Available online at www.sciencedirect.com

ScienceDirect

IFAC PapersOnline CONFERENCE PAPER ARCHIVE

IFAC PapersOnLine 58-26 (2024) 118–123

Clustering of Learners' Data – A Case Study on Learning Analytics

Susanne Franke.* Martin Trommer. Sven Hellbach.** Tobias Teich.****

* *University of Applied Sciences Mittweida, Technikumplatz 17, 09648 Mittweida, Germany (e-mail: franke2@hs-mittweida.de).*

** *University of Applied Sciences Zwickau, Kornmarkt 1, 08066 Zwickau, Germany (e-mail: martin.trommer@fh-zwickau.de).*

Abstract: With the increasing amount of digital learning offers, there is a high demand for individualized, adaptive learning pathways. The paper explores the role of learning analytics to improve qualification processes in educational institutions. E-learning, as a crucial component in educational and organizational learning, is examined for its role in enhancing learner success and motivation. Focusing specifically on Artificial Intelligence, the study aims to investigate how analysis approaches can provide valuable insights into the conceptualization and implementation of individualized learning pathways. In particular, the experimental environment, the use case for data provision and necessary data preparation are described. Furthermore, the application of different clustering methods to learners' data gathered in the context of e-learning is presented and the findings are discussed.

Keywords: Clustering Methods, Learning Analytics, Individualized Learning.

1. INTRODUCTION

Learning and qualification processes are highly relevant not only for educational institutions but also for companies as they need to provide onboarding for new employees, re- and upskilling courses to efficiently deploy the available personnel. The shift from Smart Manufacturing to Industry 5.0 also underlines the significance of human development and collaboration with technological systems. Smart Manufacturing refers to the integration of advanced technologies, such as the Internet of Things, data analysis, and Artificial Intelligence (AI), into or administrative processes or manufacturing to enhance efficiency, productivity, and flexibility. Industry 5.0 is characterized by the seamless collaboration between humans and advanced technologies, particularly emphasizing human-centric design and enhancing human well-being and skill development, see Rannertshauer et al. (2022).

In both educational and organizational learning, e-learning has gained increasing significance. The demands comprise higher learner's motivation, increased learning success and efficiency (for both teachers and learners), see Kerres (2001).

Artificial Intelligence has become an important and rapidly evolving technology to support these processes. It is our goal to investigate how AI approaches can support the evaluation of data provided by the learners. Therefore, we created learning pathways and provided the learning content in an e-learning course. Subsequently, we applied clustering algorithms to group learners by identifying patterns within their learning behavior (i.e., data on the visited learning content such as the viewing timestamp of each page and submissions) and their achievements in tasks and assignments on the learning content. The findings allow a refinement of the learning pathways and a more precise provision of learning material for learners with similar behavior and learning results. Additionally, the teacher can use the results to improve the preparation of subsequent classroom lectures which follow the digital learning courses.

The paper is structured as follows. Section 2 provides theoretical background information. In Section 3, the experimental environment as the basis for data provision is described. The main results are formulated and discussed in Section 4. The paper concludes with a summary and an outlook on future topics in Section 5.

2. THEORETICAL BACKGROUND

Srinivasa et al. (2022) define Learning Analytics as "…the measurement, collection, analysis and reporting of data about learners and their contexts, for purposes of understanding and optimizing learning and the environments in which it occurs". Using theories and methods from, e.g., statistics, cognitive psychology, pedagogy, and computer science, Learning Analytics is an essential tool to provide individually tailored learning content in the form of learning pathways. According to Scott (1991), learning pathways are defined as a sequence of intermediate steps from learner's initial knowledge state to a desired knowledge state. They allow shaping the learner's learning journey efficiently and effectively, thereby ensuring that the learner develops a comprehensive understanding of a subject or a proficiency in a skill. Learning pathways can incorporate various resources, activities, assessments, and support mechanisms specifically tailored to the learner's needs and goals, see Yang et al. (2017).

