encyclopedia

MDPI

Entry

Educational Data Mining: A Foundational Overview

Ilias Papadogiannis [1,*], Manolis Wallace [1] and Georgia Karountzou [2]

[1] ΓΑΒ LAB—Knowledge and Uncertainty Research Laboratory, University of the Peloponnese, 22131 Tripolis, Greece; wallace@uop.gr
[2] Directorate of Primary Education of Arcadia, 22131 Tripolis, Greece; gkarountzou@uop.gr
* Correspondence: i.papadogiannis@go.uop.gr

Definition: Educational data mining (EDM) is a novel scientific area that focuses on developing and applying methods to analyze datasets generated within educational settings. This paper outlines the evolution, significance, and applications of EDM. With the increasing popularity of e-learning in web-based educational systems, EDM has expanded to include a variety of analytical methods and data sources. Some key methodologies addressed include classification, regression analysis, clustering techniques, association rule mining, and Natural Language Processing, among others. Additionally, this paper looks at how EDM can facilitate data-driven decision-making among other areas such as curriculum development and customization of learners' experiences. It also touches on issues related to the challenges of the scientific field. Finally, some projections about EDM's future trends are made, especially concerning its integration into AI technologies and development trends like augmented reality or virtual reality, which imply greater possibilities for changes than any other series witnessed before within this sphere.

Keywords: educational data mining; education; algorithms

Citation: Papadogiannis, I.; Wallace, M.; Karountzou, G. Educational Data Mining: A Foundational Overview. *Encyclopedia* **2024**, *4*, 1644–1664. https://doi.org/10.3390/encyclopedia4040108

Academic Editor: Bin Dong

Received: 10 September 2024
Revised: 23 October 2024
Accepted: 24 October 2024
Published: 31 October 2024

1. Evolution of EDM

Educational data mining is a fairly new discipline that aims to develop new techniques for examining datasets obtained from the educational environment and applying these techniques in order to shed new light on students and educational settings. Over the past decades, EDM has expanded significantly, reflecting its increasing significance in the field of education. Initially, it involved the use of data mining techniques on educational data to answer some important questions [1]. However, with e-learning growth along with web-based education systems emergence, EDM's scope has expanded and now covers a wide range of data sources as well as methods [2]. Consequently, it has given rise to more advanced models and approaches for analyzing student behavior and learning outcomes [3].

The development of EDM is characterized by an interdisciplinary approach. It integrates machine learning techniques, didactics, and cognitive psychology, among others; thus, it facilitates a more holistic understanding [4]. This interdisciplinary nature has allowed EDM to address complex challenges, such as personalized learning and the predictive modeling of student performance. A key milestone in the development of EDM was the organization of the first educational data mining conference in 2008, which provided a platform for researchers to share information and developments in the field [5]. Since then, the number of conferences, publications, and related research has grown exponentially, highlighting the growing academic and practitioner interest in EDM.

New terms and subfields have emerged alongside EDM, such as learning analytics (LA), academic analytics, and big data in education, reflecting the increasing focus and application of data analytics in educational contexts [6]. While EDM is mainly concerned with technological challenges, LA focuses on data-driven decision-making from what is being taught and integrates social and pedagogical dimensions [7]. Academic Analytics

(AcAn), on the other hand, focuses on the use of broader data with the aim of using it mainly for administrative and decision-making purposes. Finally, the use of big data, which is now available, allows for the integration of large amounts of data from gamification and virtual reality tools [8].

2. The Significance of EDM in Education

EDM is very important for modern educational systems, and it is also vital because it makes it possible for raw data to be transformed into actionable information. According to Refs. [9,10], there has been an increase in the volumes of educational data due to e-learning platforms and Internet-based education systems, leading to the creation of large databases. The volume complexities involved in this type of data make it impossible to gain any meaningful understanding via conventional means. By making use of cutting-edge tools, the process may be made easier, leading to improved precision besides increasing the speed at which analyses are performed, thus promoting more informed decision-making alongside strategic planning [11].

In an organizational context, EDM has revealed the knowledge hidden deep within their data. The importance of EDM in teachers' understanding of students' learning processes is very significant [3]. An example is the examination of students' behavioral patterns in classrooms and how they interact with instructional materials. This develops more adaptive approaches to the learning process. A very important contribution is the prediction of students' academic performance based on specific individual characteristics; prediction models that integrate academic, social, and educational data help teachers identify students who are likely to drop out early and provide them with immediate assistance, thus improving overall academic outcomes [7].

Another important contribution of EDM is its ability to provide support for data-driven decision-making. In this way, educational institutions can make informed decisions. Better decisions are expected to have an impact on the optimization of educational processes and student success [6]. This approach is not only limited to improving academic performance but also to improving the overall educational environment [12].

Furthermore, EDM provides teachers and academic authorities with feedback, which, in turn, supports improvements in curriculum design, methods, and resource allocation. These kinds of feedback loops lead to building an education system that effectively responds to the increasing demands of both students and teachers.

Further, the importance of EDM in promoting a tradition of constant enhancement within learning institutions is beyond any doubt. In terms of understanding educational processes in a more detailed manner, EDM helps find areas that are not efficient or effective and therefore provides specific remedy actions. This will lead to better education while making sure that resources are managed as efficiently as they can be. Consequently, this allows EDM to nurture a fairer and more proficient education system based on data rather than hunches or customs [1].

The tools provided by EDM enable teachers and schools to comprehend, anticipate, and enhance students' education. This enables one to make use of data and transform it into decisions, thus leading to individualized, prosperous, and efficient schools. In times to come when education will be enriched using educational data, EDM will be an integral part.

3. EDM Literature Reviews

The scientific field has been reviewed in the literature several times. Few of these, however, provide systematic reviews of the entire scientific discipline. According to Papadogiannis [13], some important reviews have been conducted in previous decades, as listed in Table 1. In the following years, literature reviews were published, and from the examination of which, conclusions can be drawn about the latest developments in the scientific field of EDM [13,14]. In order to present the methods, thematic categories, data, and applications of the EDM discipline in this introductory presentation, data from the two above recent literature reviews have been used [13,14].

Table 1. EDM Literature reviews.

Reference	Year	Findings
Romero and Ventura [15]	2007	• Presents the majority of relevant work since 2005. • Classifies surveys based on their objective. • Enrollment numbers are increasing each year. • Possible future developments in the field are identified. • Integrating data mining in education is crucial for both researchers and external users.
Baker and Yacef [16]	2009	• Identifies an increasing pace of expansion in the field. • Recognizes a rise in available data. • Illustrates the shift in the objectives and techniques used. • Emphasizes the importance of open educational data and the significant room for growth (EDM).
Romero and Ventura [5]	2010	• Categorizes potential users of effects. • Categorizes tasks based on applied techniques and their goals. • Highlights key future issues. • Ease of use for non-specialist effects in decision-making. • Need for integration into simple environments. • Standardization of data and models.
Papamitsiou and Economides [17]	2014	• Papers categorized by methods, goals, and learning settings. • Major directions of the field are highlighted. • Suggestion for integrating technologies like game-based and mobile learning. • SWOT analysis conducted.
Pena-Ayala [18]	2014	• Detailed presentation of articles by method, algorithms, and objectives was made. • Two different approaches to educational data mining are identified. • SWOT analysis conducted. • A great potential was identified due to the widespread use of information systems in education.
Thakar, Mehta, and Manisha [19]	2015	• Key areas of research identified: ○ Identifying weak students and predicting student failure. ○ Assessing students in specific courses. ○ Assessing students' understanding. • The lack and need for a unified approach was highlighted.
Sukhija et al. [20]	2015	• Presents tools, techniques, and outcomes of studies from 2001 to 2015. • Identifies challenges for educational data mining (EDM): ○ Lack of coherent data sets at the overall education system level. ○ Lack of flexibility in data sets. ○ Lack of trust from educational authorities in EDM results. • Highlights that most research was limited to small-scale experiments.
Del Rio and Insuasti [21]	2016	• Focused on predicting students' academic performance in higher education (2011–2016). • Presented the main methods used by authors. • Highlighted key predictors of academic performance. • Listed the software tools used.

The literature review of Ozyourt [22] shows an exponentially increasing trend of research on EDM. According to Romero and Ventura [23], more than 6000 articles have been published on learning analytics and more than 3000 on EDM, which shows the growing interest of researchers in educational data analysis [24]. The exponential increase in the number of articles published on EDM and LA is also evidenced by the data we extracted from the Crossref database. We searched for "education data mining" AND "learning analytics" in the time period 2000–2024 and exported the results in a spreadsheet.

The results were then cleaned up, and those that were relevant to EDM. The exponential growth of articles as well as a decline in the number of articles during the COVID-19 crisis is shown in Figure 1. In simpler terms, this trend shows that the quantity of EDM and LA research has been increasing at an exponential pace over time. Each year's growth builds on the previous year's studies at an accelerating rate, captured by the exponentiation of the constant 0.1646. This has resulted in the number of studies multiplying rapidly from year to year. The high R^2 value of 0.89 indicates that this exponential function fits the data extremely well, suggesting that this pace of expansion is likely to continue in the near future. This rapid growth reflects the increasing importance, acceptance, and application of data mining and analytics approaches in education. Even challenges such as the COVID-19 crisis have not been able to slow the long-term upward trend.

