Revista Intersaberes ISSN: 1809-7286
Vol. 19 224en5002 2024

FROM OPACITY TO CLARITY: DEMYSTIFYING MACHINE LEARNING IN EDUCATION WITH EXPLAINABLE AI

DA OPACIDADE À CLAREZA: DESMISTIFICANDO O APRENDIZADO DE MÁQUINA NA EDUCAÇÃO COM IA EXPLICÁVEL

DE LA OPACIDAD A LA CLARIDAD: DESMITIFICAR EL APRENDIZAJE AUTOMÁTICO EN LA EDUCACIÓN CON IA EXPLICABLE

Wellington Rodrigo Monteiro
Ph.D., Universidade Positivo
https://orcid.org/0000-0001-8450-8714
E-mail: wellington.r.monteiro@gmail.com

Eduardo Ayrosa
Ph.D., Universidade Positivo
https://orcid.org/0000-0002-9833-3756
E-mail: eduardo.ayrosa@up.edu.br

ABSTRACT

The adoption of machine learning (ML) algorithms in education is increasing, aiming to enhance teaching, learning, and administrative processes. These algorithms are crucial in personalized learning, student performance prediction, and curriculum design. However, their widespread use can lead to challenges like bias, lack of transparency, and excessive reliance on automated decisions. Educators often need help understanding the inner workings of ML models. This paper examines Explainable AI (XAI) as a solution to these issues in education. XAI techniques can provide educators and administrators with valuable insights into ML algorithms, facilitating more informed decision-making. We discuss the difference between transparent and opaque algorithmic choices and demonstrate the tangible benefits of XAI in education. Transparent models enable educators to leverage their expertise effectively, discover hidden patterns, and improve student outcomes.

Keywords: artificial intelligence; explainable artificial intelligence; machine learning; education.

RESUMO

A adoção de algoritmos de aprendizado de máquina (ML) na educação está aumentando, com o objetivo de aprimorar os processos de ensino, aprendizado e administração. Esses algoritmos são cruciais para a aprendizagem personalizada, a previsão do desempenho dos alunos e a elaboração de ementas. No entanto, seu uso generalizado pode levar a desafios como parcialidade, falta de transparência e dependência excessiva de decisões automatizadas. Os educadores precisam geralmente de ajuda para entender o funcionamento interno dos modelos de ML. Este artigo examina a IA explicável (XAI) como uma solução para esses problemas na educação. As técnicas de XAI podem fornecer aos educadores e administradores percepções valiosas sobre os algoritmos de ML, facilitando a tomada de decisões mais informadas. Discutimos a diferença entre escolhas algorítmicas transparentes e opacas, demonstrando os benefícios tangíveis da XAI na educação. Os modelos transparentes permitem que os educadores aproveitem seus conhecimentos de forma eficaz, descubram padrões ocultos e melhorem os resultados dos alunos.

Palavras-chave: inteligência artificial; inteligência artificial explicável; aprendizado de máquina; educação.

RESUMEN

La adopción de algoritmos de aprendizaje automático (ML) en la educación es cada vez mayor, con el objetivo de mejorar los procesos de enseñanza, aprendizaje y administración. Esos algoritmos son cruciales para el aprendizaje personalizado, la predicción del rendimiento de los alumnos y el diseño de planes de estudio. Sin embargo, su uso generalizado puede plantear problemas como la parcialidad, la falta de transparencia y la dependencia excesiva de las decisiones automatizadas. A menudo, los educadores necesitan ayuda para comprender el funcionamiento interno de los modelos de ML. Ese artículo examina la IA explicable (XAI) como solución a esos problemas en la educación. Las técnicas de XAI pueden proporcionar a los educadores y administradores información valiosa sobre los algoritmos de ML, facilitando una toma de decisiones más informada. Discutimos la diferencia entre opciones algorítmicas transparentes y opacas y demostramos los beneficios tangibles de la XAI en la educación. Los modelos transparentes permiten a los educadores disfrutar su experiencia de forma eficaz, descubrir patrones ocultos y mejorar los resultados de los estudiantes.

Palabras clave: inteligencia artificial; inteligencia artificial explicable; aprendizaje automático; educación.