AI, as a subfield of computer science, is concerned with the formulation of algorithms designed to perform tasks traditionally attributed to human intelligence, such as learning and problem-solving. A comprehensive study of the statistical foundations and an introduction to Machine Learning methodologies is given by Bishop (2006). Shalev-Shwartz et

10.1016/j.ifacol.2024.10.281

al. (2014) focus on the algorithmic perspective and aspects like complexity analysis, while Goodfellow et al. (2016) provide an introduction to Deep Learning. AI applications are widely used in various fields like, e.g., fraud detection, see Dhieb et al. (2020), or medical diagnosis, see Teich et al. (2023).

In the context of learning analytics, it is our goal to identify learners with similar learning behavior and learning progress based on their learning data. This corresponds to identifying patterns in a data set with no existing labels, which can be solved by clustering algorithms.

Clustering as an unsupervised Machine Learning approach tries to identify hidden structures in finite, unlabeled data and to group the data points into a certain number of clusters, see Cherkassky et al. (1998). In particular, the goal is to obtain clusters where the data points within one group should be as homogeneous as possible whereas the data points belonging to different groups should be maximally heterogeneous. Here, the mathematical definition of homogeneity and heterogeneity (which can for example be expressed by a dissimilarity function d: $X \times X \to \mathbb{R}_+$, e.g., the Euclidean distance, for a data space $X \subseteq \mathbb{R}^n$) may differ for different clustering algorithms. Moreover, there exist different categorizations of clustering algorithms. In the following, we want to focus on partitional, hierarchical and density-based clustering.

Partitional clustering separates data into a predefined number of clusters, where the sum of the pairwise dissimilarity between the data points of one cluster and a specific cluster representative (cluster centers or centroids) is minimized while the dissimilarities between different clusters are maximized.

Density-based clustering forms clusters where data points of one cluster are high density areas. The method assigns each data point to one of the following types: core points, border points, and noise, see, e.g., Kriegel et al. (2011).

Hierarchical clustering uses the matrix of pointwise dissimilarities to create a hierarchical structure of the data. The result is usually visualized in a dendrogram, where the root node represents the whole data set and each leaf node represents a single data point. The internal nodes depict the proximity of data points to each other, whereas the dendrogram's height represents the distance. Depending on the threshold determining the "cutting" level, different numbers of clusters can be obtained. The relationship between data points can be extracted from their relative position to each other.

The partitional clustering method k-means, in its original form suitable for numerical data, has some major disadvantages. The first one is the random selection of initial centroids (i.e., cluster centers), which may lead to different outcomes of the clustering of the underlying data set. Solution approaches to solve this problem are, e.g., given by Arbelaitz et al. (2013). Second, the number of clusters k is a hyperparameter which needs to be set in advance. Suitable choices for k can be deducted from the elbow method. Third, k-means is, just as hierarchical clustering methods, sensitive to outliers and noisy data, which has to be considered during data preprocessing. It can be deviated by choosing enhancements of k-means as in, e.g., Ball et al. (1967). While density-based clustering methods such as DBSCAN do not need the number of clusters in advance and are able to label outliers as noise, they usually require finetuning of the hyperparameters search radius and neighborhood density.

Depending on the choice of algorithm and the specific choice of involved hyperparameters, the clustering results may vary significantly. Approaches to validate hierarchies, partitions, and individual clusters are, e.g., presented by Jai et al. (1988).

Hence, suitable algorithms have to be identified for the respective task to be solved, a thorough testing for tuning hyperparameters has to be performed and different clustering methods have to applied for the robustness of conclusions.

3. LEARNING PLATFORM: ARCHITECTURE AND USE CASE DESCRIPTION

The major part of data necessary for learning analytics stems from the learner's activities within an e-learning course. The conceptualization and implementation of the learning platform's architecture focused on a learning management system (in our case: Moodle) as both activity and data provider, see Trommer et al. (2024). The data includes information on when each learner visited which page but also which input was made by each learner (e.g., file uploads, results of quizzes taken). The data is transferred to a Learning Record Store via xAPI, which provides the main part of the database. The database can be enriched by data from other sources like quiz results from classroom teaching. The analytics tools, implemented in Python, have both reading and writing access to the database for retrieving the data for analytic purposes and for storing learning analytics' results. The teacher is able to use those results in preparation of classroom teaching lessons following the e-learning courses and to improve the whole module's quality.

