.

[1,2,3,4] *Engineering, Systems and Applications Laboratory, ENSA, Sidi Mohamed Ben Abdellah University, Fez, Morocco*
[1]*ORCID ID: 0000-0003-3989-5203*
[4]*ORCID ID: 0009-0007-6437-107X*
** Corresponding Author Email: mounia.rahhali@usmba.ac.ma*

3. Big Data Concept

3.1. The Definition of Big Data

Big data is a huge and complex set of data, yet growing exponentially with time. It is a data with enormous size and complexity that could not be perceived, collected, controlled, and treated by tools and traditional data processing applications [9].

The common significant volume of data is produced by social data, machine data, and transactional data, but the percent of beneficial data is lessened, in comparison with other kinds of data sources that are more important, like governmental establishments, education establishments.

3.2. Characteristics of Big Data

The following features can characterize big data:

Volume: The name Big Data itself is relevant to a size that is large. Volume refers to the size of the data from Terabytes (TB) to Petabytes (PB) [10] and represents a highly important role in defining the sense out of data. Therefore, 'Volume' is one feature that requires to be regarded while dealing with Big Data.

Velocity: The term 'velocity' refers to methods of transferring big data including batch, near time, real-time, and streams. Velocity also refers to the speed of generation of data time. The data can be analyzed, treated, stored, and managed at a fast rate, or with a lag time between events. Stock exchanges and Weather reports are some of the real-time examples.

Variety: Variety of big data refers to various sources of data including structured, unstructured, semi-structured data. The data format can be in the form of photos, videos, emails, monitoring devices, audio, PDFs, etc.

Structured: Structured data is mostly classified data that is designed by columns and rows in a database. Databases that include tables in this shape are named relational databases.

Unstructured: is data that is not arranged in a pre-built way or does not have a pre-built data pattern. Videos, audio and binary data files might not have a special temple. They are ascribed to as unstructured data.

Semi-structured: is data that does not establish in a relational database but that has a few organizational features that make it simpler to resolve. Example: XML data, JavaScript Object Notation format and graph databases.

3.3. Big data applications in education

Thanks to Big Data and the increase of computing and digital technology in education, data analysis improving institution and education systems everywhere. This section presents a short overview of some applications of Big Data in the education area.

Performance Prediction

Student's success can be predicted by analyzing student's interaction in an education environment with other learners and instructors. The educational establishments are usually inquiring that how many learners will succeed/fail for required arrangements. Educational data mining is fully an efficient process for attaining this goal [11].

Data Visualization

Reports on educational data become more and more complicated as educational data increase in volume [12]. Data Visualization is a method for searching and analyzing digital data using graphs.

Learning materials recommendation

A recommender system is software that aids learners to distinguish the most suitable and appropriate learning items from a large number of items. It is designed to supply appropriate resources to a learner using certain user and resource information, by applying data mining methods and tools [13].

Types of Educational Environments

- Learning Management Systems

The learning management system notion began from e-Learning, is a software application for the management, documentation, tracking, and reporting of training programs, or learning and development programs. They also offer a large diversity of canals and workspaces to simplify knowledge participation and discussion between all the members in a course [6].

- Massive Open Online Courses

A massive open online course (MOOC) is an online course with the option of free and open registration. MOOCs have interactive user groups that enable teachers, instructors, and teaching assistants to create a community, in addition to conventional course resources such as animations, presentations, and videos.

- Intelligent Tutoring Systems

Intelligent Tutoring Systems (ITS) are instructional systems that produce immediate customized instruction or feedback to learners [6] using artificial intelligence (AI) techniques in machine programs to simplify learning. These systems are established on cognitive psychology as an underlying theory of learning [14].

- Adaptive and Intelligent Hypermedia Systems

One of the first and most prevalent types of dynamic hypermedia is adaptive and intelligent hypermedia systems (AIHS).It can be represented as a learning management system (LMS) or also an individual learning platform that adjusts the education to singular learner variances, such as cognitive skills, learning styles, emotional states, etc [15].

