Towards Explainability in Modern Educational Data Mining: A Survey

Basile Tousside, Yashwanth Dama and Jörg Frochte
Bochum University of Applied Science, 42579 Heiligenhaus, Germany

Keywords: Educational Data Mining, Data Mining, Explainable Artificial Intelligence.

Abstract: Data mining has become an integral part of many educational systems, where it provides the ability to explore hidden relationship in educational data as well as predict students' academic achievements. However, the proposed techniques to achieve these goals, referred to as educational data mining (EDM) techniques, are mostly not explainable. This means that the system is black-boxed and offers no insight regarding the understanding of its decision making process. In this paper, we propose to delve into explainability in the EDM landscape. We analyze the current state-of-the-art method in EDM, empirically scrutinize their strengths and weaknesses regarding explainability and making suggestions on how to make them more explainable and more trustworthy. Furthermore, we propose metrics able to efficiently evaluate explainable systems integrated in EDM approaches, therefore quantifying the degree of explanability and trustworthiness of these approaches.

1 INTRODUCTION

Data mining, also referred to as Knowledge discovery in data (KDD), has become a standard when dealing with large datasets. It is the process of extracting useful information from a large set of data, tuning those information into valuable insights and predictions. When data mining is applied to educational datasets – often collected via a learning management system (LMS) platform, it is referred to as educational data mining (EDM).

The growth of online learning over the past two years due to COVID-19 has made it much easier to collect a mass of educational data set in universities and other educational facilities. This mass of data has whetted the appetite of educational data mining researchers, producing a boost in the field. However, educational data mining is not a new paradigm in itself. Its origin goes back to the year 2000 when it was briefly addressed during research on the intelligence of tutorial systems whose results were presented during the workshop on "Applying Machine Learning to ITS Design/Construction" hosted in Montreal, Canada.

During the 2000s, several other workshops on data mining in education were held, including the "Educational Data Mining" workshop created in 2005 and hosted in Pittsburg, USA. Two years later, a similar and complementary workshop named workshop on "Applying Data Mining in e-Learning" was held for the first time in Crete, Greece, in 2007. Most of these workshops have evolved into conferences and are known as such nowadays. This is the case of the "International Conference on Educational Data mining", which takes place this year in London, United Kingdom. In recent years, the topic of EDM has become increasingly important and several new conferences have been born. The most prominent among these conferences are the "International Conference on Artificial Intelligence in Education (AIED)", the "International Conference on Learning Analytics & Knowledge (LAK)" and the "International Educational Data Mining Society" founded in 2011. Over the years, the idea and purpose of the EDM has evolved significantly. In its early days, it was limited to predicting students' performance in specific courses. During the last few years, it evolved into improving the educational process and explaining educational strategies for better decision-making (Silva and Fonseca, 2017). This was achieved via the development and adoption of statistical, machine-learning and data-mining methods to study educational data generated by students and instructors.

Currently, the main metric used in EDM is the overall prediction accuracy. However, educational data mining exhibits a multi-targeted problem and only focusing on the accuracy might be misleading. In this paper, we show that in addition to accuracy, explainability needs to be considered to better address problems in educational data mining.

Tousside, B., Dama, Y. and Frochte, J.
Towards Explainability in Modern Educational Data Mining: A Survey.
DOI: 10.5220/0011529400003335
In *Proceedings of the 14th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2022) - Volume 1: KDIR*, pages 212-220
ISBN: 978-989-758-614-9; ISSN: 2184-3228