The use case presented in this paper refers to a course on Scientific Writing and Spreadsheet Calculation. The course covers one semester of teaching and addresses first semester non-computer scientist students. The cohort encompassed 43 students.

The learning material (which was already available from earlier semesters in the form of scripts, videos, assignments, and tasks) was transformed into a learning pathway consisting of different components (and enriched if necessary): sections (containing the following components), learning content (in the form of, e.g., text and videos), folders (to provide files for download), quizzes, digressions (optional content to gain deeper knowledge on certain topics), assignments (complex tasks to be solved), and feedback. The components containing mandatory learning content are represented in a learning pathway of linear structure. Regarding the progression in the course, simple rule-based progression was implemented as well:

1. Depending on the learner's result in a quiz (always following the main learning content), a specific recommendation was presented, e.g., to repeat the quiz, to have a closer look again at the learning content or to join the upcoming classroom tutorial.

2. Following an assignment, a downloadable solution approach was only available after the upload of the learner's solution and automatic check whether a certain number of points was achieved.

In addition to the linear learning pathway, voluntary content was included as well. An exemplary excerpt of the course's learning content is depicted in Fig. 1: The learning pathway includes the main components Files (F), Content (C), Quiz (Q), Digression (D) and Assignment (A). The junction Rule-based (R) implies that the learning content presented next to the learner depends on the result in the component he or she visited (in this case, depending on the quiz result, the learner revisits the learning content, repeats the quiz or continues). The junction X-OR (X) allows the learner to decide which path he or she wants to follow (in this case, the learner can view further content in the digression or continue with the assignment). We used a slightly adapted approach to formulate learning pathways as presented by Bargel et al. (2012). Within those main components, detailed conjunctions of sub-components are implemented as well.

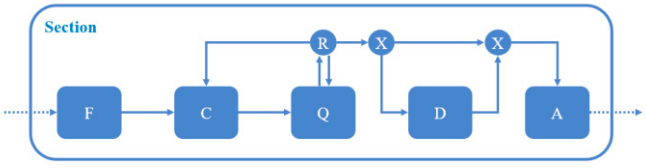

Figure 1. Excerpt of a learning pathway.

Note that the learners always had full access to the whole learning content (except for the solutions of the assignments) and were able to move non-linearly (including backwards) within the learning pathway. This approach was chosen after careful consideration to gain additional information on the learners' motivation to finish content early. Additionally, the learning volume was uniformly distributed throughout the semester. For each quiz, a final deadline to finish at least one attempt was given (which was also evenly distributed throughout the semester). The deadline for uploading the assignments was set to the end of the semester.

By applying Learning Analytics to the learner's data we now investigate: if the difficulty levels of the tasks were appropriate, how the learners used the platform, which knowledge could be used to improve the classroom teaching and how the learning pathways could be optimized.

4. ANALYSIS AND EVALUATION

The goal of this analysis is to get a deeper understanding on the learners' achievements and to derive optimization potential for the e-learning course as one step towards individualized learning pathways. Hereby, cohort-specific and learner-specific evaluations can be distinguished. In the following, we will concentrate on a cohort-specific analysis, which includes evaluations regarding the tasks solved by the learners.

We follow the procedure by Xu et al. (2005) to apply clustering algorithms to a given data set.