4. Knowledge Discovery And Data Mining

Knowledge Discovery in Databases (KDD) is the process of finding helpful, efficient, novel, and understandable knowledge from a collection of data.

The KDD process has attained its height in the latest 10 years. It presently houses several various methods to explore, which involve Bayesian statistics, inductive learning, knowledge acquisition for expert systems, semantic query optimization. The final purpose is to obtain important knowledge from low-level data. Artificial intelligence also helps KDD by finding empirical rules from observations and experimentation.

Data Mining (DM) is the gist of the KDD process, including the deducing of algorithms that explore the data, fusing previous knowledge on data sets, and rendering precise solutions from the found results. Presently, it is qualified as science and technology for exploring data to identify previously present unknown models [10].

4.1. The KDD Process

Problem Specification and Understanding

This is the first preliminary action and needs prior understanding and knowledge of the domain to be implemented in. It prepares the scene for understanding what should be done with the several decisions (about transformation, algorithms, representation, etc.). This assumption is very important which, if established incorrectly, can attend to incorrect interpretations and negative consequences on the end-user.

Data Selection

Having set the goals and objectives, the data collected needs to be selected and divided into significant sets based on availability, accessibility importance, and quality. These parameters are important for data mining because they build the base for it and will influence what sets of data models are created. If some essential characteristics are missing, then the whole study may fail [11].

Data Preprocessing

This step requires handling missing values and removal of noise or outliers in order to enhance the reliability of the data and its effectiveness. Certain algorithms are used for searching and eliminating unwanted data based on attributes specific to the application.

Data Mining

This is the gist process of the entire KDD where the methods are used to extract valid data patterns. This step involves the selection of the most proper DM type, for example, classification, regression, or clustering. This often depends on the KDD aims, and on the prior steps.

Evaluation

Once the aim and models have been taken from multiple data mining techniques, these models demand to be described in separated forms such as bar graphs, pie charts, histograms, etc. In this step, we evaluate and interpret the worked models, laws, and reliability to the goal defined in the first step.

Result Exploitation

In Knowledge display, knowledge is served to the user utilizing diverse knowledge illustration methods. In fact, the hit of this step defines the efficiency of the full KDD process.

5. Data Preprocessing

When we converse concerning data, we normally consider several massive datasets with a large number of rows and columns. That is a possible situation, it is not constantly the case; data could be in so several forms: Structured Tables, Audio, Images, Records, Videos, etc. Data preprocessing is a confirmed way of solving such problems. Data preprocessing is a data mining technique that includes converting raw data into an intelligible structure. [10] Differentiate between data preparation and data reduction due to the increased significance that the recent set of methods have been performing in recent years and some of the evident differentiation that can be extracted from this perception.

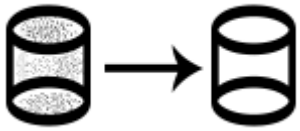
5.1. Data pre-processing steps

Data preparation

Data preparation is usually a necessary step. It transforms previously unusable data into new data that implements a DM method.

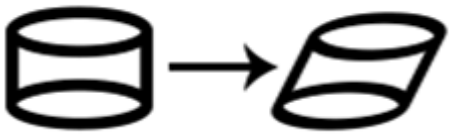
Data cleaning

Data cleaning or data cleansing involves methods of detecting and correcting (or removing) errors in data saved in databases or in files. The data existing in the databases can have different kinds of errors like typing errors, missing information, inaccuracies, etc. The inappropriate portion of the treated data can be replaced, changed, or removed.



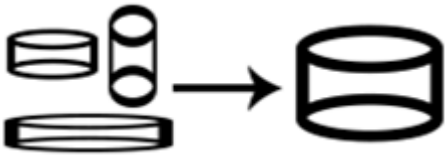
Data transformation

Data transformation is the process of converting data from one specific format or order to another one. Data transformation is important to transfer the data into a format that data mining can react with to obtain actionable insights.



Data integration

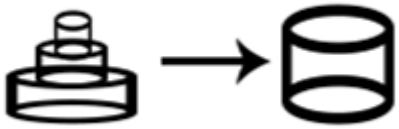
Data integration includes merging data residing in various sources and giving users a united vision of them. This method must be neatly presented in order to avert repetitions and disparities in the resulting data set.