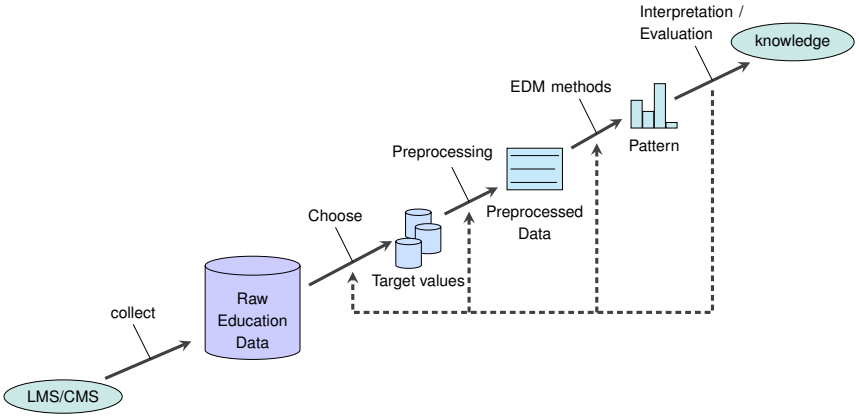

Figure 1: Educaton Data Mining (EDM) pipeline: Overview of how EDM techniques are applied from data gathering in learning environments such as learning management system (LMS) platform to knowledge discovery. The dashed lines indicates revisiting previous state of the process, when the result of the current state is not suitable.

Explainability is a relatively new paradigm in data mining that aims to shed light on the decision-making process in predictions made by intelligent systems. In the context of EDM, it aims to explain for example, why a data mining algorithm predicts that a student will not pass the mathematics subject. This is a real challenge since modern algorithms are so sophisticated that their decisions have become too complex to express in humans understandable terms. However, significant progress has been made in the explainable artificial intelligence community and explainability techniques have been successfully applied to ML models. The question of whether these techniques can be used in EDM is a major concern in this paper. A review of those techniques will be presented in Section 2. Furthermore, section 3 will give a detailed overview of the state of explainability in educational data mining.

Another artificial intelligence discipline which is closely related to EDM is Learning Analytics (LA). According to the call of the First International Conference on Learning Analytics and Knowledge (LAK), LA can be defined as the measurement, collection, analysis and reporting of data about learners and their contexts, to understand and optimise learning and the environments in which it occurs (Calvet Liñán and Juan Pérez, 2015). The learning analytic process is the same as the data mining process and can be summarized as: an iterative process in which data is extracted from an educational environment and preprocessed before applying quantitative methods to help instructors in decisions making. A schematic representation of this process is shown in Figure 1.

In this paper, we focus on EDM and address the limitations of the prediction accuracy, which is the current state-of-the-art metric in evaluating and analysing educational data mining techniques. Furthermore, we give an overview of the state of explainability in the EDM landscape. As will be seen in Section 3, when designing EDM approaches, explainability can be combined with prediction accuracy to overcome the shortcomings of the latter. The main contributions of this paper is as follows:

- We Provide a survey on EDM and compare existing EDM techniques.
- We define explainability and review the recent improvement made in explainable artificial intelligence.
- We review the current state of explainability in modern EDM approaches.
- We discuss the multi-dimensional requirement for educational data mining, pointing out the limitation of the prediction accuracy metric and propose to integrate explainability in future EDM techniques in order to fill the accuracy gaps.

2 EXPLANABILTIY IN GENERAL

Despite being opaque and difficult to interpret, data mining models have spread out in many applications from different domains such as education, healthcare, criminal justice or autonomous driving. They help educators predicting students' grade, doctors making diagnostics or judges taking decisions. This raises concern and demand for humanly understandable decisions of these artificial intelligent systems.

To satisfy this demand, the concept of eXplainable Artificial Intelligence (XAI) has emerged with the goal of producing AI models whose decisions are interpretable and understandable by a non-AI-expert human. A doctor should, for example, understand

why an AI model diagnoses a tumour rather than simply operates on a patient because the model said so. A judge using AI-based software to pass judgement should be able to explain the motives and facts behind that decision. An educator or visa officer should understand why a model predicts that a student will not succeed in his studies, instead of simply rejecting the application because the model recommends to do so.

On the other side, there is a growing number of legal regulations restricting how companies and organizations make use of AI-based decisions. The EU's General Data Protection Rights (GDPR), – which contains a "right to an explanation" article – for example requires a human to examine and confirm the conclusions reached by an AI algorithm before applying them. Another example is the "algorithmic accountability bill" of the city of New York, which requires the fairness and validity of AI-based decisions to be verifiable.