4.1 Basis: Data Samples

We start with a thorough description of the data base. The learners' data is stored in a database as explained in Section 3. For evaluation, this data is extracted in csv format. Besides the ID (which serves as primary key), the main important information is provided by timestamp, actor, verb, result, and object. The timestamp is given in date time format W3C. Actor contains information on the user and the account. Verb provides information on the activity performed by the user, i.e., did the user view a site or perform a login or logout. Object contains the name of the actual object the action was performed on (e.g., the user viewed "Learning Content 1") and the platform used (in our case, this is the learning management system "Moodle"). Result contains information depending on verb and object. If a task within a quiz was "answered", result contains information on the response (i.e., the chosen or provided answer of the learner), the completion of this task and the success of solving this task. For a "completed" quiz, the achieved score (absolute and relative), the minimum and maximum score possible, the completion of the quiz, the success (depending on an a priori threshold set by the teacher), and the duration of providing all answers in this quiz are depicted. Finally, context provides meta data for example on the session ID, the language, and the parent section as well as the respective short IDs of the current object. Within the latter five variables, the data is stored in json format. Note that the database contains much more (meta) data than described above. Exemplary content of the variables is summarized in Table 1.

Table 1. Data provided by the learning management system

Variable	Content	Data type	Example
Time-stamp	Timestamp	Date time object	2024-01-19T13:00:03+00:00
Actor	User ID	String	"User52"
Verb	Activity	String	"enrolled to", "logged into"
Result	Achieved score	Float	6.5
	Max. possible score	Float	10
	Completion	Boolean	True
	Duration	String	PT240S

For a thorough analysis of the learners' activities, it is essential to have additional information on the questions. This includes question type (e.g., multiple choice, free text, matching), assignment of the question within the course (e.g., an assignment to a chapter), and in the context of multiple choice the correct answers and the distractors.

Within Moodle, it is possible to construct quizzes by randomly selecting a defined number of tasks from a question pool. The question data, including the data mentioned above, is stored in xml format, can be extracted and integrated with the data originating from the learning management system.

This data can be enriched with further information provided by the teacher, who for example has a predefined classification of the tasks' levels of difficulty or a different construction of tasks with distractors of varying difficulty levels for similar questions.

4.2 First Step: Feature Selection and Extraction

Data preprocessing included distribution and outlier handling (for k-means and hierarchical methods), transformation as well as normalization of numerical data. It became apparent that a few learners did participate only in a very small number of activities. These learners had to be excluded from the cohort-specific evaluation such that the data base investigated reduced to 40 instances. Feature selection aims at reducing the dimensionality of data sets by identifying the most relevant features and reducing the subsequent data analysis to those ones. Removing redundant and non-informative features can be achieved e.g. by identifying as well as removing highly correlated features. Dimensionality reduction projects the given data to a lower-dimensional subspace to capture the most relevant information, which can result in new combinations of the original input features. A method typically applied is Principal Component Analysis (PCA).

We start with the cohort-specific evaluation of the learners' data. This data includes information on the overall learning pathway of the learners and their results. The relevant features were deduced from the provided data base (note that based on the data depicted in Table 1, we already performed a preselection, e.g., account was removed, or used the respective features to identify relevant content in other columns, e.g., using activity to identify instances containing quiz results). In this context, it easily becomes obvious that a few features are redundant: a pairwise comparison led to a correlation coefficient r=0,96 for the features total number of quizzes taken and number of quizzes taken in Section 3; hence, the redundant information provided by number of quizzes taken in Section 3 was excluded from the feature set. Specifically, the features, with the respective data type depicted in brackets, are:

- For each mandatory learning content: time spent in this content (float; calculated from *timestamp* the view of the content was started and finished (date time)), total number of visits (integer).
- For each voluntary learning content: time spent in this content (float; calculated from *timestamp* the view of the content was started and finished (date time)), total number of visits (integer).
- For each quiz: number of rounds this quiz was performed (integer; calculated from a combination of *activity* and *page name*), maximum/minimum/average points achieved in this quiz (float), threshold for success (float), average duration to conduct one quiz (float), minimum/maximum/average time needed to solve one quiz (float).

This provided the basis for a deeper analysis of these features and their interconnections. PCA clarified which quizzes had higher impact. Analogously, the impact of the final voluntary learning content was very small (only very few learners visited it), hence it was removed from the feature set.