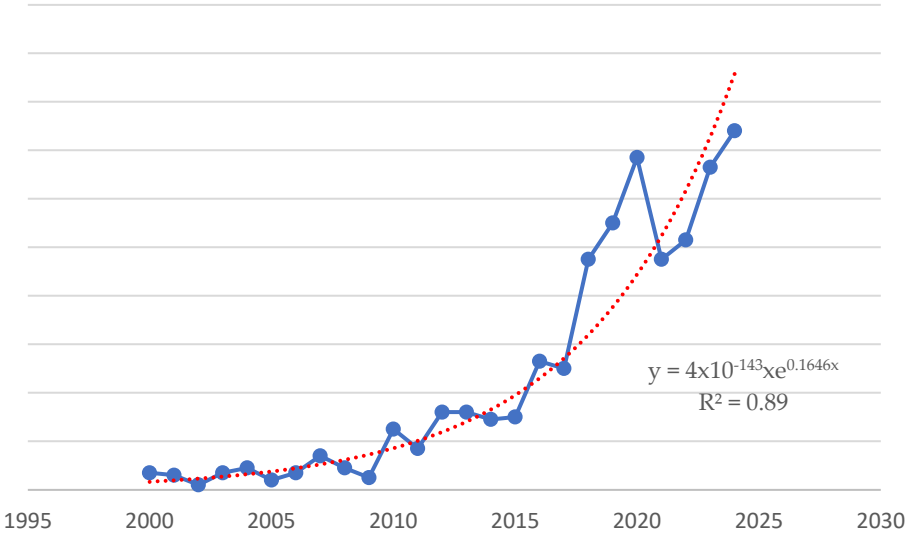

Figure 1. Number of EDM and LA studies from Crossref database.

4. EDM Data and Tools

EDM uses different types and sources of data, including those collected for purposes of research or those obtained from e-learning management platforms as well as MIS systems. Traditionally, sources such as students' attendance records, grades, and demographic details provide insight into educational developments [1,25]. The increasing use of electronic learning platforms has led to massive amounts of data produced by students while they interact with digital platforms. Learning management systems (LMS), massive open online courses (MOOCs), and intelligent tutoring systems (ITS) record learners' actions during online courses, results achieved, etc. [23]. Such information is important for drawing conclusions about student participation and learning patterns.

New educational settings like virtual and augmented reality and gamification tools are some of the additional sources of data. These settings track pupil movements or activities, allowing for a new way of looking at how learners gain knowledge in modern learning environments. Educational data can be divided into three levels [13]:

- Micro-level data in education are mainly generated through interactions between students and learning platforms, such as MOOCs, simulations, and games. This data can capture detailed learner actions and contexts, enabling real-time interventions like feedback or skill adjustment. It is validated through real-time observations or retrospective coding [23,24]. Common methodologies, including Bayesian Knowledge Tracing and Performance Factor Analysis, are used to assess students' knowledge and predict outcomes [25].
- Meso-level data are mainly generated through student texts in platforms like LMS and social media. Natural Language Processing (NLP) helps examine students' cognitive, social, behavioral, and emotional processes. It supports automated grading and feedback,

improves course design, and enhances student participation, though challenges remain in tool reliability and contextual factors [16].

- Macro-level data are collected over longer periods and include demographics, course enrollments, and academic records. These data are primarily used for institutional decision-making, supporting early warning systems that identify at-risk students and guidance systems that recommend courses [26,27]. Macro-level data are also utilized for administrative analyses to assess curriculum effectiveness and patterns of student success or dropout [28].

This hierarchical nature allows for analysis at multiple (micro, meso, and macro) levels, and overall can provide a comprehensive picture of the learning process [12]. The complexity and volume of educational data require sophisticated methods for cleansing and pre-processing them to produce reliable results [10]. While these diverse data sources are a wealth for analysts, they also present challenges in terms of pre-processing. Data must be collected, cleaned, and anonymized in a way that ensures privacy and accuracy.

The data utilized in EDM can be broadly categorized into several types, including demographic data, interaction data, performance data, and psychometric data. Each type serves a specific purpose in understanding and enhancing the educational process. The following is an attempt to group educational data based on their type [2]:

- Demographic data refer to ways in which students are divided by the information about their backgrounds (e.g., their age, gender, socio–economic status, and educational history). Demographic data are generally used to uncover some regularities and interconnections between sociodemographic features and the academic success of students. For example, the researchers can decide to investigate how the poverty rate affects the availability of education resources or what role the demographic aspects have in the level of a student's performance [29]. This piece of information is crucial for building personalized learning experiences that highlighting the connection between the students and educators.
- Interaction data are data that students receive from their interactions with different educational technologies; for instance, learning management systems, online courses, and educational software [30]. These data contain specific information about how often, for how long, and in what style students interact with a digitized learning process. Interaction data are a crucial part of the understanding of how students actually interact with e-learning, which can thus be helpful in the design of more effective instructional materials and interventions [31]. Thus, from an LMS, clickstream data analysis is a powerful tool that can allow educators to figure out which study resources are important to students; hence, the right content delivery can be achieved.
- Performance data are grades, scores, and other assessments that tell us how well a student performed. This class of data can determine the effects of the educational programs on students and look where the children need the most help [23]. Performance data can be collected from traditional assessments, such as exams and quizzes, as well as from more dynamic sources such as real-time analytics from online learning platforms [31]. The interpretation of student performance data calls for the provision of quality performance measures that will be used in decision-making and predicting the student's achievement.
- Psychometric data cover measuring students' cognitive abilities, their personality traits, and their emotional states. This kind of data is typically gathered through questionnaires, psychological assessments, and observation studies [30]. Psychometric data forms a basis for understanding the latent variables that on the one hand make learning happen in the first place, e.g., motivation, self-efficacy, and stress levels [29]. Other than providing the data of psychology, the data of other educational types come from the researchers and make possible the buildup of models shown by the students and the processes of the training.

According to Choi [14], data used in the studies are mainly academic (including grades), behavioral, and demographic, among others. The frequencies of use of EDM data types are shown in Table 2.

Table 2. EDM data types.

Type	Percent
Academic performance data	36.2%
Behavioral interaction data	20.3%
Programming data	20.3%
Demographic data	11.6%
Contextual data	10.1%
Psychometric data	1.4%

Today, there are a variety of tools available for conducting EDM research, including Rapidminer, Weka, SPSS, Knime, Orange, and Spark Lib, as well as programming languages such as R and Python. However, educators find it challenging to use these tools since it is tough to identify suitable methods and parameters. The lack of competence is to be expected given that most cases require specialist knowledge, with the exception being perhaps visualization. In addition, researchers can now access a wide range of open datasets online (Table 3). Although these data sets address specific issues, they are extremely beneficial for researchers and those seeking to further their understanding of the sector.

Table 3. EDM datasets available.

Datasets	URL	Description
ASSISTments Data Set 2012–2013	https://sites.google.com/view/assistmentsdatamining/home (accessed on 5 October 2024)	Competition data set using real-world educational data.
Canvas Network dataset	https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/1XORAL (accessed on 5 October 2024)	Data from Canvas Network
DataShop	https://pslcdatashop.web.cmu.edu/index.jsp?datasets=public (accessed on 5 October 2024)	A large repository of learning interaction data.
Educational Process Mining Dataset	https://archive.ics.uci.edu/dataset/346/educational+process+mining+epm+a+learning+analytics+data+set (accessed on 5 October 2024)	Students' logs from activities through a logging application while learning with an educational simulator.
HarvardX-MITx dataset	https://dataverse.harvard.edu/dataverse/mxhx (accessed on 5 October 2024)	Deidentified student-level data from the first year of HarvardX and MITx courses.
KDD Cup 2010 Dataset	https://pslcdatashop.web.cmu.edu/KDDCup/ (accessed on 5 October 2024)	Data from an education data mining challenge in 2010.
Educational Data Set Prize	https://educationaldatamining.org/data-set-awards/ (accessed on 5 October 2024)	Contains data about courses, students, and their interactions with Virtual Learning Environment.
NUS Multisensor Presentation Dataset	https://scholarbank.nus.edu.sg/handle/10635/137261 (accessed on 5 October 2024)	It contains real-world presentations recorded in a multisensory environment.
Open University Learning Analytics Dataset	https://analyse.kmi.open.ac.uk/open_dataset (accessed on 5 October 2024)	It contains data about courses, students, and their interactions with Moodle for seven selected courses.
Student Performance Dataset	https://archive.ics.uci.edu/ml/datasets/Student+Performance (accessed on 5 October 2024)	Student achievement data in secondary education of two Portuguese schools.
xAPI-Educational Mining Dataset	https://www.kaggle.com/aljarah/xAPI-Edu-Data (accessed on 5 October 2024)	Academic performance dataset from e-learning system.

5. EDM Methods

EDM relies on a range of techniques, tools, and methods. These techniques are basically DM techniques adapted to the specific characteristics of educational data. Below are the main methods used by researchers.