Due to these critical needs for ethics, fairness and transparency of AI systems, there has been a considerable increase in research interest in explainability methods for interpreting and understanding black-boxed automated decisions. This has led to some convincing model explanation algorithms, especially in the supervised learning domain (Burkart and Huber, 2021). Some of these algorithms have a promising future in the EDM field. In Section 3, we will elaborate on which XAI algorithms we believe could be integrated to which EDM approach to make it more explainable.

One problem that can pose difficulties in the acceptance of this explainable methods is that the explainability capability of these algorithms is sometimes subjective since there is currently no consensus on a definition of explainability. In 2.1 and 2.2, we will review some conceptual and formal popular meaning and definition behind explainability.

Another open challenge in explainable artificial intelligence is to quantify explainability. We believe a standard metric will be difficult to attempt until a consensus has been made on a concrete definition of explainability. However, some promising evaluation metrics have emerged recently and will be presented in 2.3.

2.1 Conceptual Definition of Explainability

Although the goal of explainable artificial AI is clear (making decisions or actions of AI systems – EDM systems for example – explainable to a human observer), there is currently no consensus on how to define explainability. This is emphasized in a recent paper (Miller, 2019), where the authors point out the fact that most works in eXplainable AI are based only on the researchers' intuition of what constitutes a good explanation. However, people select, represent or comprehend explanations depending on the discipline they are related to. From psychology, according to (Lombrozo, 2006), explanations are the currency in which we exchange beliefs. This definition emphasizes the need for high confidence in explanation. More recently, (Miller, 2019) formulates explainability as the degree to which a human can understand the cause of a decision. In a similar vein, (Kim et al., 2018) define explainability as the degree to which a human can consistently predict the model result. A definition of explainability in relation to AI is given by (Islam et al., 2019) as the extent of transferable qualitative understanding of the relationship between model input and prediction in a recipient-friendly manner. (Ribeiro et al., 2016) link explainability to visual perception. The authors of the paper argue that explaining a decision means presenting visual artifacts that provide qualitative understanding of the relationship between the instance's components (e.g. words in text, patches in an image) and the model's prediction. We believe a concise definition of XAI should take these rich and diverse viewing angle into consideration.

2.2 Formal Definition of Explainability

Consider a standard supervised learning problem – as it is almost always the case in educational data mining – and let denotes $f : \mathcal{X} \mapsto \mathcal{Y}, f \in \mathcal{F}$ the supervised model trained with structured data-set $D_{train} = \{x_n, y_n\}_{n=1}^{m}$, which maps input feature $x \in \mathcal{X}$ to targets $y \in \mathcal{Y}$. In this paper, we focus on local explanation, which is the most common formal definition of model interpretability. If $\mathcal{F}$ is complex, a local explanation can be generated to understand its behavior in some neighborhood $N_x \in P[\mathcal{X}]$, where $P[\mathcal{X}]$ denotes the space of probability distribution over $\mathcal{X}$. In explainable artificial intelligence (XAI) literature, systems that produce local explanations – local explainers – are often denoted as $\sigma : \mathcal{X} \times \mathcal{F} \mapsto \xi$, where ξ is the set of possible explanations (Plumb et al., 2019). It is worth mentioning that the choice of the explanations system ξ greatly depends on whether or not feature in D_{train} are semantics. In XAI, features are said to be semantic if one can reason about them and understand the meaning of change in their values (e.g. a student's age, grade or number of siblings).

To understand the decision made by a data mining model, the vast majority of local explainers try to pre-

dict how the model's output would change if its input were perturbed to some degree (Ribeiro et al., 2016; Plumb et al., 2018). The explainer output space can then be formalized as $\xi_o := \{g \in \mathcal{G} \mid g : \mathcal{X} \mapsto \mathcal{Y}\}$, where $\mathcal{G}$ denotes a class of interpretable functions. Since the vast majority of features encountered in EDM literature is semantic, we believe integrating local explainers into EDM methods should be a fairly convenient and promising task.