INTRODUCTION

In contemporary educational institutions, relying solely on human decision-making can be time-consuming and error-prone. This is particularly true in crucial areas like learning environment design and personalized learning path development, where ensuring effective and efficient learning is paramount amidst growing student populations and diverse learning needs (Zhang; Aslan, 2021). In parallel, educators struggle with heavy workloads, managing large class sizes and diverse student needs, often exceeding their own capacities (Griffin, 2022). This situation can limit the individualized attention students receive, potentially hindering their learning progress. In fact, students might find their current learning experience unsustainable in the long term, potentially leading to reduced engagement and motivation (Beattie; Thiele, 2016).

The growing excitement surrounding generative AI (GenAI) technologies, such as ChatGPT, has positioned them as potential game-changers in how educators can enhance their teaching abilities by providing adaptative teaching strategies and customized suggestions (Chiu, 2023). However, educators must recognize that GenAI does not represent the next step of AI as an "evolution" of machine learning (ML), a subset of AI algorithms and solutions. In fact, it is just another subtype of ML (Nah *et al.*, 2023). A more apt analogy would be to compare GenAI to interactive whiteboards and ML to standard desktop computers. Much like interactive whiteboards have not entirely replaced desktop computers in classrooms but instead found their unique niches, GenAI and ML each have distinct applications and strengths. Many tasks are more efficiently executed in the classroom on a desktop computer. However, specific scenarios exist where an interactive

whiteboard might enhance learning outcomes. This analogy extends to the relationship between GenAI and ML: each has its preferred use cases and areas where it excels.

Regardless, not all educational challenges are best addressed with AI solutions similar to ChatGPT. In other words, ChatGPT and its competitors are not one-size-fits-all solutions in educational applications. While GenAI excels in generating new content instead of predicting, such as creating personalized problem statements, use cases, or educational stories (Chiu, 2023), ML analyzes existing educational data to identify patterns and inform decisions. ML can be used to develop several applications, such as providing personalized learning paths with intelligent tutoring systems, providing immediate and tailored feedback to students, predicting students' grades at the end of their courses based on their current trends and performance, automatic content selection for tests and classes, automation of quality control for learning artifacts for courses, predicting churn rates, predicting student dropout, and automatic recommendation of additional learning resources (Korkmaz; Correia, 2019; Bonifro *et al.*, 2020; Luan; Tsai, 2021). Consequently, ML algorithms are more suitable for a broader range of AI-related applications in education. However, this raises a crucial question: How are these ML algorithms initially assembled in educational contexts?

Predictive ML, similar to experienced educators, utilizes historical data to enhance its decision-making capabilities (Korsakiene *et al.*, 2015). Just as educators gain knowledge through diverse teaching experiences, observing student behavior, and identifying learning patterns over time, ML algorithms learn from past educational data to predict future student performance or success.

Consider, for instance, the development of a student dropout prediction algorithm (Bonifro *et al.*, 2020). An experienced educator might find that certain factors influencing student dropout are consistent across different schools or curricula, supported by various studies within specific educational settings. However, complexity arises: not all schools operate identically or share the same resources or data availability regarding these influential factors.

As another illustrative example, consider the development of an algorithm to provide personalized learning experiences for students to improve their engagement (Nabizadeh *et al.*, 2020). Not all students, courses, and institutions operate in the same

way. The impact of these personalized learning experiences on student engagement might vary significantly between different grade levels, subject areas, and student learning styles (Santally; Senteni, 2013). While personalized learning may greatly influence engagement among students in advanced math courses, factors like access to technology or classroom environment are likely more pertinent to elementary school students (Haverinen-Shaughnessy; Shaughnessy, 2015; Fabian; Topping; Barron, 2016; Bernacki; Walkington, 2018).

Additionally, other factors beyond ML algorithms can influence student engagement, such as the availability of extracurricular activities on student engagement, which can differ drastically between schools, depending on their availability and support system (Buckley; Lee, 2021; Munir; Zaheer, 2021).

Therefore, the same factors (e.g., technology availability, classroom environment, teaching experiences) have different weights in different institutions, classes, and students. Consequently, it is essential to acknowledge that there are better routes than a one-size-fits-all approach to ML algorithms in education. Each institution can achieve better results by having its own ML algorithm tailored to its context, objectives, and goals.

For example, a university interested in developing an algorithm to predict student dropout may have smaller datasets (L'heureux *et al.*, 2017; Zhou *et al.*, 2017) tailored to its different schools. Law students may have different behaviors than IT students, for example. Therefore, training specialized algorithms in different schools can lead to more accurate predictions and interventions tailored to the unique needs of different student groups.