Classification is a supervised machine learning method that aims to predict Y based on a set of explanatory variables X. The target categorical variable Y can take on one of c preset values. A predictive classification function g classifies each X into one of the defined classes. The overall objective, as with any supervised learning method, is to minimize the expected loss or risk through g by leveraging labeled training data to learn patterns between X and Y. The expected loss or risk is given by:

$$l(g) = \text{ELoss}(Y, g(X))$$

where Loss $(y, \hat{y})$ is some loss function that quantifies the impact of classifying a response $y = g(x)$. In case there is no loss for a correct classification and a unit loss for a misclassification, the optimal classifier is the following:

$$g(x) = \text{argmax } P[Y = y|X = x]$$

There are many different classification algorithms. Some of the most commonly used in EDM are [22]:

- Decision tree is a tree-like structure where each internal node represents a decision based on a feature, each branch represents an outcome of that decision, and each leaf node represents a class label. The tree is constructed by recursively splitting the data based on the feature that best separates the classes at each step, usually using metrics like Gini index or information gain.
- The Naive Bayes classifier is based on Bayes' Theorem and assumes that features are independent of each other (hence the term "naive"). Despite this unrealistic assumption, Naive Bayes works well in practice, especially for text classification tasks like spam detection.
- Support vector machines aim to find a hyperplane that best separates the data points of different classes in a high-dimensional space. SVMs are particularly effective when the data are linearly separable. In cases where the data are not linearly separable, SVM uses a kernel trick to transform the data into a higher dimension where separation is possible.
- The KNN algorithm is an instance-based learning method where the classification of a new data point is determined by the majority class among its K-Nearest Neighbors in the feature space. It relies on distance metrics like Euclidean distance to measure the similarity between data points.
- Random Forest is an ensemble learning method that combines multiple decision trees to improve classification accuracy. Each tree is trained on a random subset of the data, and the final prediction is made based on the majority vote of the individual trees. This approach helps mitigate the overfitting problem associated with single decision trees.
- Neural networks are inspired by the biological neural networks of the human brain. They consist of layers of interconnected nodes (neurons) that process input features to make predictions. Deep learning models, which are a type of neural network with many hidden layers, have gained popularity for complex classification tasks like image recognition and Natural Language Processing.

Evaluating the performance of the models is crucial in data mining. Some common evaluation metrics of classification algorithms are as follows: (a) Accuracy, which is the ratio of correctly predicted instances to the total instances; it is simple to compute but may not be appropriate when the data is imbalanced. (b) Precision and recall are also important, with precision measuring the proportion of true positives among all predicted positives, and recall measuring the proportion of true positives among all actual positives. (c) F1 score, which is the harmonic mean of precision and recall, provides a balanced measure that works well for imbalanced datasets. (d) The Receiver Operating Characteristic (ROC) curve, which plots the true positive rate against the false positive rate, and the Area Under the Curve (AUC) are used to summarize the model's performance.

In the context of EDM, classification can be used to predict student performance levels, identify at-risk students, or categorize students according to their learning style [24]. Some common algorithms in this regard are decision trees, Random Forests, and support vector machines. Decision trees are popular because they are easier to interpret and extract rules, which help educators understand the factors influencing predictions [7]. Figure 2 presents a decision tree based on four levels of performance of secondary school students in Greece. The tree predicts the level of performance in the third grade, based on the first and second grade levels [21].

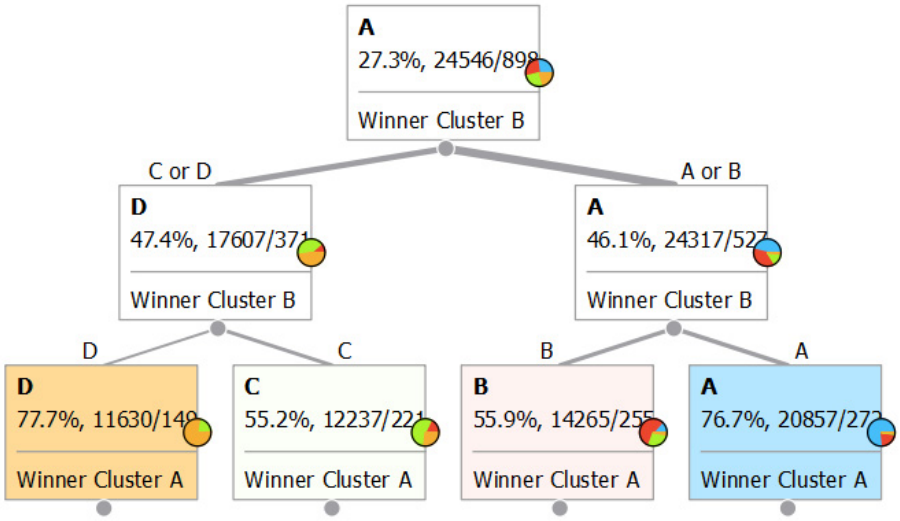

Figure 2. A decision tree, based on Greek students' secondary educational dataset, using Orange data mining software, Version: 3.35.0 [21], (Note: The presentation of pie charts rather then specific numbers is the focus of this figure).

Classification is usually used more often than regression, mainly due to the ease of implementation of the algorithms and their satisfactory accuracy. The application rate of classification algorithms ranges from 26.2% to 48.7% [13]. The higher rate is reported by Choi [14], where 68.8% of the total number of studies reviewed used classification algorithms.

Common classification algorithms employed in EDM include decision trees, Random Forests, support vector machines (SVMs), Naive Bayes, and K-Nearest Neighbors (KNN). Each algorithm offers unique strengths. Decision trees are favored for their interpretability, allowing educators and academic staff to understand the factors influencing predictions and extract actionable rules. SVMs are particularly effective for binary classifications like pass/fail predictions [13].

- Regression is a basic technique in supervised learning used to model the relationship between dependent and independent variables to predict continuous variables. More specifically, regression tries to predict a variable y using a function $g(x)$, where x is an explanatory vector $x = [x_1, \ldots, x_p] \top$. The optimal function g^* has to be learned from the training set by minimizing the training loss:

$$l(g) = \frac{1}{n}\sum_{i=1}^{n}(y_i - g(x_i))^2$$

Some of the most frequent types of regression are as follows: (a) Linear regression, where the relationship between the dependent and independent variables is linear. In the simplest case, the relationship is estimated from a straight line with one explanatory variable. (b) Polynomial regression is a form of regression where the relationship between the independent and dependent variables is modeled as an nth-degree polynomial. (c) Logistic regression is used when the dependent variable is binary (0,1). (d) Ridge regression is a

variation of linear regression that introduces a regularization term (penalty) to the loss function to prevent overfitting. (e) Lasso regression is similar to ridge regression but uses a different penalty (absolute value of coefficients). (f) Elastic Net regression combines the properties of both Ridge and Lasso regression. Common evaluation metrics include:

- Mean Squared Error (MSE) calculates the average of the squared differences between predicted and actual value.

$$MSE = \frac{1}{N} \times \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

where y is the actual value of the predicted variable and $\hat{y}$ is the estimation of y form the model.

- Root Mean Squared Error (RMSE) is the square root of MSE, providing error estimates in the same units as the target variable.

$$RMSE = \sqrt{\frac{1}{N} \times \sum_{i=1}^{n} (y_i - \hat{y}_i)^2}$$

Mean Absolute Error (MAE) measures the average absolute differences between predicted and actual values, offering a more interpretable measure of error without penalizing large errors as severely as MSE.

$$MAE = \frac{1}{N} \sum_{i=1}^{n} |y_i - \hat{y}_i|$$

R-squared (R^2) evaluates the proportion of the variance in the dependent variable that the model explains, with values closer to 1 indicating better model fit.

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \overline{y}_i)^2}$$

Additionally, Adjusted R-squared adjusts for the number of predictors in the model, providing a more accurate measurement when dealing with multiple independent variables. An example of a simple regression model is presented in Figure 3.

Regression is used in educational data mining to predict outcomes such as student grades or time spent on assignments and more. Several types of regression algorithms are used, such as linear regression, which assumes a direct linear relationship between the input and output variables, and polynomial regression, which adjusts a polynomial curve to capture more complex relationships. The ridge regression adds a penalty term to reduce overfitting, particularly in cases of multilinearity, while support vector machine (SVM) regression adapts the support vector classification approach for continuous forecasting or logistic regression for binary outcomes [1,13,14].

In EDM studies, regression has been used in about 18.8% [14] of studies or even less [13]. Among the algorithms used, linear regression was the one that performed better than other regression types and also became the preferred choice. Cross-validation techniques such as 5-fold and 10-fold cross-validation were often used to validate the accuracy of the models [13].

- Clustering—an unsupervised learning technique—classifies similar instances into groups without predetermined labels. Clustering techniques aim to minimize within-cluster variance by iteratively assigning data points to the nearest cluster center. Clustering methods can be broadly categorized into five types: Partitioning methods, such as K-means and K-medoids, aim to divide data into a predefined number of clusters. K-means assigns points to clusters based on their proximity to centroids, which are iteratively refined, while K-medoids are more robust for outliers. Hierarchical methods build

a tree-like structure (dendrogram) of clusters through either agglomerative (bottom-up) or divisive (top-down) processes.