2.3 Quantifying Explainability

While explainability tells the how and why regarding the decision of an AI model, it does not tell us much about how much trust we can have in that decision. As a result, the question of how to quantify explainability is still an open challenge. In recent years, some metrics have been proposed to quantify proposed explainability systems. The rest of this section will present two of the most popular metrics.

Fidelity Metric. The fidelity metric evaluates explainability by measuring how accurately g models f in a neighborhood N_x (Ribeiro et al., 2016; Plumb et al., 2019). It can be formally defined as:

$$F(f,g,N_x) := \mathbb{E}_{\tilde{x}\sim N_x}[(g(\tilde{x}) - f(\tilde{x}))^2]. \quad (1)$$

In this context, a lower fidelity value (lower is better) means that the explainability system is able to accurately determine which patterns were relevant to the model for making this prediction. In the context of EDM, an explanation with good fidelity accurately identifies which pattern in the data the model used to predict the student grade or overall student academic year achievement, for example.

Stability Metric. The stability metric has been proposed by (Alvarez Melis and Jaakkola, 2018) and consists of evaluating the degree to which the explanation changes between points in the neighborhood N_x:

$$\mathcal{S}(f,\sigma,N_x) := \mathbb{E}_{\tilde{x}\sim N_x}[\,||(\sigma(x,f) - \sigma(\tilde{x},f))||_2^2\,]. \quad (2)$$

In this setting, the lower the stability value (lower is better), the more reliable and trustworthy the model explainability.

In the next section, we will analyse explainability in EDM. Doing so, we will give some suggestions on how to integrate explainability system in EDM approaches where it is missing. For the evaluation of such explainability system, the tools described in this section can be used.

3 EXPLAINABILITY IN EDUCATIONAL DATA MINING

Explainable AI (XAI) is a relatively new field that has not yet really penetrated the EDM field, although it would be of great importance in achieving the EDM objectives. Generally, the objective of EDM is to improve the educational system by assisting administrators and educators in decision-making tasks, like, grades prediction (Bhopal and Myers, 2020), predicting student dropout (Gardner et al., 2019), education admission decisions (Marcinkowski et al., 2020) or forecasting on-time graduation of students (Hutt et al., 2019). In all these tasks, the current state of research in EDM considers accuracy as the main metric to evaluate the quality of a model. This allows to gain great insight since accuracy is important. However, EDM often infers a multi-target problem, therefore, just focusing on accuracy is misleading and will not be sufficient to reach the goal. This accuracy shortcoming can be overcome by integrating explanability in EDM approaches, as will be shown in this section.

The rest of this section will explore if and how explainability is integrated into state-of-the-art educational data mining techniques. A summary of this exploration on the most prominent state-of-the-art EDM techniques is given in Table 1. Before delving into explainability in the current state of EDM research landscape, let briefly discuss the motivation for using explainability in EDM.

3.1 Advantages of using Explainability in EDM

To maintain the integrity of the data and validate the ethical decision made by a model, explainable AI should be used in EDM for deriving an understanding of the important features that might be directly effecting the students' performance and help them in choosing suitable courses of interest for example. Furthermore, instead of focusing on the students' performance and result-oriented examinations, the consideration and understanding of various attributes affecting students' performance and behavioural areas is an important aspect for instructors. This is not a trivial task since the educational dataset is dynamic, as it consists of varying attributes such as location, courses chosen, behavioural factors, historical performance records and many more for each and every individual student. There is a massive necessity for supporting students' achievement, whether good or weak, answering the question why and also discern the information for betterment of students' performance. Explainability also determines the model's prediction

strategy's intrinsic bias (Gaur et al., 2020). Moreover, explainability opens new research areas and ensures decision-making based on domain knowledge.

In a nutshell, eXplainable Artificial Intelligence (XAI) would not only provide the user with the performance predictions but also necessitates the evaluation of underlying features which play the medium for reaching out to specific predictions.