Data normalization

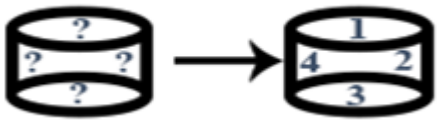
Data normalization gets rid of a numeral of exceptions that can make the analysis of the data more complex. Once these errors are come out and discarded from the system, more advantages can be achieved by other applications of data analytics.

Normalizing the data tries to present all features with similar weight and it is especially helpful in statistical learning techniques.



Missing data imputation

Several datasets may include missing values for several reasons. They are usually encoded as NaNs, blanks. In most cases, adding a rational estimate of a proper data value is more useful than leaving it blank. Training a model with a dataset that has a lot of missing values can drastically affect the machine-learning model's performance.



Noise identification

Are data with a huge volume of further unmeaning information in it named noise [10]. Its principal goal is to discover random errors or variations in a measured variable. It also involves any data that a user system cannot explain and understand perfectly.

Data reduction

In this instance of data reduction, the data generated normally keeps the fundamental structure and integrity of the initial data, only the quantity of data is decreased.

Feature selection

Feature selection is the method of decreasing the number of input variables when generating a predictive model. For [13], Feature selection is the concept of collecting a little subset of variables that ideally is important and adequate to define the aim concept.

Instance selection

Instance selection is the thought of choosing a subgroup from the group by keeping the underlying division undamaged so that the sampled data is a deputy of the features of the total data group. The random selection of examples is ordinarily perceived as Sampling and it is currently in the huge plenty of DM models for managing inside effectiveness and for avoiding over fitting [10].

Discretization

Discretization is one of the more utilized data pre-processing methods. It converts quantitative data into qualitative data by splitting the numerical characteristics into a restricted figure of uncorrelated intervals [16]. Discretization is the method of inserting values into buckets so that there are a restricted number of likely states.

Feature extraction

Feature extraction is a method of dimensionality decrease by which an original set of raw data is decreased to more flexible collections for handling.

5.2. Data Pre-processing problems in education

Missing data

In data, a missing value can indicate a variety of things. Maybe the data was unavailable or inapplicable, or perhaps the event did not happen. The user who filled the data may not have known the appropriate value or may have ignored it.

Missing values are handled differently by different data mining techniques. Missing values are usually ignored, excluded, replaced with the mean, or inferred from known values.

Manual input

Manual entry errors are troublesome since they might result in missing data as well as data inconsistencies.

Consider the case of a budget. Rather than utilizing exclusively numerical numbers, someone wishes to input the budget using other non-numerical symbols. To handle with this data, you must first eliminate these signs in order to convert it to numerical form.

Data inconsistency

When two objects (or tuples in the relational model) obtained from multiple data sources are recognized as versions of one another yet some of the values of their associated properties change, there is a data value level discrepancy [17].

Regional formats

When we receive raw data, we must determine the format in which the data is demonstrated. For example, if we are working with dates, what format should we use? Should it be day/month or day/month? Is a comma or a period used to demonstrate my decimal numbers?

Wrong data types

Databases are excellent at storing various kinds of data. This occurs for optimization purposes, as well as to avoid human error. As a result, we must exercise caution when attempting to add something to our database. For a human, "3" and "three" are the same thing, but not for a computer.

File manipulation

To view the contents of a file, we occasionally need to open it. Additionally, the software we use for it sometimes corrupts our data. With Excel, for instance, this happens. Excel has the ability to adjust dates, big figures, etc. When dealing with CSV and text format files, file manipulation might also become problematic.

Missing anonymization

Not every data is suitable for all purposes. In fact, before analysis, some data must be deleted or anonymised. Maintaining privacy, utilizing security, and preventing bias all depend on this.

5.3. Data Preprocessing tools

Open Refine

Previously named Google Refine is a standalone open-source desktop application for operating with messy data: cleaning it; transforming it from one format into another, it works on rows of data that have cells under columns, which is very alike to relational database tables. (https://openrefine.org/)

Import is supported from the following formats: TSV, CSV, JSON, XML, RDF triples, Google Spreadsheets. Export is supported in the following formats: TSV, CSV, HTML table, Google Spreadsheets.