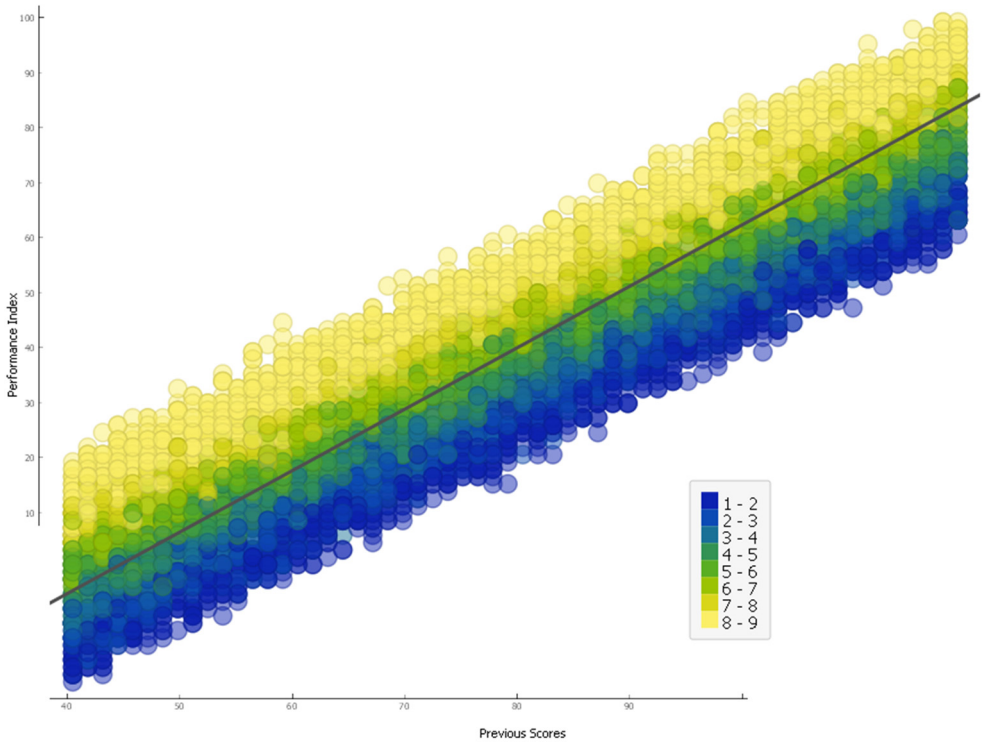

Figure 3. Simple linear regression, students' performance dataset, using Orange data mining software, Version: 3.35.0 (dataset available on: https://www.kaggle.com/datasets/nikhil7280/student-performance-multiple-linear-regression?resource=download (accessed on 23 October 2024)).

Density-based methods, like DBSCAN, identify clusters by detecting dense regions in data and treating sparse points as noise. Model-based methods, such as Gaussian Mixture Models (GMM), assume that data are generated from a mixture of probability distributions and seek to estimate the parameters that best fit the data. Finally, grid-based methods partition the data space into a grid of cells and perform clustering on these cells, providing an efficient solution for large datasets.

Some clustering evaluation metrics are as follows: (a) The Silhouette Score quantifies how well a data point is aligned with its own cluster relative to other clusters, with higher scores indicating better cluster cohesion and separation. (b) Intra-cluster and inter-cluster distances further complement this assessment by measuring the compactness of clusters and their separation from one another. (d) The Davies–Bouldin Index analyzes the relationship between within-cluster dispersion and inter-cluster separation, where lower values signify more effective clustering. (e) The Rand Index and its variant, the Adjusted Rand Index (ARI), are used to compare the agreement between predicted clusters and actual labels, if available, by evaluating the accuracy of pairwise point assignments.

In EDM, clustering has been applied to identify student groups with similar learning behaviors or performance levels, enabling personalized learning interventions. Some of the algorithms widely used to perform clustering are K-means, C-means, and hierarchical clustering methods [13,31].

Clustering allows teachers to classify students into performance-based groups, such as high-achieving or at-risk students, based on engagement or other academic data [28]. In this way, grouping helps to personalize learning and adopt measures specific to each group of students. Despite its utility, clustering remains underutilized in EDM, appearing in only 4.2% of analyzed studies or less [13,14]. An interesting example is the use of the X-means algorithm to group Greek students based on their academic performance, which

resulted in four levels of performance (Figure 4) [28]. Given the potential of clustering to enhance personalized learning, future research could focus on integrating it with other methodologies to better understand and improve student performance outcomes. Due to its unsupervised nature, clustering is typically not evaluated using cross-validation methods. Instead, its success is assessed by how well it separates distinct groups and reveals meaningful relationships between variables.

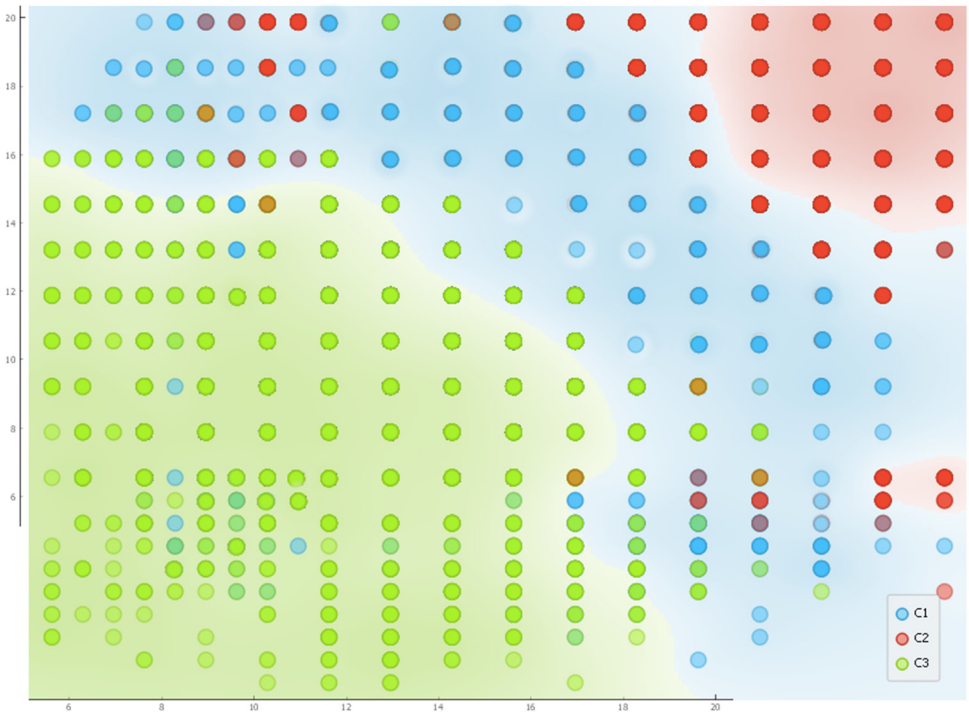

Figure 4. K-means clustering example, using Orange data mining software, Version: 3.35.0 [21]. Three clusters between Mathematics and Physics are presented in the second class of Greek junior high schools.

- Association rules mining (ARM) discovery reveals interesting relationships between variables in large data sets [32]. It is particularly effective in analyzing large amounts of student-related data, such as academic performance and behavioral data, to uncover hidden relationships. For example, ARM can be used to identify which combinations of courses or study habits are often associated with academic success or failure [14]. This information could be used to guide the adaptation of curriculum and teaching methods.

Algorithms like Apriori and FP-growth are commonly employed to generate association rules. Association rules highlight relationships between variables, such as how particular grades in prerequisite courses predict performance in subsequent ones. Association rule mining has been used to improve decision-making, allowing educational managers to design interventions tailored to the needs of specific student groups based on the rules identified. The strength of these rules is often evaluated using evaluation metrics like support, confidence, and lift, which provide insights into the reliability and relevance of the discovered patterns [32].

- Sequential pattern mining (SPM) is a method that focuses on identifying and analyzing sequences of events or behaviors over time. In the context of education, SPM helps in understanding how students' progress through learning activities, capturing patterns in their interactions with learning management systems or standard routes through a curriculum [26]. For example, SPM can reveal common sequences of mistakes or successful strategies in problem-solving, which can then be used to guide instructional design and provide personalized feedback [33].

SPM tracks the order of activities, such as when students watch video lectures, complete assignments, or participate in discussions, to predict academic outcomes. This analysis is valuable in courses involving complex tasks like programming, where understanding how students approach a problem, including the sequence of errors and corrections, can lead to better learning interventions. Sequential pattern mining (SPM) was used in a small fraction of studies, accounting for approximately 2.1% of the papers reviewed [13,14].

- Natural Language Processing (NLP) is a subfield of computational linguistics that aims to enable machines to understand, interpret, and generate human language. It bridges the gap between humans' natural language and computers' binary code.

In EDM, NLP is particularly used to analyze textual data generated by students in various educational contexts. NLP techniques allow for the processing of unstructured text data, transforming them into information that can be used for predicting student performance, detecting at-risk students, and enhancing feedback mechanisms. NLP has been applied to evaluate students' free-form comments and written assignments, using tools like Word2Vec and artificial neural networks to identify patterns in language usage and correlate them with learning outcomes [22]. K-Nearest Neighbors (KNN), GGANN, GGNN, and TBCNN algorithms are used for this scope. Furthermore, sentiment analysis can be employed to gauge students' emotional responses to course materials, providing educators with valuable information to tailor their teaching strategies and improve overall student engagement. According to Choi, only 4.2% of studies between 2018 and 2023 used this method [14].

NLP is also used to process and analyze discussion forum interactions, social media posts, and feedback surveys to gain deeper insights into student engagement and emotions (Figure 5). This kind of analysis allows educators to assess the cognitive and emotional states of students, thereby providing timely interventions when necessary [33]. By identifying linguistic patterns that indicate confusion, frustration, or disengagement, NLP can offer real-time feedback to both students and educators, helping to personalize learning experiences and improve student outcomes [3,22].

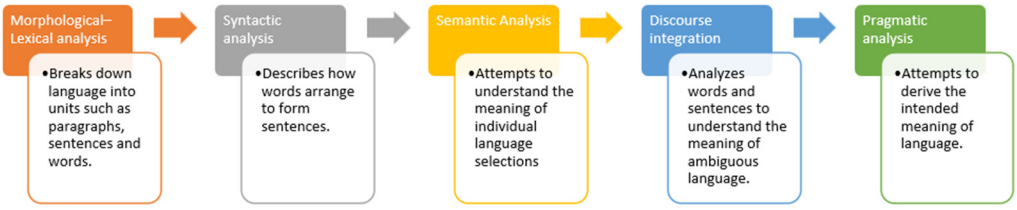

Figure 5. Natural Language Processing steps.