3.2 EDM Approaches Adressing Explainability

Recent research in educational data mining used attention models to leverage a certain degree of interpretability and visibility of words in sentences to which the model predicts responses. This is often achieved by translating the model into responses to questions where sentences are separated by *OR* clauses. In (Gaur et al., 2020) for example, the authors describe possible ways of infusing knowledge graphs in deep learning (DL) models using knowledge-aware loss function or knowledge-aware propagation function. The former function can be used to calculate the deviations in the information at each and every layer of the DL model and for every epoch during learning. The latter can interpret the lost information and transfers the lost information using mathematical procedures. The paper describes a use case on students' academic performance with academic domain knowledge as knowledge graph. This allows to infuse deep learning to the model for not only making predictions but also estimate the incorrect decisiveness of meta concepts when solving a test and provide us with the students' knowledge. For example, a student is asked to solve a problem encapsulating concepts of quadratic equations and physics, the correctness to the answer explains his domain knowledge in either one of the concepts or both. This method provides a certain level of explanation of the model decision making process.

Another paper by (Hasib et al., 2022) approaches the explainability problem in EDM in a very different way. The authors described the approach of utilizing five different data mining algorithms such as Logistic Regression, K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Naive Bayes and XGBoost for classification task to predict the accurate categorical estimations of five classes (Excellent, Good, Satisfactory, Poor and Failure) of high-school student dataset on two courses. The proposed experiments showcase an excellent accuracy in utilizing the SVM approach and provide a comparative analysis of the five algorithms used. To explain the integrity of predicted results, the author trained LIME (Local Interpretable Model-agnostic Explanations), a popular local explainer model, to understand the changes in the data and the effect of these changes on the prediction results for five classification algorithms. However, LIME is better suited for image data-set and has shown some limitations when applied to EDM data.

3.3 EDM Approaches Lacking Explainability

The vast majority of approaches in educational data mining concentrate on the accuracy of the model being deployed, therefore largely neglecting its explainability. This result in EDM approaches performing well in regard of their accuracy, but they are not explainable. This lack of explainability goes against recent legislation like the EU's "General Data Protection Right" (GDPR), for example, which requires the decision making process of an AI algorithm to be humanly understandable.

One of such EDM technique exclusively focusing on accuracy has been proposed in (Asif et al., 2017), where the authors concentrated on predicting the academic achievements of students at the end of a four year study program. This is a multi-target problem since it includes predicting the student grade at each course for example. The main finding of the paper was the importance for the educators to concentrate on a small number of courses showing particularly good or poor performance. Doing so, support can be offered to under-performing students as well as advice or opportunities to high-performing students. Their model which mainly built on decision tree and random forest does not tackle the explainability issue. However, this will be an interesting feature, especially in the case of a multi-target problem as described in their paper. Fortunately, their method is lightweight and integrating an explainability module like EXpO (Explanation-based Optimization), which has been presented in (Plumb et al., 2019), should be fairly straightforward.

An approach with a similar objective has been proposed in (Burgos et al., 2018). The goal here was to predict the overall grade a student would get in the subsequent semester. Here again, the approach focuses on the model prediction accuracy at the expense of its explainability. Although this is a single-target problem in contrast to the previously mentioned approach by (Asif et al., 2017), which was a multi target problem, understanding the algorithm's decision-making process would be a valuable asset. This should be conveniently achievable with an explainer like FTSD (Functional Transparency for Structured Data) proposed in (Lee et al., 2019).

Table 1: State-of-the-art Approaches for Educational Data Mining. For each method, we describe the task tackled, the data mining methods used, and whether or not explainability has been addressed.