Data Wrangler

"Wrangler" is a vital software and interactive tool used for data cleaning and transformation. It can operate in two modes: Users can either easily paste data into its web interface or export procedures as Python code using the web interface to handle large volumes of data in a random manner [5].

(http://vis.stanford.edu/wrangler/)

Rapid Miner's

RapidMiner is a data science platform designed for performing analytics. It provides a user-friendly graphical interface for creating analytics models and allows for the addition of code if needed.

It has restricted functionality for creating new features from existing ones (for example, creating multiplicative interactions) and for feature extraction [4]. By Rapid Miner, one can obtain different data sources, make data preparation, data cleansing, compute detailed statistics represented with graphs, and make predictive statistical analysis. (https://rapidminer.com/)

WEKA

Weka is a data-mining tool developed in java by the University of Waikato. It has many advantages such as free of charge, portability, ease of use and finally a large set of machine learning models such as Neural Networks, Decision Trees, or even K-means. Weka enables the implementation of many statistical analysis and machine learning techniques, from the pre-processing step to the prediction step. (http://www.cs.waikato.ac.nz/ml/weka/)

SPSS

SPSS (Statistical Package for the Social Sciences) is software applied for statistical analysis. SPSS modeller has functionality for generating novel features out of existing features, for feature selection, data filtering and feature space reduction [4] (https://www.ibm.com/fr-fr/products/spss-statistics).

KEEL

KEEL (Knowledge Extraction based on Evolutionary Learning) is a free and open source Java software tool that can be employed for a variety of knowledge data discovery activities. To construct experiments with various datasets and computational intelligence algorithms and to evaluate the performance of the algorithms, KEEL offers a simple GUI based on data flow. It includes a diverse set of conventional knowledge extraction algorithms, preprocessing methods, machine learning based on artificial intelligence, hybrid models, statistical approaches for contrasting experiments, and more (http://www.keel.es/).

Python

The open source programming language that computer scientists employ the most is Python. In the areas of software development, data analysis, and infrastructure management, this language has emerged as a leader.

R

R is a programming language and open-source data science and statistics software supported by the R Foundation for Statistical Computing.

6. Importance Of Data Preprocessing In E-Learning

A survey conducted by CrowdFlower, a company that offers a "data enrichment" platform for data scientists, revealed that approximately 80 data scientists were surveyed, and the results showed that they devote a significant amount of time to the following activities:

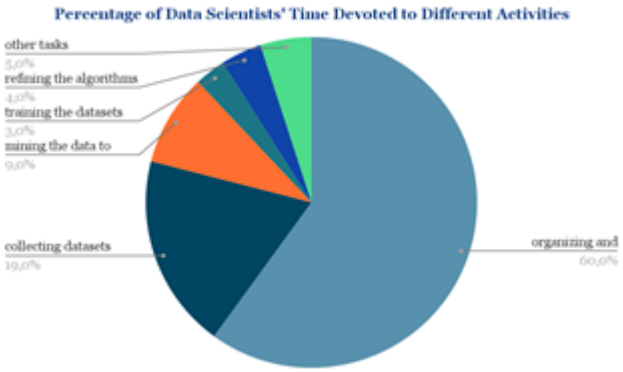


Fig 1: Persentage of Data Scientits Time Devoted to Diferent Activities

Data preprocessing becomes extremely important when it comes to implementing various machine-learning algorithms. Since data preprocessing can significantly impact the results of the learning model, it is crucial that all attributes are on an equal scale. The main goal of data preprocessing is to reduce or remove any small data anomalies that may be associated with experimental error, especially if they occur regularly [18].

Indubitably, preprocessing of educational data is usually very alike to the preprocessing assignment in other fields. Nevertheless, it is significant to indicate that the preprocessing of educational data has specific properties that distinguish it from data preprocessing in different particular fields like:

- Learning management systems, such as Moodle, keep a large number of features of learners, courses, and exercises [6]. This quantity of attributes can result in decreasing the efficiency of a learning model due to over fitting issues. One solution to this problem is to choose just the most significant features.