- Social Network Analysis (SNA) examines the relationships and interactions between entities, such as students and educators, in educational settings. By mapping the flow of communication and collaboration, SNA provides insights into how students form learning communities, share knowledge, and develop a better school climate that can influence academic performance [34]. This approach is particularly useful for analyzing online discussion forums, collaborative learning environments, and other learning networks. The relationships between individuals are represented as graphs, where nodes represent individuals and edges represent the interactions between them [34].

SNA has been applied to predict student performance by examining the strength and structure of social interactions (Figure 6). It is also found that centrality predicts future academic performance over and above a prior GPA. Students who are more centrally located within a learning network tend to perform better academically [34]. Conversely, students who are more isolated or have fewer social interactions may be at higher risk of dropping out or performing poorly. SNA allows educators to identify at-risk students early and intervene by encouraging more collaborative behaviors, thereby improving learning outcomes [6,13].

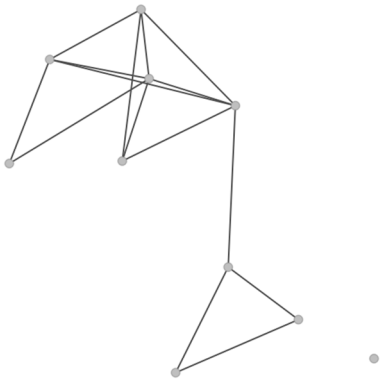

Figure 6. An example network with 10 nodes and 15 edges, created by Orange data mining software, Version: 3.35.0 using Network Generator node.

- Data visualization can help both students and teachers. Data visualization tools can effectively present information to students in an intuitive way, helping them understand their learning progress in real time. These tools also support classroom instruction, teaching interventions, and evaluations, while enabling educators to adjust their teaching goals, methods, and management strategies to enhance decision-making [35].

Data visualization techniques help to visualize data sets and to interpret them graphically, indicating complex data patterns in a simple way. Efficient data visualizations help expose existing data trends, outliers, and patterns since they reveal changes occurring over time [36]. Tools commonly used for presenting these findings include heatmaps that change depending on density, dashboards, and interactive graphs, which are commonly used to present findings in an accessible manner [22]. These tools not only enhance understanding but also facilitate informed decision-making by allowing educators and learners to engage with the data dynamically.

Data visualizations significantly enhance instructional strategies by allowing educators to quickly identify trends and patterns, which aids in making informed decisions about curriculum adjustments [36]. They enable targeted interventions by visualizing data, helping teachers tailor support to meet individual student needs and fostering a more personalized learning environment. Additionally, visualization tools facilitate the creation of reports that track student progress, promoting a culture of continuous improvement. Future data visualization tools may provide real-time insights into student understanding, enabling educators to adapt their teaching methods on the fly, ultimately leading to better educational outcomes.

6. EDM Topics

The EDM is used to support many different educational needs at the teaching and administrative levels. In order to provide an accurate representation of the thematic domains of the EDM discipline, the literature reviews [13,14,22,23] were used to record the progress of the research topics over the years.

- A dominant theme of EDM is pattern learning, which includes studies that analyze how students learn and behave in educational contexts. The main goal is to understand the learning process by examining patterns in the way students interact with learning systems and educational content. Much of the research has focused on discovering sub-groups of students with similar learning styles and evaluating the effectiveness of teaching methods [4,30]. The results of the research can help educational authorities in designing effective curricula and interventions tailored to the needs of students. According to Ozyurt et al., 27.22% of the studies between 2008 and 2022 have pattern learning as an objective [22].

- Recommendation systems in EDM can enhance personalized learning by providing students and educators with tailored suggestions that improve educational outcomes. These systems use algorithms to recommend courses, learning materials, and resources that align with individual student needs and preferences. By analyzing historical data, such as academic performance and behavior patterns, recommendation systems can predict future learning requirements and suggest the most relevant educational content. For instance, they help students select the most suitable courses or assignments, thereby optimizing their learning path and improving academic performance. Moreover, these systems can assist educators in developing customized learning plans for students, improving the overall quality of education by fostering more personalized and efficient learning experiences [37].

Educational recommendation systems (ERS) not only focus on students but also support other actors in the educational environment, such as teachers and administrators. These systems can suggest research materials, resources, and professional development opportunities that align with the interests and career goals of educators. Additionally, recommendation systems can be applied to course management, helping instructors tailor curricula to the needs of their students and identify the most effective teaching methods. By automating these processes, ERS contributes significantly to the improvement of educational quality indicators [10,37].

- Sentiment analysis and feedback help educational institutions to evaluate the quality of courses and programs from the student's perspective. Within this theme, studies have focused on developing approaches to analyze textual data from sources, such as open-ended survey responses, online discussions, and assignments using Natural Language Processing and machine learning [38]. Sentiment analysis provides valuable insights for educators on how to enhance student satisfaction, engagement, and motivation.
- Analysis of data from MOOCs and online learning platforms involves the use of several machine learning methods applied to online and blended learning models. The field continues to be important as virtual and hybrid education become dominant, especially in the COVID-19 era [39]. Researchers in this area are trying to optimize the design of online courses, promote participation, and enhance the digital learning experience.
- Learning analytics uses techniques to process massive educational and classroom-level data to gain insights into learning behaviors, understand learner profiles, and predict educational-related outcomes. This area has become increasingly important in recent years as technology-enhanced learning and large data sets are generated [40]. Learning analytics help to develop ideas for personalizing learning experiences, monitoring progress, and improving program design.
- Performance prediction, in the form of final grades, exam grades, and assessment of the risk of failure also informs targeted interventions. Advances in machine learning have enabled more accurate predictions. By analyzing characteristics such as demographics, prior academic history, measures of effort, and engagement, performance prediction develops performance prediction models and can thus identify at-risk students [28,41]. This allows for targeted academic support and optimization of teaching methodologies.
- Finally, student clustering applies unsupervised machine learning techniques to discover groups of students with similar characteristics without predefined labels. Clustering supports the development of personalized, differentiated teaching by recognizing student diversity and grouping students accordingly.

In order to examine the longitudinal trends in each topic, the data from a literature review [30] were used and the percentage of each topic over time was calculated; Figure 7 provides a visual representation of the trends.

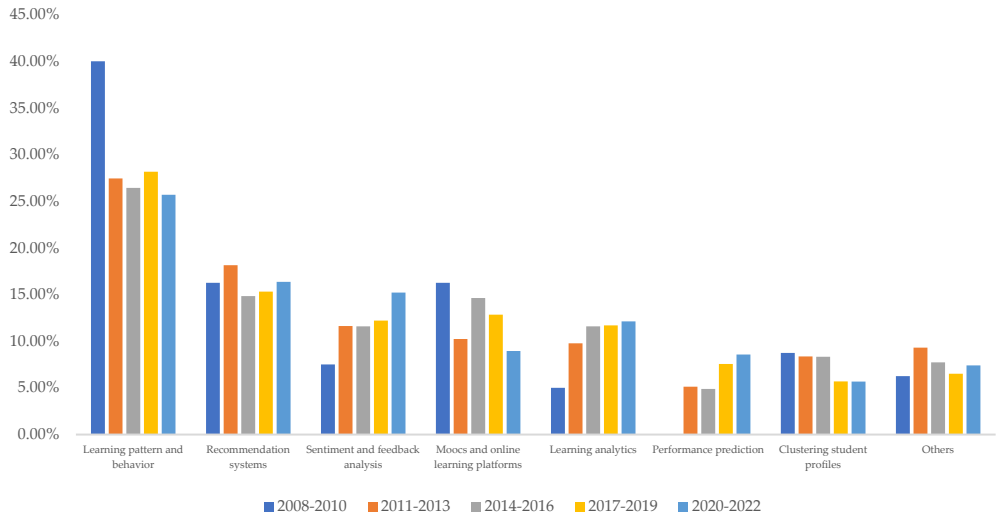

Figure 7. EDM topics over time.

Figure 7 illustrates the shifting focus of research topics in EMD over the period from 2008 to 2022. Initially, learning patterns and behavior dominated the field, peaking at 40.00% in 2008–2010. However, its importance has gradually decreased, stabilizing at 25.70% in 2020–2022, although it remains the dominant topic overall. Recommendation systems have maintained a constant level of interest, fluctuating slightly between 16.25% and 16.36% over the years. Sentiment and feedback analysis, on the other hand, has seen a significant increase, rising steadily from 7.50% in 2008–2010 to 15.21% in 2020–2022, reflecting the growing importance of understanding learner sentiment. MOOCs and online learning platforms showed a significant drop in interest from 16.25% in 2008–2010 to 8.95% in 2020–2022. Learning analytics showed a significant upward trend, increasing from 5.00% to 12.13% over the same period. Another emerging area is performance prediction, which was absent in the first period but has grown steadily to reach 8.57% in 2020–2022. On the other hand, the clustering of student profiles has seen a gradual decline from 8.75% in 2008–2010 to 5.68% in 2020–2022. These trends suggest an increasing focus on areas such as sentiment analysis, learning analytics, and performance prediction, which have gained significant attention in recent years.

7. EDM Basic Applications

We should emphasize that EDM encompasses a wide and diverse range of applications, adaptable to various educational environments and contexts. The proper analysis and examination of data regarding student performance and learning preferences by teachers leads to increased engagement and better academic results for the students, which leads to a conducive learning environment [13,30]. These findings help to design adaptive learning systems that are able to be modified dynamically and create a unique teaching/learning experience both within the traditional classroom setting as well as through online e-learning platforms. By creating and maintaining detailed student profiles based on their interactions with such systems, educators can make predictions about future performance trends, adjust their instructional methods accordingly, and assist with personalized learning that meets individual student needs [2].