Furthermore, the versatility of ML extends to various educational applications, thus demonstrating its potential to transform traditional teaching practices. A prominent example is the use of ML algorithms in personalized learning (Santally; Senteni, 2013). These algorithms analyze historical data, including student performance, learning styles, and past learning behaviors, to identify the resources and instructional approaches most likely to benefit individual students. More advanced applications aim to predict which learning approaches will lead to significant academic growth for each student. This predictive capability empowers educators to make more informed decisions, enhancing

the effectiveness of personalized learning strategies and improving overall student outcomes.

The second example considers ML algorithms aimed at predicting student performance. These algorithms can predict which students are at risk of falling behind or exceeding expectations based on their current trajectories, considering the analysis of the historical performance of other students, learning patterns, current students' performance, and other demographic data (Yousafzai; Hayat; Afzal, 2020). Additionally, similar algorithms can be used for "what-if" scenario modeling, simulating potential changes in student performance outcomes based on adjustments to factors like learning environment, instructional approaches, support services, and access to resources (Wachter; Mittelstadt; Russell, 2018).

The third example involves ML algorithms designed to minimize student attrition (Beer; Lawson, 2017). Algorithms can rely on data from themes such as work, personal information, academic support, financial data, and institutional support to predict the individual student likelihood of not completing their enrollment for reasons such as withdrawing from a course, failing to attend classes, or canceling their program (Beer; Lawson, 2017). These predictions can guide educators in implementing targeted interventions to foster student engagement and prevent attrition-related issues, which can also impact the institution's reputation (Beer; Lawson, 2017).

The fourth example involves algorithms used to improve student engagement within the classroom. These algorithms analyze data such as student gender, age, interest in the classroom topics, and abilities (Goldberg *et al.*, 2021). Based on this data, the algorithms can predict which students or learning groups are experiencing lower engagement. With this information, educators can adapt their teaching method accordingly, react to disruptions, and improve the effectiveness of their instruction time (Goldberg *et al.*, 2021).

However, a significant challenge arises regardless of the specific strategy chosen for developing a new ML algorithm. While contemporary ML models in education often demonstrate high accuracy compared to traditional methods, their decision-making processes can be challenging to understand and explain. This lack of transparency means they are often considered *black-box models*. The core issue with black-box models is the

inability to explain clearly how an algorithm arrives at a prediction (Samek; Müller, 2019). For instance, consider a student dropout prediction algorithm that produces contrasting predictions for two seemingly similar students. To build trust in the algorithm, educators must understand its predictions' rationale. Unquestioned trust in the algorithm's recommendations is often unacceptable and impractical (Samek; Müller, 2019; Arrieta *et al.*, 2020). Even for legal and ethical reasons (Samek; Müller, 2019), it is vital to clearly explain the rationale behind the predictions of ML algorithms used in education. Additionally, promoting professional development is another compelling reason — teachers could gain valuable insights and discover new patterns or teaching strategies by understanding the logic behind these algorithms (Samek; Müller, 2019; Arrieta *et al.*, 2020). This knowledge sharing can be particularly beneficial for less experienced educators on the team.

Therefore, how can we promote transparent decision-making in education with AI tools while maintaining high levels of automation and accuracy? The answer lies in a recent advancement in AI research: Explainable AI (XAI) (Samek; Müller, 2019; Gunning *et al.*, 2019; Gunning; Aha, 2019; Arrieta *et al.*, 2020). Introduced at the end of the last decade, XAI opens up the inner workings of "black-box" ML algorithms to educators and other professionals. Utilizing XAI techniques with ML models in educational contexts can foster trust, ensure accountability, discover new pedagogical insights, and fulfill legal requirements depending on location (Samek *et al.*, 2019; Pessach; Shmueli, 2022). Therefore, this paper demonstrates the benefits of understanding XAI methods and how educators and administrators can apply these techniques to ML algorithms in their schools and systems.

EXPLAINING THE DATA SCIENCE PIPELINE FOR EDUCATORS

To better understand XAI's relevance, it is essential to understand how ML algorithms are built in the first place. Although data science teams use several project methodologies to build these ML algorithms for education solutions, the most common is CRISP-DM (Schröer; Kruse; Gómez, 2021). Its lifecycle is shown in Figure 1.