As for each quiz, the duration is explicitly provided in the data base, whereas the time spent in a learning section must be derived from the timestamps (see Table 1) which mark the beginning of the time in the respective object. Here, it has to be taken into account that learners do not always log out properly or leave the page after having finished a learning session. Hence, depending on the content, we set specific thresholds marking realistic durations and excluded the values exceeding this threshold from our further analysis (for learning content pages containing videos, the threshold value was based on the video's duration; for text pages, a threshold based on the length of the content was created). Additionally, from the deadline to solve this quiz the remaining time until the deadline was reached could be calculated.

Regarding the tasks, the following data is available: question type (categorical), topic (categorical), percentage of correct answers (float), average time per learner spent in the respective learning content (float). Depending on the question type, additional information was available: for example, the number of distractors and their respective difficulty levels for selection of missing word and matching tasks or the number of correct items (both absolute in relative with respect to the total amount of items presented) for multiple choice tasks. For each question type and for each question, the difficulty also has to be taken into account. For reasons of comparability, this information was condensed to a new feature called difficulty of ordinal scale {1, 2, 3, 4, 5} with 5 being the highest ranking. Altogether, 89 instances were available.

The database with the learners' activity data (as shown in Table 1) consisted of about 20.000 instances for one learning unit; one quiz consisting of ten questions led to about 15 instances (including data on the start and end of the quiz, the results for each question and the quiz result).

4.3 Second Step: Application of Selected Algorithms

To get a deep understanding of the patterns in the data, clustering algorithms from different categories as described in Section 2 were applied to the two data groups learners and tasks. As partitioning method, the standard k-means algorithm was implemented for k=3 (and also for different number of clusters for validation, which is described in Section 4.4). A hierarchical approach was implemented in the form of agglomerative clustering, i.e., starting with each datapoint as its own cluster, these clusters are iteratively merged. The distance between clusters was calculated using Ward's variance minimization (instead of measuring the distance directly), which is suitable for quantitative features. The resulting tree showed high similarities between different data points, which were assigned to three clusters. Density-based clustering was implemented using DBSCAN, which did not perform well due to the data's structure (DBSCAN is not able to distinguish clusters of similar density, see Ester et al. (1996)). A deeper analysis of these results follows in the upcoming sections.

4.4 Third Step: Validation

As it was already mentioned in Section 2, the application of different clustering algorithms or of the same algorithm using

varying hyperparameters usually leads to different clustering results (e.g., the number of clusters or the assignment of data points to certain clusters or as noise). Hence, the results have to be assessed with respect to validity and usefulness.

The first question refers to the number of clusters. Regarding k-means, the elbow method suggested three clusters as the most suitable choice for learners as well as tasks. This was confirmed by a post analysis of the applied hierarchical clustering method: The threshold for agglomerating clusters, which is applied to the distance matrix, can be varied. This is visualized in the dendrogram depicted in Fig. 2 for tasks: Each vertical line represents one data point. The coloring marks the data points' affiliation to the clusters, and the height of the dendrogram indicates the order in which the data points were agglomerated. The horizontal dotted line represents the threshold or "cutting" level, in this case leading to the three clusters. The solution was evaluated in connection with scatter plots for different features and a heat map.

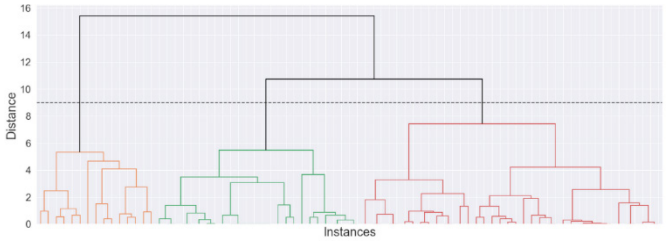

Figure 2. Hierarchical clustering.