On the other hand, early warning systems show a remarkable ability to track students who are at risk of falling behind by analyzing datasets of students' academic performance and engagement in the learning process over time. Such advanced tools are key to informing teachers and students in a timely manner of the need for additional support and resources, thus facilitating interventions that prevent students from dropping out and providing equal opportunities for students [10].

EDM plays a significant and transformative role in the complex and multifaceted process of curriculum design by uncovering and identifying which specific courses or

teaching methods are most beneficial for student learning outcomes. By analyzing student feedback combined with performance data, educational authorities are empowered to improve their curricula in ways that better meet the diverse needs of students and their individual learning goals, thereby fostering an enhanced educational experience [3,10].

EDM can also improve pedagogical evaluation methods, as it provides useful information about students' learning experiences and clarifies any ambiguities they may have about what happened in that particular lesson. In addition, personalizing feedback through automated assessment systems can help students identify and understand their mistakes, thus improving learning processes and academic performance, as well as reducing teacher workload [4].

EDM can be used as a management tool that helps in the allocation of resources in the best manner. Although the question of how much to allocate requires wider agreements, through analysis of student enrollments, course approval, and resource use, administrators are able to make informed decisions about staffing, scheduling, and investments in educational technologies [6]. Visualization technology provides real-time visualizations of student data, allowing teachers and administrator staff to monitor and quickly identify trends. These dashboards support data-driven decision-making and promote a continuous improvement mindset [1].

The above list of EDM applications is indicative. Moreover, the ever-increasing number of publications in the scientific field constantly reveals new approaches and applications of this discipline.

8. Challenges

EDM offers significant potential for improving educational outcomes but faces several important challenges. These challenges arise from the nature of the methods used, the complexity of educational contexts, the nature of educational data, and ethical considerations. This diversity of potential risks requires careful scheduling and preprocessing, as inaccurate analyses drive misleading conclusions [4].

According to Pardo and Siemens [42], educational data often contain sensitive information about students, teachers, and other people that needs strict privacy protections under laws like FERPA and GDPR. Other than legal compliance, ethical concerns include transparency around how data can be used as well as ensuring that its analysis supports students without causing them any harm, like reinforcing their biases or interfering with their autonomy [1]. Thus, it is imperative that strong ethical frameworks and guidelines are established for responsible handling of such matters.

It is important for the results to have explanatory power. Although neural networks and other complicated models may produce the right predictions, teachers are often unable to understand and trust them [7]. This type of "black box" approach hindered use in educational contexts where decision-making requirements call for clarity. The development of simpler models that can be explained easily remains a challenge.

Scalability is another challenge as educational data continues to grow exponentially [24]. It is vital that these models are able to scale with an increasing amount of information without sacrificing any efficiency or correctness. In this case, it may require optimizing algorithms, processing massive datasets using cloud computing, and ongoing examination of algorithm performance after new data sets are introduced. There is a rising demand for real-time data analysis, particularly in settings where adaptive learning takes place. Having efficient tools for processing data through advanced algorithms is one aspect of real-time analysis. It involves addressing issues that have to do with speed and processing infrastructure and making sure that the real-time information is accurate as well as useful [1].

Educational data are influenced by many cultural, social, and institutional factors. These factors can significantly affect outcomes and must be taken into account to ensure the reliability of results [10]. Developing models that take these contextual variables into account increases complexity by requiring collaboration with educators and experts to

incorporate contextual variables into data analysis processes, but it is necessary to provide information applicable to different educational settings. EDM involves collaboration across many disciplines, including computer science, education, psychology, and statistics. Bridging these fields is difficult but essential [4]. The interdisciplinary approach allows for the development of more integrated models that can address complex educational challenges. However, it also requires effective communication and collaboration between stakeholders. Continuous study is needed to resolve these issues. Researchers must develop new machine learning algorithms and data processing systems that are both innovative and ethical. Furthermore, communication between teachers, researchers, and decision-makers promotes the evolution of best practices. Addressing moral concerns and aligning EDM with educational objectives will ensure that EDM is used responsibly.

9. Future Trends

EDM had a remarkable development linked to the rapid developments of algorithms, combined with the increase in the amount of available data. As the field continues to grow, several trends are shaping.

The integration of artificial intelligence (AI) with EDM is expected to be pronounced [43]. The use of AI and the development of recommendation systems can bring significant improvements to the quality of education that students receive. On the other hand, the integration with AI increases the challenges associated with the accumulation, control, and handling of the data, while the critical issue of personal data security is raised. Breakthrough AI technologies have the ability to use large amounts of data and present many new and exciting learning opportunities. From this point of view, EDM functions as a driver by providing insights into how the students interact and engage in virtual environments and thus contributes to the improvement of the educational content [44].

The integration of data mining and deep learning techniques into augmented reality (AR) and virtual reality (VR) applications has significantly enhanced the educational experience by offering immersive, interactive learning environments. These technologies facilitate a deeper level of engagement and personalization, enabling educators to tailor educational content more effectively to individual student needs. In particular, deep learning, through Natural Language Processing (NLP) and sentiment analysis, has been employed to analyze large datasets from educational platforms and social media [45]. This analysis provides critical insights into student attitudes, behaviors, and learning preferences, allowing for the optimization of AR and VR educational content. For instance, the analysis of social media data has proven valuable in understanding public sentiment and perspectives on the use of AR and VR in education, thereby informing the development of more effective learning tools [46].

In AR, deep learning models enhance adaptability by recognizing the real-world reactions of students and providing real-time feedback, creating dynamic, visually engaging educational experiences. These applications allow students to interact with complex concepts in a tangible manner, improving comprehension and retention. Similarly, VR benefits from deep learning by predicting student behaviors and learning preferences, enabling the creation of personalized, immersive virtual environments. These environments can adjust to individual learning paces and styles, leading to more effective educational outcomes. Through the use of data-driven insights, AR and VR technologies not only improve the learning process but also foster a richer, more interactive platform for students, contributing to higher engagement and better academic performance [46,47].

However, the already established fields of use of EDM also attract the interest of researchers. One of the most significant trends in EDM is the advancement of predictive analytics, particularly in the areas of student performance and retention. Predictive regression and classification models are becoming more sophisticated. The use of massive amounts of data is helping to better predict student performance and identify students who may be struggling or dropping out. In particular, the emphasis on predicting dropout risk has grown significantly, as highlighted by Refs. [46,47]. These models use different

types of data to predict future academic performance. Educational institutions are likely to follow this trend as a means of improving the academic success of their students through data-driven interventions, which could also help them to improve their retention rates [23]. Students' performance continues to be the primary subject of research [48–59].

Educational environments generate an increasing amount of unstructured data. These data can be in the form of text, video, social media interactions, and more. This necessitates the integration of unstructured data analysis into EDM [22]. While traditional educational data mining methods have primarily focused on structured data, the rise of big data has necessitated the implementation of techniques that can handle unstructured data. The processing and analysis of unstructured data provide an opportunity to gain further insight into student behavior and learning processes. Another reason for the increasing use of unstructured data is the use of NLP [45]. The data used are texts, comments, or posts in forums and educational platforms. A common objective of all these methods is to improve student performance [22].

Learning analysis continues to be a main focus of interest. Data on the interaction between educational content and everyday classroom teaching are used to develop frameworks and models of effective teaching methods [7,31]. Sentiment analysis has also become increasingly popular in recent years. Teachers and administrative staff can better understand what students want and where they need to improve by analyzing feedback from course evaluations, discussion forums, and social media. These details can then be used to make data-driven decisions to enhance the learning experience [7,49]. LA continues to attract the interest of researchers [60–65].

Personalized recommendation systems are also becoming more and more sophisticated and provide students with accurate and tailored learning recommendations. The integration of NLP techniques into text data provides information that allows students to identify problems in their learning path [47]. This makes recommender systems an effective tool for promoting self-directed learning. The evolution of recommendation systems is expected to continue as institutions prioritize the quality of education they provide [23,50]. The importance of recommendation systems is also linked to the growing trend of lifelong learning. As people seek to learn new skills throughout their lives through e-learning systems, recommendation systems can help them find the most relevant and effective learning pathways. Overall, recommendation systems are the primary research topic in 2024 [66–71].

EDM is a powerful tool to support decision-making in educational environments. EDM can optimize curriculum design, personalize learning experiences, and improve resource allocation. By providing data-driven information, it enhances institutions' ability to make informed, strategic decisions that lead to better educational outcomes. Decision support has been a trend in recent years, particularly in higher and online education [72–79].

As EDM continues to evolve, educational institutions can leverage these tools to improve student engagement, retention, and overall learning success. The continuous advancements in EDM promise to reshape the landscape of modern education, fostering a more data-driven and personalized learning environment. These trends indicate a dynamic evolution in the EDM landscape, focusing on practical applications and interdisciplinary collaboration.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/encyclopedia4040108/s1.