When considering the application of AI in education, educators and educational leaders collaborate to determine if AI can effectively address a specific challenge or opportunity (the Business Understanding step) (Schröer; Kruse; Gómez, 2021). This process

often starts by identifying a specific issue or area for improvement, such as enhancing student engagement or optimizing personalized learning pathways. It is essential to note that only some educational challenges necessitate complex AI solutions. While there may be concerns about over-reliance on AI, more straightforward solutions might be more appropriate and cost-effective in many cases. These could involve adjusting existing teaching practices, developing informative dashboards to visualize student progress, or even creating standard tools or scripts to automate routine tasks.

In this step, it is also helpful to set the project's objectives closer to reality and in a way that allows data to be retrieved and results to be measured quickly. Now, if there is a need to *predict* something, it is possible to move forward to understand whether there is available data within the education institution to move forward. Sometimes, there needs to be more data available to create a good ML algorithm, or there are considerable issues with data quality (such as unavailable or wrong data). If there are deal-breaking issues with the data, the objectives set in the Business Understanding step should be changed accordingly before moving forward.

Every predictive AI algorithm requires a representative dataset that should be prepared by a human beforehand. This tabular dataset can be similar to an Excel spreadsheet composed of rows and columns (Shwartz-Ziv; Armon, 2022). Each row can represent a data point (in our illustrative example, a student), and each column represents a data attribute such as previous grades, attendance patterns, participation in extracurricular activities, or demographic information (Shwartz-Ziv; Armon, 2022). The ML algorithm will be trained over that dataset (Russell; Norvig, 2021); therefore, including as many examples as possible is essential while keeping the highest quality standards. Values should be standardized, anomalies should be analyzed and treated if needed, and typing errors must be fixed. In this Data Preparation stage, the responsibility should be shared between educators and data professionals (Viaene, 2013): educators comprehend whether that data correctly represents business rules and trends, and data professionals understand statistics techniques and visual tools to measure the quality of that information quantitatively. Any filters, business rules, and data standardization procedures in this first analysis should also be carried over during the prediction of the ML algorithms.

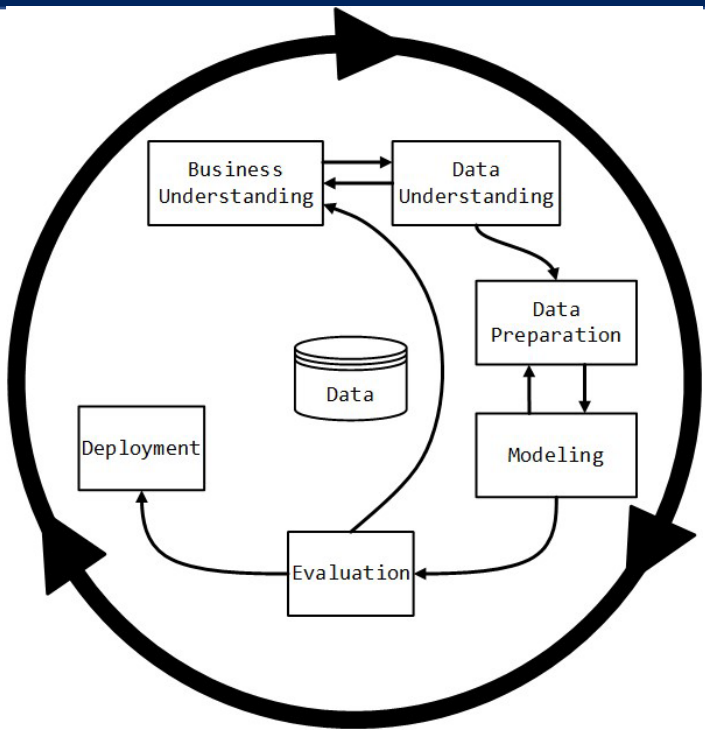

Figure 1: CRISP-DM lifecycle. It is a continuous process where the AI model used in education applications is regularly updated and improved as it receives new data, ensuring that the insights remain relevant and accurate. Source: the author.