Methods	Tasks	Methods	Explainability
(Hasib et al., 2022)	Regression, Classification	KNN, XG-Boost, SVM, Logistic Regression and Naive Bayes	√
(Hoffait and Schyns, 2017)	Classification	RF, LR, ANN	×
(Fernandes et al., 2019)	Classification	Gradient Boosting	×
(Gaur et al., 2020)	Regression	Knowledge Graphs infused DL model	√
(Cruz-Jesus et al., 2020)	Classification	DT, RF, SVM	×
(Toivonen and Jormanainen, 2019)	Classification	DT	×
(Kaur et al., 2015)	Regression	Classification Algorithms	×
(Asif et al., 2017)	Regression	DT, RF	×
(Waheed et al., 2020)	Classification	ANN	×
(Xu et al., 2019)	Regression	DT, SVM	×
(Burgos et al., 2018)	Regression	SVM, FNN	×
(Rebai et al., 2020)	Regression, Classification	RT, RF	×

Another view of the EDM problem has been proposed by (Waheed et al., 2020). They suggest a model analyzing students' records related to their navigation through a learning management system (LMS) platform. The authors demonstrate that a student's clickstream activity has a significant impact on its performance. More precisely, students who navigate through courses achieve better results. Their approach which is based on deep neural networks is complex and its decision is difficult to explain or interpret. Incorporating a neural network based local explainers like RRR (Right for the Right Reason) (Shao et al., 2021) – an explainer, which regularizes a black-box model via a regularizer that involves a model's explanations – into their model could enhance the trustworthiness of the results.

Tackling a similar problem as the previously mentioned approach, (Xu et al., 2019) proposed to model the relationship between Internet usage behaviors and academic performance of undergraduate students via decision tree and SVM. The final model therefore predicts student academic performance from Internet usage data. The authors found a positive correlation between academic performance and internet connection frequency, whereas internet traffic volume was negatively correlated with academic performance. Although they do not mention the explanation of the model in their article, their method is already intuitive in itself and its result should be fairly explainable. Nonetheless, the integration of a fairly simple explainer should not be a complicated task.

From a different perspective, (Rebai et al., 2020) proposed to identify the key factors that impact schools' academic performance and to explore the relationships between these factors. Their regression tree model featured that factors, which are the most correlated to high student performance are competition, school size, class size and parental pressure. Their model does not provide any explanation on its decision-making process. Moreover, because we could not find any implementation of their method, it is difficult to assess whether the integration of an explainability module would be good achievable or a relatively complex task.

In a related move, a model trained with demographic characteristics of students of a federal district of Brazil was proposed by (Fernandes et al., 2019). The approach which aims at predicting student's academic performance uses classification models based on the Gradient Boosting Machine (GBM). The results indicate that demographic features like neighbourhood, school and age are crucial indicators of student success or failure. In their paper, the authors provide usecase to understand the functioning of their model. However, understanding how the model works do not help to understand the model decision process when predicting a specific class. We suggest the integration of a model agnostic method like SHARP (Lundberg and Lee, 2017) to make their approach more interpretable.

A similar and more recent study based on student's demographic characteristics has been proposed in (Cruz-Jesus et al., 2020), where the authors develop a model able to predict the academic achievement of public high school students in Portugal. The dataset consisted of 16 demographics on 110,627 students among which gender, age, internet access, computer possession or class attendance. The authors implement many variants of their approach using various data mining techniques such as K-nearest neighbours, logistic regression, support vector machine or artificial neural networks. These implementations achieve a student performance prediction accuracy ranging from 51.2% to 81.1%. Interestingly, the variants of their implementation resulting in low accuracies offer a greater flexibility regarding the integration of an explainability module. On the other side, the implementations that demonstrate the best accuracies are more complex and integrating an explainability module without compromising the models accuracies is not a trivial task. This difficulty is well illustrated in their ANN model, which is a complex black box model, thus making the integration of external component difficult. A technique that could be used to infuse explainability into their neural network is the one proposed in (Plumb et al., 2019), which regularizes an ANN for explanation quality at training time.

To wrap up the exploration and importance of explainability in EDM, Figure 2 compares the neighborhood fidelity metric (presented in 2.3) of the simple FTSD explainer, which was particularly designed for structured data, and the predictive mean square error of several models trained on the UCI Educational Process Mining (EPM) regression dataset.