Author Contributions: Conceptualization, M.W. and G.K.; methodology, I.P.; formal analysis, I.P.; investigation, I.P.; writing—original draft preparation, I.P; writing—review and editing, M.W. and G.K.; supervision, M.W.; project administration, M.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data are contained within the article and Supplementary Materials.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Romero, C.; Romero, J.R.; Ventura, S. A Survey on Pre-Processing Educational Data. In *Educational Data Mining*; Springer: Cham, Switzerland, 2014; pp. 29–64.
2. Romero, C.; Ventura, S. Educational Data Science in Massive Open Online Courses. *WIREs Data Min. Knowl. Discov.* **2017**, *7*, e1187. [CrossRef]
3. Bakhshinategh, B.; Zaiane, O.R.; ElAtia, S.; Ipperciel, D. Educational Data Mining Applications and Tasks: A Survey of the Last 10 Years. *Edu. Inf. Technol.* **2018**, *23*, 537–553. [CrossRef]
4. Baker, R.S.J.d; Inventado, P.S. Educational Data Mining and Learning Analytics. In *Learning Analytics: From Research to Practice*; Larusson, J.A., White, B., Eds.; Springer: Berlin/Heidelberg, Germany, 2014; pp. 61–75.
5. Romero, C.; Ventura, S.; Pechenizky, M.; Baker, R. *Handbook of Educational Data Mining*; CRC Press, Taylor & Francis Group: Boca Raton, FL, USA, 2010.
6. Siemens, G.; Baker, R.S.J.d. Learning Analytics and Educational Data Mining: Towards Communication and Collaboration. In Proceedings of the 2nd International Conference on Learning Analytics and Knowledge, Vancouver, BC, Canada, 29 April–2 May 2012; pp. 1–3.
7. Siemens, G. Learning Analytics: The Emergence of a Discipline. *Am. Behav. Sci.* **2013**, *57*, 1380–1400. [CrossRef]
8. Cerezo, R.; Lara, J.-A.; Azevedo, R.; Romero, C. Reviewing the differences between learning analytics and educational data mining: Towards educational data science. *Comput. Hum. Behav.* **2024**, *154*, 108155. [CrossRef]
9. Chan, K.I.; Lei, P.I.S.; Pang, P.C.-I. A literature review on educational data mining with secondary school data. In Proceedings of the 9th International Conference on Education and Training Technologies, Macau, China, 21–23 April 2023; ACM: New York, NY, USA, 2023; pp. 1–6. [CrossRef]
10. Baker, R.S. *Big Data and Education*, 2nd ed.; Teachers College, Columbia University: New York, NY, USA, 2015.
11. Ray, S.; Saeed, M. Applications of Educational Data Mining and Learning Analytics Tools in Handling Big Data in Higher Education. In *Applications of Big Data Analytics*; Springer: Cham, Switzerland, 2018; pp. 135–160. [CrossRef]
12. Bousbia, N.; Belamri, I. Which Contribution Does EDM Provide to Computer-Based Learning Environments? In *Studies in Computational Intelligence. Educational Data Mining*; Springer: Cham, Switzerland, 2014; pp. 3–28. [CrossRef]
13. Papadogiannis, I.; Poulopoulos, V.; Wallace, M. A Critical Review of Data Mining for Education: What Has Been Done, What Has Been Learnt and What Remains to Be Seen. *Int. J. Educ. Res. Rev.* **2020**, *5*, 353–372. [CrossRef]
14. Choi, W.-C.; Lam, C.-T.; Mendes, A.J. A systematic literature review on performance prediction in learning programming using educational data mining. In Proceedings of the 2023 IEEE Frontiers in Education Conference (FIE), College Station, TX, USA, 18–21 October 2023; pp. 1–9. [CrossRef]
15. Romero, C.; Ventura, S. Educational data mining: A survey from 1995 to 2005. *Expert Syst. Appl.* **2007**, *33*, 135–146. [CrossRef]
16. Baker, R.S.; Yacef, K. The state of educational data mining in 2009: A review and future visions. *J. Educ. Data Min.* **2009**, *1*, 3–17.
17. Papamitsiou, Z.; Economides, A.A. Learning analytics and educational data mining in practice: A systematic literature review of empirical evidence. *J. Educ. Technol. Soc.* **2014**, *17*, 49–64.
18. Peña-Ayala, A. Educational data mining: A survey and a data mining-based analysis of recent works. *Expert Syst. Appl.* **2014**, *41*, 1432–1462. [CrossRef]
19. Thakar, P.; Mehta, A.; Manisha. Performance Analysis and prediction in educational Data mining: A research travelogue. *Int. J. Comput. Appl.* **2015**, *110*, 60–68.
20. Sukhija, S.; Singh, S.; Riar, C.S. Isolation of starches from different tubers and study of their physicochemical, thermal, rheological and morphological characteristics. *Starch-Stärke* **2016**, *68*, 160–168. [CrossRef]
21. Del Río, C.A.; Insuasti, J.A.P. Predicting academic performance in traditional environments at higher-education institutions using data mining: A review. *Ecos Acad.* **2016**, *4*, 185–201.
22. Ozyurt, O.; Ozyurt, H.; Mishra, D. Uncovering the educational data mining landscape and future perspective: A comprehensive analysis. *IEEE Access* **2023**, *11*, 120192–120208. [CrossRef]
23. Romero, C.; Ventura, S. Educational data mining and learning analytics: An updated survey. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **2020**, *10*, e1355. [CrossRef]
24. Rodrigues, M.W.; Isotani, S.; Zárate, L.E. Educational data mining: A review of evaluation process in e-learning. *Telemat. Inform.* **2018**, *35*, 1701–1717. [CrossRef]
25. Pham Kim, C. Evaluating Student Teachers in Micro-Teaching with Analysis of Video Recording Lesson by Boris Software at Vietnam National University. *Sci. Publ. Cent. Sociosphere* **2017**, *8*, 67–74. [CrossRef]
26. Ferreira-Mello, R.; André, M.; Pinheiro, A.; Costa, E.; Romero, C. Text mining in education. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **2019**, *9*, e1332. [CrossRef]
27. Chaturapruek, S.; Dalberg, T.; Thompson, M.E.; Giebel, S.; Harrison, M.H.; Johari, R.; Stevens, M.L.; Kizilcec, R.F. Studying undergraduate course consideration at scale. *AERA Open* **2021**, *7*, 233285842199114. [CrossRef]
28. Papadogiannis, I.; Wallace, M.; Poulopoulos, V.; Karountzou, G.; Ekonomopoulos, D. A First Ever Look into Greece's Vast Educational Data: Interesting Findings and Policy Implications. *Educ. Sci.* **2021**, *11*, 489. [CrossRef]