- Educational data are usually collected from different sources such as (enrollment, quiz, course activity statistics, and Participation in discussions) since they have been created in

several places at various times.

● Usually, the learners do not finish all the exercises, tests, etc. As a result, there are oftentimes missing data values. [19] Distinguished three main sources of missing data values in research: no coverage, total nonresponse, and item nonresponse.

● No coverage which happens when some parts of the community of inference are not incorporated in the study sampling stage and therefore have no opportunity of being chosen in the pattern.

● Total non-response happens when a sampled element does not engage in the study.

● Item nonresponse happens when a responding sampled component failure to give satisfactory answers.

Finally, the following key instructions can be drawn regarding the pre-processing of educational data [6] :

Pre-processing is always a crucial first step in any application or process involving data mining. This activity is crucial since the produced DM models' interest, value, and usability are directly related to the caliber of the data employed.

There are many usable pre-processing operations and various learning contexts, which provide various forms of data.

It is not necessary to use all pre-processing tasks or stages in every situation, as their application may or may not be essential depending on the data and problem at hand. While there is no fixed formula or guideline for each pre-processing task, multiple methods can be utilized in each phase. The user must make decisions on which methods to employ based on various factors, such as the data characteristics, available techniques and algorithms, and the ultimate goal of the data-mining problem to be solved.

7. Conclusions

In conclusion, this mini review has highlighted the importance of pre-processing educational big data in data mining. Pre-processing is a crucial step in data mining as it helps to transform raw data into a more useful format for analysis. By using appropriate pre-processing techniques, researchers can ensure that the data they are analyzing is accurate, complete, and consistent. In addition, pre-processing can also help to reduce the amount of noise and irrelevant information in the data, which can improve the accuracy of the analysis. Overall, the use of pre-processing techniques in educational big data mining has the potential to enhance our understanding of how students learn and how educational institutions can improve their teaching methods.

Our next work will require applying data mining techniques on an educational data set with more distinctive attributes to get more accurate results.

Declarations

- Ethical Approval: 'Not applicable'
- Informed Consent: 'Not applicable'
- Statement Regarding Research Involving Human Participants and/or Animals: 'Not applicable'
- Consent to Participate: 'Not applicable'
- Consent to Publish: 'Not applicable'
- Funding: 'Not applicable'
- Author's Contribution: 'Not applicable'
- Competing Interests: 'Not applicable'
- Availability of data and materials: 'Not applicable'

References

[1] R. A. Huebner, « A survey of educational data-mining research », p. 13, 2013.

[2] C. Silva et J. Fonseca, « Educational Data Mining: A Literature Review », in *Europe and MENA Cooperation Advances in Information and Communication Technologies*, vol. 520, Á. Rocha, M. Serrhini, et C. Felgueiras, Éd., in Advances in Intelligent Systems and Computing, vol. 520. , Cham: Springer International Publishing, 2017, p. 87-94. doi: 10.1007/978-3-319-46568-5_9.

[3] E. A. Amrieh, T. Hamtini, et I. Aljarah, « Preprocessing and analyzing educational data set using X-API for improving student's performance », in *2015 IEEE Jordan Conference on Applied Electrical Engineering and Computing Technologies (AEECT)*, Amman, Jordan: IEEE, nov. 2015, p. 1-5. doi: 10.1109/AEECT.2015.7360581.

[4] S. Slater, S. Joksimović, V. Kovanovic, R. S. Baker, et D. Gasevic, « Tools for Educational Data Mining: A Review », *J. Educ. Behav. Stat.*, vol. 42, nº 1, p. 85-106, févr. 2017, doi: 10.3102/1076998616666808.

[5] S. Oni, Z. Chen, S. Hoban, et O. Jademi, « A Comparative Study of Data Cleaning Tools »:, *Int. J. Data Warehous. Min.*, vol. 15, nº 4, p. 48-65, oct. 2019, doi: 10.4018/IJDWM.2019100103.