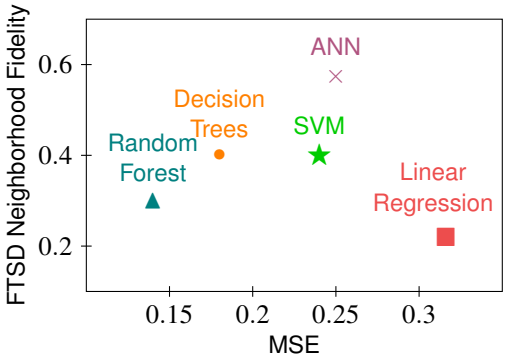

Figure 2: Neighborhood Fidelity of the FTSD explainations (lower is better) vs. predictive MSE of several model trained on the UCI "EPM" regression dataset.

3.4 Discussion and Outlook

An empirical conclusion that can be made out of this section is that when designing EDM approaches, the educational data mining community should try to find a trade-off between prediction accuracy and model explanation. Good reasons for this are on one side the constantly growing demands and legislation for humanly understandable decision of artificial intelligent system and on the other side the easier debugging and enhancement of the trustworthiness of those systems. Improving the trustworthiness of EDM approaches by making their decision process understandable is of crucial importance since EDM decisions are applied to humans, for example when deciding to accept or reject an application for a study visa or an application for a specific program.

In view of this, more attention should be paid on providing EDM approaches that are explainable, even at the accuracy cost. In designing these human understandable models, the EDM community could inspire from the few literatures in the XAI community (Alvarez Melis and Jaakkola, 2018; Plumb et al., 2019; Al-Shedivat et al., 2020) that have thematized on this accuracy-explainability trade-off.

4 CONCLUSION

In this work, we addressed the explainability problem in educational data mining. We made several observations, both conceptual and empirical, which open future research directions. In particular, we noticed that explainability is not well addressed in EDM although it would be of great importance in satisfying regulations about artificial intelligent systems as well as making EDM more trustworthy. Furthermore, we took a look at the current state of explainable artifi-

cial intelligence (XAI) research both in terms of conceptual formalism and evaluation metrics and propose some XAI methods that could be incorporated into specific state-of-the-art EDM approaches.

In the long horizon, achieving explainable EDM would have a great impact on education and should be among significant goals for a healthier and trustworthy EDM practice. Our work is a foundational research and does not lead to any direct applications.

ACKNOWLEDGEMENTS

This work is part of the Digital Mentoring project, which is funded by the Stiftung Innovation in der Hochschullehre under FBM2020-VA-219-2-05750.

REFERENCES

Al-Shedivat, M., Dubey, A., and Xing, E. P. (2020). Contextual explanation networks. *J. Mach. Learn. Res.*, 21:194–1.

Alvarez Melis, D. and Jaakkola, T. (2018). Towards robust interpretability with self-explaining neural networks. *Advances in neural information processing systems*, 31.

Asif, R., Merceron, A., Ali, S. A., and Haider, N. G. (2017). Analyzing undergraduate students' performance using educational data mining. *Computers & Education*, 113:177–194.

Bhopal, K. and Myers, M. (2020). The impact of covid-19 on a level students in england.

Burgos, C., Campanario, M. L., de la Peña, D., Lara, J. A., Lizcano, D., and Martínez, M. A. (2018). Data mining for modeling students' performance: A tutoring action plan to prevent academic dropout. *Computers & Electrical Engineering*, 66:541–556.

Burkart, N. and Huber, M. F. (2021). A survey on the explainability of supervised machine learning. *Journal of Artificial Intelligence Research*, 70:245–317.

Calvet Liñán, L. and Juan Pérez, Á. A. (2015). Educational data mining and learning analytics: differences, similarities, and time evolution. *International Journal of Educational Technology in Higher Education*, 12(3):98–112.

Cruz-Jesus, F., Castelli, M., Oliveira, T., Mendes, R., Nunes, C., Sa-Velho, M., and Rosa-Louro, A. (2020). Using artificial intelligence methods to assess academic achievement in public high schools of a european union country. *Heliyon*, 6(6):e04081.

Fernandes, E., Holanda, M., Victorino, M., Borges, V., Carvalho, R., and Van Erven, G. (2019). Educational data mining: Predictive analysis of academic performance of public school students in the capital of brazil. *Journal of Business Research*, 94:335–343.