29. Chen, Y.; Chang, H.-H. Psychometrics Help Learning: From Assessment to Learning. *Appl. Psychol. Meas.* **2018**, *42*, 3–4. [CrossRef] [PubMed]
30. Zhang, Y.; Yun, Y.; An, R.; Cui, J.; Dai, H.; Shang, X. Educational data mining techniques for student performance prediction: Method review and comparison analysis. *Front. Psychol.* **2021**, *12*, 698490. [CrossRef]
31. Romero, C.; Ventura, S. Data Mining in Education. *WIREs Data Min. Knowl. Discov.* **2013**, *3*, 12–27. [CrossRef]
32. Njiru, T. Association rule mining in educational data: Unveiling patterns for enhanced learning outcomes. *Preprints* **2024**. [CrossRef]
33. Xu, R.; Chen, J.; Han, J.; Tan, L.; Xu, L. Towards emotion-sensitive learning cognitive state analysis of big data in education: Deep learning-based facial expression analysis using ordinal information. *Computing* **2020**, *102*, 765–780. [CrossRef]
34. Polatcan, M.; Balcı, A. Social capital wealth as a predictor of innovative climate in schools. *Int. J. Contemp. Educ. Res.* **2022**, *6*, 183–194. [CrossRef]
35. Lu, M. Research on data visualization analysis in education curriculum quality management and student development. In Proceedings of the Annual Conference on Computers, Ottawa, ON, Canada, 16–18 October 2020. [CrossRef]
36. Hansen, L.; Holanda, M.; Borges, V.R.P.; Da Silva, D. Visual analysis of educational data: A case study of introductory programming courses at the University of Brasília. In Proceedings of the 2022 IEEE Frontiers in Education Conference (FIE), Uppsala, Sweden, 8–11 October 2022; IEEE: Piscataway, NJ, USA, 2022; pp. 1–6. [CrossRef]
37. Aucancela, M.; Briones, A.G.; Chamoso, P. Educational recommender systems: A systematic literature review. In Proceedings of the Barcelona Conference on Education 2023: Official Conference Proceedings, Barcelona, Spain, 19–23 September 2023; pp. 933–951. [CrossRef]
38. Kandhro, I.A.; Chhajro, M.A.; Kumar, K.; Lashari, H.N.; Khan, U. Student feedback sentiment analysis model using various machine learning schemes: A review. *Indian J. Sci. Technol.* **2019**, *12*, 1–9. [CrossRef]
39. Asad, R.; Altaf, S.; Ahmad, S.; Mahmoud, H.A.; Huda, S.; Iqbal, S. Machine learning-based hybrid ensemble model achieving precision education for online education amid the lockdown period of COVID-19 pandemic in Pakistan. *Sustainability* **2023**, *15*, 5431. [CrossRef]
40. Hernandez-de-Menendez, M.; Morales-Menendez, R.; Escobar, C.A.; Ramírez Mendoza, R.A. Learning analytics: State of the art. *Int. J. Interact. Des. Manuf.* **2022**, *16*, 1209–1230. [CrossRef]
41. Gadde, S.S.; Anand, D.; Sasidhar Babu, N.; Pujitha, B.V.; Sai Reethi, M.; Pradeep Ghantasala, G.S. Performance prediction of students using machine learning algorithms. In *Lecture Notes in Mechanical Engineering*; Applications of Computational Methods in Manufacturing and Product Design; Springer: Singapore, 2022; pp. 405–411. [CrossRef]
42. Pardo, A.; Siemens, G. Ethical and Privacy Principles for Learning Analytics. *Br. J. Educ. Technol.* **2014**, *45*, 438–450. [CrossRef]
43. Ankora, C.; Aju, D. Integrating Educational Data Mining in Augmented Reality Virtual Learning Environment. In *Advances in Computing Communications and Informatics*; Bentham Science Publishers: Sharjah, United Arab Emirates, 2022; pp. 1–18. [CrossRef]
44. Liu, N.; Chen, Y.; Yang, X.; Hu, Y. Do Demographic Characteristics Make Differences? Demographic Characteristics as Moderators in the Associations between Only Child Status and Cognitive/Non-cognitive Outcomes in China. *Front. Psychol.* **2017**, *8*, 423. [CrossRef]
45. Shaukat, S.M. Exploring the potential of augmented reality (AR) and virtual reality (VR) in education. *Int. J. Adv. Res. Sci. Commun. Technol.* **2023**, *3*, 52–57. [CrossRef]
46. Lampropoulos, G.; Keramopoulos, E.; Diamantaras, K.; Evangelidis, G. Augmented reality and virtual reality in education: Public perspectives, sentiments, attitudes, and discourses. *Educ. Sci.* **2022**, *12*, 798. [CrossRef]
47. Khan, A.; Ghosh, S.K. Student performance analysis and prediction in classroom learning: A review of educational data mining studies. *Educ. Inf. Technol.* **2021**, *26*, 205–240. [CrossRef]
48. Dol, S.M.; Jawandhiya, P.M. Review of EDM for Analyzing the Performance of Students in Educational Settings. In Proceedings of the 2022 6th International Conference on Computing, Communication, Control and Automation (ICCUBEA), Pune, India, 26–27 August 2022; IEEE: Piscataway, NJ, USA, 2022. [CrossRef]
49. Mavridis, A.; Symeonidis, A.L. A review of sentiment analysis applied to education. *J. Educ. Technol. Soc.* **2021**, *24*, 48–58.
50. Raza, S.; Rahman, M.; Kamawal, S.; Toroghi, A.; Raval, A.; Navah, F.; Kazemeini, A. A Comprehensive Review of Recommender Systems: Transitioning from Theory to Practice. *arXiv* **2024**, arXiv:2407.13699. Available online: https://arxiv.org/abs/2407.13699 (accessed on 29 September 2024).
51. Wei, H.; Cong, W.; Wu, A.; Zhou, G. Prediction method of higher education college students' employability based on data mining. In *Learning and Analytics in Intelligent Systems*; Springer: Cham, Switzerland, 2024; pp. 144–154. [CrossRef]
52. Pliuskuvienė, B.; Radvilaitė, U.; Juodagalvytė, R.; Ramanauskaitė, S.; Stefanovič, P. Educational data mining and learning analytics: Text generators usage effect on students' grades. *New Trends Comput. Sci.* **2024**, *2*, 19–30. [CrossRef]
53. Hanumanthappa, S.; Prakash, C. Machine learning based education data mining through student session streams. *Int. J. Reconfigur. Embedded Syst.* **2024**, *13*, 383. [CrossRef]
54. Tosun, S.; Bakan Kalaycıoğlu, D. Data mining approach for prediction of academic success in open and distance education. *J. Educ. Technol. Online Learn.* **2024**, *7*, 168–176. [CrossRef]
55. Bussaman, S.; Nasa-Ngium, P.; Sararat, T.; Nuankaew, W.S.; Nuankaew, P. Influence analytics model of the general education courses toward the academic achievement of Rajabhat university students using data mining techniques. In *Smart Innovation, Systems and Technologies*; Springer: Cham, Switzerland, 2024; pp. 117–129. [CrossRef]

56. Shen, S. Exploration of the management mode and quality evaluation of entrepreneurship education in colleges and universities based on data mining. *Trans. Comp. Educ.* **2024**, *6*, 1. [CrossRef]
57. Chen, Z. Intelligent evaluation system for labor education quality based on data mining. In Proceedings of the 2024 IEEE 7th Eurasian Conference on Educational Innovation (ECEI), Bangkok, Thailand, 26–28 January 2024. [CrossRef]
58. Papadogiannis, I.; Wallace, M.; Poulopoulos, V.; Vassilakis, C.; Lepouras, G.; Platis, N. An Assessment of the Effectiveness of the Remedial Teaching Education Policy. *Knowledge* **2023**, *3*, 349–363. [CrossRef]
59. Papadogiannis, I.; Wallace, M.; Poulopoulos, V. Examining Pupils' Achievement in Primary and Secondary Schools in Greece. *Eur. J. Eng. Technol. Res.* **2022**, *2022*, 10–18. [CrossRef]
60. Roski, M.; Ewerth, R.; Hoppe, A.; Nehring, A. Exploring data mining in chemistry education: Building a web-based learning platform for learning analytics. *J. Chem. Educ.* **2024**, *101*, 930–940. [CrossRef]
61. Gagnon, D.J.; Swanson, L.; Harpstead, E. Open game data: Defining a pipeline and standards for educational data mining and learning analytics with video game data. In Proceedings of the 2024 IEEE Conference on Games (CoG), Milan, Italy, 5–8 August 2024; pp. 1–8. [CrossRef]
62. Pan, J. Research on the online learning mechanism of education based on data mining. In Proceedings of the 2024 International Conference on Informatics Education and Computer Technology Applications (IECA), Beijing, China, 26–28 January 2024; pp. 38–41. [CrossRef]
63. Hajjej, F.; Ayouni, S.; Alohali, M.A.; Maddeh, M. Novel framework for autism spectrum disorder identification and tailored education with effective data mining and ensemble learning techniques. *IEEE Access* **2024**, *12*, 35448–35461. [CrossRef]
64. Chen, J. Construction of E-learning English wisdom classroom based on educational big data mining. *Comput.-Aided Des. Appl.* **2024**, *21*, 251–264. [CrossRef]
65. Yu, S.; Zhang, Z.; Kang, K.; Zhu, L.; Jiang, X. Discussion on individualized teaching strategies of international Chinese education based on data mining. In *Learning and Analytics in Intelligent Systems*; Springer: Cham, Switzerland, 2024; pp. 574–583. [CrossRef]
66. Zhang, A. Research and practice of E-learning education and teaching mode based on data mining technology. *Comput.-Aided Des. Appl.* **2024**, *21*, 32–44. [CrossRef]
67. Rybalchenko, A.; Abildinova, G. Personalizing the learning process through data mining in higher education. *Sci. Herald Uzhhorod Univ. Phys. Ser.* **2024**, *56*, 1580–1588. [CrossRef]
68. Zhang, L. Data mining and learning behaviour analysis of French online education data-driven teaching based on generative adversarial network improvement Apriori algorithm. *Int. J. Wirel. Mobile Comput.* **2024**, *1*, 1. [CrossRef]
69. Ji, X.; Sun, L.; Xu, X.; Lei, X. Construction and innovative exploration of personalized learning systems in the context of educational data mining. *Int. J. Inform. Commun. Technol. Educ.* **2024**, *20*, 1–14. [CrossRef]
70. Sareminia, S.; Mohammadi Dehcheshmeh, V. Developing an intelligent and sustainable model to improve E-learning satisfaction based on the learner's personality type: Data mining approach in high education systems. *Int. J. Inform. Learn. Technol.* **2024**, *41*, 394–427. [CrossRef]
71. Chen, X.; Cao, C. Research on building community education platform based on data mining technology. In *Learning and Analytics in Intelligent Systems*; Springer: Cham, Switzerland, 2024; pp. 398–406. [CrossRef]
72. Wang, Y. University moral education management system using ensemble learning in data mining. In Proceedings of the 2024 International Conference on Data Science and Network Security (ICDSNS), Tiptur, India, 26–27 July 2024; pp. 1–4. [CrossRef]
73. Han, L. Prediction and analysis of students' behavior based on data mining in educational administration. In *Learning and Analytics in Intelligent Systems*; Springer: Cham, Switzerland, 2024; pp. 229–238. [CrossRef]
74. Liu, W.; Qin, X.; Yang, L. High quality management of higher education based on data mining. *Int. J. Bus. Intell. Data Min.* **2024**, *25*, 424–450. [CrossRef]
75. Li, S.; Ma, B.; Meng, D. Reflections on strategies for psychological health education for college students based on data mining. *Int. J. Bus. Intell. Data Min.* **2024**, *25*, 394–408. [CrossRef]
76. Kawesha, F.; Phiri, J. Data mining and machine learning-based predictive model to support decision-making for the accreditation of learning programmes at the higher education authority. In *Lecture Notes in Networks and Systems*; Springer: Cham, Switzerland, 2024; pp. 351–361. [CrossRef]
77. Pan, W. Study on quality evaluation method of multimedia distance education based on Data Mining. *Int. J. Contin. Eng. Educ. Lifelong Learn.* **2024**, *34*, 194–203. [CrossRef]
78. Wang, L.; Wang, B.; Huang, H.; Zhang, X. Research on the teaching reform of Data Mining and Data Analysis based on the concept of 'outcomes-Based Education'. *High. Educ. Pract.* **2024**, *1*, 1–6. [CrossRef]
79. Zhong, Q. Intelligent optimization of labor education curriculum based on data mining technology. In Proceedings of the 2024 IEEE 4th International Conference on Electronic Communications, Internet of Things and Big Data (ICEIB), Taipei, Taiwan, 19–21 April 2024. [CrossRef]