Additionally, this dataset enables an AI to understand the patterns exclusive to that dataset and, by extension, to that educational institution: Is there a link between a student's previous grades in a given subject and their likelihood of seeking additional support? Should we analyze grades concerning learning styles and preferences? Are attendance patterns over the last 30 days (or perhaps 60 days) a significant predictor of academic struggle? Should factors such as socioeconomic background or participation in extracurricular activities be considered, or would they introduce biases and create unfair predictions? How can we ensure the dataset is diverse enough to avoid accidentally training the algorithm with only limited examples? Each decision taken for any of these examples implies training a different version of an ML algorithm, which, in turn, directly impacts its performance and fairness. Therefore, the complexity of having an ML algorithm starts before its existence. An experienced data scientist working alongside a knowledgeable education professional can discover valuable insights by analyzing this dataset and reducing the risk of spurious correlations or selecting irrelevant data attributes (Calude; Longo, 2017).

Now that a good representative dataset has been produced and prepared, the next step is to train the ML algorithm appropriately in the Modeling stage (Schröer; Kruse; Gómez, 2021). This step means choosing the best ML architecture to discover hidden patterns in the data and predict new cases. In the same way that several schools of thought exist in fields of study such as Pedagogy and Psychology for education strategies, where each strategy has its advantages and disadvantages, the same happens in AI (Domingos, 2015). There is no single ML algorithm, but instead dozens of different families (Domingos, 2015). In that sense, it is expected to hear from data scientists terms such as *gradient boosting, random forest, neural networks*, and *support vector machines* to refer to these different strategies. It is also the data scientist's responsibility to understand the best strategy for different education problems, considering characteristics such as data size, speed, cost, and generalization. In scenarios where a data scientist is unavailable, it is also possible to resort to enterprise products that automate part of the task for education professionals: working as "citizen data scientists", these professionals can submit a curated dataset to an automated ML product that will attempt to find the best possible ML algorithm (Mullarkey *et al.*, 2019).

However, who decides how many algorithms should be prepared (i.e., *trained*)? Creating AI for education in the shape of new, tailor-made ML algorithms is the shared task of data professionals such as data analysts, data engineers, analytics engineers, data scientists, and machine learning engineers (Patil; Bhavsar, 2021; Schröer; Kruse; Gómez, 2021). These separate technical roles require a particular skill set to prepare and deploy robust ML algorithms, keeping in mind business particularities and advanced statistical concepts. Sometimes, the algorithm trained does not perform as expected; in that case, the data professionals walk back one step and refine the dataset before training a new algorithm again (Schröer; Kruse; Gómez, 2021).

As soon as a good algorithm has been found and evaluated against the project objectives set in the Business Understanding step, the next phase is to deploy that model on a secure IT server (Schröer; Kruse; Gómez, 2021). It is good to take an ML algorithm that is running locally on the computer of a data scientist or an educator and put it on a server (Singh, 2021). After all, that algorithm should remain accessible to all stakeholders and be open for emergency changes if required, regardless of whether someone is unavailable or

on vacation. This step is usually the responsibility of ML engineers, professionals with deep computing and data science knowledge, to keep the ML algorithms running without significant issues for educators and students affected by them.

THE PROBLEM OF PURSUING "GREAT" AI

Even though these algorithms can exhibit high accuracy, enhance educational performance, and offer rapid predictions based on extensive historical data, potential challenges arise when developing new AI applications in education.

One critical concern is the potential for bias (Arrieta *et al.*, 2020; Pessach; Shmueli, 2022). Suppose the historical data used to train the algorithm reflects inherent biases (such as inequitable grading practices in the past or an overrepresentation of particular student demographics) (Friedler *et al.*, 2019). In that case, the algorithm may unintentionally learn and reinforce these biases.

This bias can result in unfair and harmful outcomes (Arrieta *et al.*, 2020; Pessach; Shmueli, 2022), such as underpredicting students' potential from specific backgrounds or systematically overlooking students requiring additional support. Even if an ML algorithm demonstrates low statistical error, it does not guarantee fairness or alignment with the school's or district's equity-focused policies.

Another danger lies in the overreliance on these algorithms. While AI can provide valuable insights, it only captures part of the complexity of human behavior and workplace dynamics. Consuming the ML predictions without understanding how that algorithm works can lead to spurious correlations or "Clever Hans" behaviors (Samek; Müller, 2019). Clever Hans was a famous horse in the early 20th century, known for his apparent ability to perform arithmetic and other intellectual tasks. However, it was later discovered that Hans responded to subtle, unintentional cues from his handler instead of understanding the tasks (Samek; Müller, 2019). Considering AI, an ML model could be trained to predict student disengagement or dropout risk. This model might analyze factors like academic performance, participation in class, attendance patterns, and access to resources. However, similar to the Clever Hans effect, the model could inadvertently capture correlations with irrelevant data points, such as the student's surname or the preferred teaching subject. This situation could lead to a false sense of predictive accuracy, with the

model's forecasts based on coincidental, non-causal associations rather than genuine indicators of disengagement risk. If there is an over-reliance on these algorithms or a lack of interest in understanding how they work, such undesired situations might arise in critical applications.