4.5 Fourth Step: Interpretation of the Results

Considering the cohort-specific analysis, the assignment of learners to three clusters provided a good understanding of the learners' behavior. We were able to identify three groups of learners, where the related available data allowed us to define the following labels:

1. Highly motivated learners: This group consists of learners who showed a positive learning behavior, whose goal was to reach high number of points in the quizzes, solved additional tasks and uploaded the assignments when the respective section was covered (and not at the end of the semester). With 12,5 %, these were the minority of learners.
2. Target-oriented learners: These are the ones focusing on the teacher's formulated goals to achieve a certain amount of points within the quizzes who showed minimalistic behavior, i.e., they did not try to improve non-optimal quiz results, they solved quizzes very close to the deadline, the time spent in the LMS's learning content was about 30 % less than the one of the first group, and they did not make use of the voluntary offers like digressions. These were 27,5 % of the cohort.
3. Typical learners: These are learners who represent the average type. Since this was the largest group with 60,5 %, the learners' behaviour within this group was quite diverse, especially with respect to the average number of points achieved in each quiz. They however tended to repeat quizzes quite often.

Regarding the evaluation of the tasks, we want to start with the following remark. According to the teacher of the use case course, the vast number of questions provided in this module were of low difficulty. This is confirmed by investigating the learners' results: about 96 % of the quizzes taken were solved successfully (i.e., the learner achieved at least 60 % of the points). Hence, the questions do not differentiate well, which reduces the questions' significance in the evaluation process. However, the different question types are mixed within the clusters. The dissimilarity stems from other factors: Even for seemingly very easy tasks, unusual structural components (like the absence of distractors which means that all provided answers were correct solutions) increased the number of mistakes significantly. If the distractors were chosen carefully, learners tended to make more mistakes. We specifically checked questions solved incorrectly by learners clustered into "highly motivated learners" and found that the formulation of two of these questions could be improved to avoid confusion.

Based on the identified clusters of learners, it is now the question how this information can be used expediently. First, new learners who provide some information (for example because they solved an entry test) can be clustered according to this knowledge such that their content along the learning pathway is adapted according to knowledge gained from the respective cluster's data. In this context, the label-defining features are of high interest since it reduces the effort of data acquisition. This approach corresponds to supervised learning, where the cluster labels (as results from unsupervised learning) are used as the target variable for the subsequently applied supervised learning.

One way to achieve this is the application of learning vector quantization, see Kohonen (1995). Here, the training phase is used to determine prototypes representing the data set's classes. The classes for previously unknown instances are then predicted in accordance with these prototypes and their receptive fields, i.e., the prototype being the best representative for this instance based on the chosen metric. Each data point is affiliated to one of the three clusters produced by the hierarchical clustering algorithm. The clusters were added as labels to the data, which was subsequently split into training and test data sets. The prototypes derived from applying a generalized relevance learning vector quantization algorithm to the training data set led to a 100 % correct classification of the remaining test data (and 89 % for learners). The classifier was validated using typical approaches such as accuracy (Shalev-Shwartz et al. (2014)).

This approach provides an efficient method to assign new tasks as well as new learners to existing clusters using basic information like task type and difficulty for tasks or using data gathered from short entry courses for learners.

5. SUMMARY AND OUTLOOK

In this contribution, we explored AI-based learning analytics in higher education. The scenario involved learning data provision and the essential data preparation. In this context, we explored the utilization of various clustering techniques on learners' data collected with the help of learning pathways in an e-learning course, followed by an examination of the

outcomes. The focus was set on getting a better understanding of learners' success by identifying similar behaviors with respect to their activities within the platform and learning results. The findings can be used both for the improvement of the learning material as well as the preparation of subsequent classroom teaching courses in presence. Besides the results presented here, a major part of our investigations was also learner-specific analysis, to derive individual patterns on preferred learning time windows, durations to complete the lessons, and the development over time. The detailed learners' time data analysis was not specified in this paper but provide valuable insight for further analysis in connection with the results presented here. Our future tasks will contain acquiring learner-specific data like preferred learning styles and times to study. Additionally, for each task within the quizzes, we plan to include a self-evaluation of the learners on how sure he or she is the given answer is correct. This allows a better evaluation if our findings are valid and detailed feedback for the learners to enhance self-reflection.

Since most of the questions used in the quizzes currently available are of low difficulty, we intend to use them as entry test, which helps to get a first evaluation of the learner's knowledge level. In connection with the intended implementation of AI-based live evaluation of the learners' data, this also allows us to improve the individualization of the tasks and recommendations.