[6] C. Romero, J. R. Romero, et S. Ventura, « A Survey on Pre-Processing Educational Data », in *Educational Data Mining*, vol. 524, A. Peña-Ayala, Éd., in Studies in Computational Intelligence, vol. 524. , Cham: Springer International Publishing, 2014, p. 29-64. doi: 10.1007/978-3-319-02738-8_2.

[7] S. Suhirman, T. Herawan, H. Chiroma, et J. Mohamad Zain, « Data Mining for Education Decision Support: A Review », *Int. J. Emerg. Technol. Learn. IJET*, vol. 9, nº 6, p. 4, déc. 2014, doi: 10.3991/ijet.v9i6.3950.

[8] S. Khan et S. Alqahtani, « Big Data Application and its Impact on Education », *Int. J. Emerg. Technol. Learn. IJET*, vol. 15, nº 17, p. 36, sept. 2020, doi: 10.3991/ijet.v15i17.14459.

[9] M. Rahhali, L. Oughdir, et Y. Jedidi, « E-Learning Recommendation System for Big Data Based on Cloud Computing », *Int. J. Emerg. Technol. Learn. IJET*, vol. 16, nº 21, p. 177, nov. 2021, doi: 10.3991/ijet.v16i21.25191.

[10] B. Furht et F. Villanustre, « Introduction to Big Data », in *Big Data Technologies and Applications*, Cham: Springer International Publishing, 2016, p. 3-11. doi: 10.1007/978-3-319-44550-2_1.

[11] Y. Jedidi, A. Ibriz, M. Benslimane, M. Tmim, et M. Rahhali, « Predicting Student's Performance based on Cloud Computing », in *WITS 2020*, vol. 745, S. Bennani, Y. Lakhrissi, G. Khaissidi, A. Mansouri, et Y. Khamlichi, Éd., Springer Singapore, 2022. doi: 10.1007/978-981-33-6893-4_11.

[12] Open University Malaysia, K. Sin, L. Muthu, et Bharathiyar University, « APPLICATION OF BIG DATA IN EDUCATION DATA MINING AND LEARNING ANALYTICS – A LITERATURE REVIEW », *ICTACT J. Soft Comput.*, vol. 05, nº 04, p. 1035-1049, juill. 2015, doi: 10.21917/ijsc.2015.0145.

[13] M. Rahhali, L. Oughdir, Y. Jedidi, Y. Lahmadi, et M. Z. El Khattabi, « E-learning Recommendation System Based on Cloud Computing », in *WITS 2020*, vol. 745, S. Bennani, Y. Lakhrissi, G. Khaissidi, A. Mansouri, et Y. Khamlichi, Éd., Springer Singapore, 2022. doi: 10.1007/978-981-33-6893-4_9.

[14] I. Padayachee, « Intelligent Tutoring Systems: Architecture and Characteristics », 2002.

[15] M. Tmimi, M. Benslimane, M. Berrada, et K. Ouazzani, « Intelligent Model Conception Proposal for Adaptive Hypermedia Systems », *Int. J. Adv. Comput. Sci. Appl.*, vol. 9, nº 8, 2018, doi: 10.14569/IJACSA.2018.090825.

[16] S. García, J. Luengo, et F. Herrera, *Data Preprocessing in Data Mining*, vol. 72. in Intelligent Systems Reference Library, vol. 72. Cham: Springer International Publishing, 2015. doi: 10.1007/978-3-319-10247-4.

[17] X. Wang, L.-P. Huang, X.-H. Xu, Y. Zhang, et J.-Q. Chen, « A Solution for Data Inconsistency in Data Integration », p. 15, 2011.

[18] K. Kira et L. A. Rendell, « A Practical Approach to Feature Selection », in *Machine Learning Proceedings 1992*, Elsevier, 1992, p. 249-256. doi: 10.1016/B978-1-55860-247-2.50037-1.

[19] J. Brick et G. Kalton, « Handling missing data in survey research », *Stat. Methods Med. Res.*, vol. 5, nº 3, p. 215-238, sept. 1996, doi: 10.1177/096228029600500302.