Gardner, J., Brooks, C., and Baker, R. (2019). Evaluating the fairness of predictive student models through slicing analysis. In *Proceedings of the 9th international conference on learning analytics & knowledge*, pages 225–234.

Gaur, M., Faldu, K., and Sheth, A. (2020). Semantics of the black-box: Can knowledge graphs help make deep learning systems more interpretable and explainable?

Hasib, K. M., Rahman, F., Hasnat, R., and Alam, M. G. R. (2022). A machine learning and explainable ai approach for predicting secondary school student performance. In *2022 IEEE 12th Annual Computing and Communication Workshop and Conference (CCWC)*, pages 0399–0405.

Hoffait, A.-S. and Schyns, M. (2017). Early detection of university students with potential difficulties. *Decision Support Systems*, 101:1–11.

Hutt, S., Gardner, M., Duckworth, A. L., and D'Mello, S. K. (2019). Evaluating fairness and generalizability in models predicting on-time graduation from college applications. *International Educational Data Mining Society*.

Islam, S. R., Eberle, W., and Ghafoor, S. K. (2019). Towards quantification of explainability in explainable artificial intelligence methods. *arXiv preprint arXiv:1911.10104*.

Kaur, P., Singh, M., and Josan, G. S. (2015). Classification and prediction based data mining algorithms to predict slow learners in education sector. *Procedia Computer Science*, 57:500–508. 3rd International Conference on Recent Trends in Computing 2015 (ICRTC-2015).

Kim, B., Wattenberg, M., Gilmer, J., Cai, C., Wexler, J., Viegas, F., et al. (2018). Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *International conference on machine learning*, pages 2668–2677. PMLR.

Lee, G.-H., Jin, W., Alvarez-Melis, D., and Jaakkola, T. (2019). Functional transparency for structured data: a game-theoretic approach. In *International Conference on Machine Learning*, pages 3723–3733. PMLR.

Lombrozo, T. (2006). The structure and function of explanations. *Trends in cognitive sciences*, 10(10):464–470.

Lundberg, S. M. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30.

Marcinkowski, F., Kieslich, K., Starke, C., and Lünich, M. (2020). Implications of ai (un-) fairness in higher education admissions: the effects of perceived ai (un-) fairness on exit, voice and organizational reputation. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pages 122–130.

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence*, 267:1–38.

Plumb, G., Al-Shedivat, M., Cabrera, A. A., Perer, A., Xing, E., and Talwalkar, A. (2019). Regularizing black-box models for improved interpretability. *arXiv preprint arXiv:1902.06787*.

Plumb, G., Molitor, D., and Talwalkar, A. S. (2018). Model agnostic supervised local explanations. *Advances in neural information processing systems*, 31.

Rebai, S., Yahia, F. B., and Essid, H. (2020). A graphically based machine learning approach to predict secondary schools performance in tunisia. *Socio-Economic Planning Sciences*, 70:100724.

Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144.

Shao, X., Skryagin, A., Stammer, W., Schramowski, P., and Kersting, K. (2021). Right for better reasons: Training differentiable models by constraining their influence functions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 9533–9540.

Silva, C. and Fonseca, J. (2017). Educational data mining: a literature review. *Europe and MENA Cooperation Advances in Information and Communication Technologies*, pages 87–94.

Toivonen, T. and Jormanainen, I. (2019). Evolution of decision tree classifiers in open ended educational data mining. In *Proceedings of the Seventh International Conference on Technological Ecosystems for Enhancing Multiculturality*, pages 290–296.

Waheed, H., Hassan, S.-U., Aljohani, N. R., Hardman, J., Alelyani, S., and Nawaz, R. (2020). Predicting academic performance of students from vle big data using deep learning models. *Computers in Human behavior*, 104:106189.

Xu, X., Wang, J., Peng, H., and Wu, R. (2019). Prediction of academic performance associated with internet usage behaviors using machine learning algorithms. *Computers in Human Behavior*, 98:166–173.