In the context of creating prototypes for each cluster to correctly classify new data, the next step is the identification of relevant features. This will minimize the effort of gathering data on new instances to be classified while it simultaneously maximizes the classification' precision.

Acknowledgments. The project on which this publication is based was funded by the German Federal Ministry of Education and Research under grant number 16DHBKI063. The authors are responsible for the content of this publication.

REFERENCES

Arbelaitz, O., Gurrutxaga, I., Muguerza, J., Pérez, J.M., and Perona, I. (2013). An extensive comparative study of cluster validity indices. *Pattern Recognition* 46 (1), 243–256.

Ball, G. and Hall, D. (1967). A clustering technique for summarizing multivariate data. *Behavioral Sciences* 12, 153–155.

Bargel, B.A., Schröck, J., Szentes, D., and Roller, W. (2012). Using Learning Maps for Visualization of Adaptive Learning Path Components. *International Journal of Computer Information Systems and Industrial Management Applications* 4, 228–235.

Bishop, C.M. (2006). *Pattern Recognition and Machine Learning*. 1st edn. Springer, New York.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). Deep Learning. The MIT Press.

Cherkassky, V. and Mulier, F. (1998). *Learning From Data: Concepts, Theory, and Methods*. Wiley, New York.

Dhieb, N., Ghazzai, H., Besbes, H., and Massoud, Y. (2020). A Secure AI-Driven Architecture for Automated Insurance Systems: Fraud Detection and Risk Measurement. *IEEE Access* 8, 58546–58558.

Ester, M., Kriegel, H.-P., Sander, J., and Xu, X. (1996). A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. *Proceedings of 2nd International Conference on Knowledge Discovery and Data Mining*.

Jain, A. and Dubes, R. (1988). *Algorithms for Clustering Data*. Prentice-Hall, Englewood Cliffs, NJ.

Kerres, M. (2001). *Multimediale und telemediale Lernumgebungen. Konzeption und Entwicklung*. 2nd edn. Oldenbourg, München.

Kohonen, T. (1995). Learning vector quantization. In: Arbib, M. (eds.), *The handbook of brain theory and neural networks*. 537–540. Cambridge, MA: MIT Press.

Kriegel, H.-P., Kröger, P., Sander, J. and Zimek, A. (2011). Density-based clustering. *WIREs Data Mining Knowledge Discovery* 1, 231–240.

Rannertshauser, P., Kessler, M., and Arlinghaus, J.C. (2022). Human-centricity in the design of production planning and control systems: A first approach towards Industry 5.0. *IFAC Paper-sOnLine* 55 (10), 2641–2646.

Scott, P.H. (1991). Pathways in Learning Science: A case study of the development of one student's ideas relating to the structure of matter. In: Duit, R., Goldberg, F., Niedderer, H. (eds.) *International Workshop on Research in Physics Learning: Theoretical Issues and Empirical Studies*. University of Bremen.

Shalev-Shwartz, S. and Ben-David, S. (2014). *Understanding Machine Learning. From Theory to Algorithms*. 1st edn. Cambridge University Press.

Srinivasa, K.G. and Muralidhar, K. (2021). *A Beginner's Guide to Learning Analytics*. Springer Nature, Cham, Switzerland.

Teich, L., Franke, D., Michaelis, A., Dähnert, I., Gebauer, R., Markel, F., and Paech, C. (2023). Development of an AI based automated analysis of pediatric Apple Watch iECGs. *Frontiers in Pediatrics*.

Trommer, M., Franke, S., Teich, T., Riedel, R., and Hellbach, S. (2024). Enhancing Educational Outcomes through Learning Pathways and Artificial Intelligence-Driven Learning Analytics. *STE2024*, Helsinki, Finland.

Xu, R. and Wunsch, D. (2005). Survey of clustering algorithm. *IEEE Transactions on Neural Networks* 16 (3), 654–678.

Yang, F. and Dong, Z. (2017). *Learning Path Construction in e-Learning*. Springer, Business Media, Singapore.