HOW TO OPEN THE SECRETS OF A SYNTHETIC BRAIN?

One way to avoid "Clever Hans" behaviors is by reducing the opaqueness of ML algorithms (Samek *et al.*, 2019). At the end of the 20th century, more straightforward techniques such as linear regression, logistic regression, support vector machines, and single decision trees were thoroughly used in enterprise ML applications due to performance constraints and because they were intrinsically interpretable to humans (Arrieta *et al.*, 2020). At first glance, the decision tree structure with its nodes and branches is easy to read by humans since its structure does not require a solid technical background to be analyzed (Arrieta *et al.*, 2020). The same applies to a logistic regression with its coefficients and intercept, which only requires basic knowledge of Mathematics (Arrieta *et al.*, 2020).

However, the breakthroughs in the 2010s enabled rapid dissemination of more complex ML algorithms that unleashed accuracy levels unseen in these intrinsically interpretable techniques at the cost of being too complex to be readily understood by humans. This is the case with techniques such as random forests, gradient-boosting machines, and deep neural networks. In these algorithms, the underlying rationale is opaque — hence the "black-box" alias. XAI technical literature often brings up the trade-off between interpretability and accuracy, as illustrated in Figure 2 (Adadi; Berrada, 2018; Gunning *et al.*, 2019; Gunning; Aha, 2019; Arrieta *et al.*, 2020). ML models with higher accuracy are located on the left side. Unfortunately, these models also have the lowest interpretability. The potential accuracy reduces as the line goes to the right, and the interpretability increases.

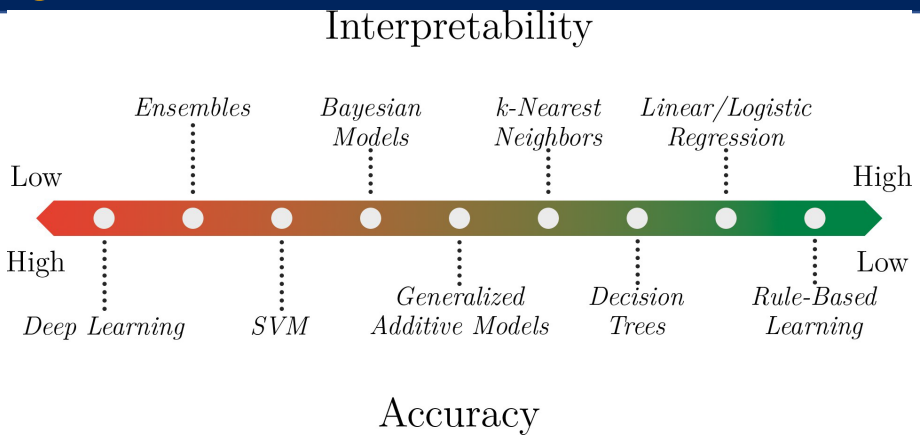

Figure 2: Continuum between interpretability and accuracy. Source: the author.

The lack of transparency in these newer ML models can lead to undetectable "Clever Hans" problems and trust issues among educators and other professionals, who may feel that automated decisions regarding their work are being made arbitrarily or unfairly (Samek *et al.*, 2019). If the professionals cannot understand how the algorithms came to a given decision, they can feel dissatisfied and perceive injustice (Binns, 2020). Moreover, black-box algorithms make it difficult to identify and correct biases (Pessach; Shmueli, 2022). With clear insight into how decisions are made, discriminatory patterns might be noticed and addressed, avoiding the perpetuation of classroom inequality. Finally, having these black-box algorithms can challenge ensuring compliance with legal and ethical standards. With transparency, it becomes easier to audit and validate the effectiveness and accuracy of AI tools (Arrieta *et al.*, 2020). This lack of accountability can result in the continued use of flawed systems, leading to poor decision-making in critical education functions.

Although the easy route is to avoid these complex techniques, not relying on these more advanced algorithms implies losing the competitive advantage bought by newer AI (Sagi; Rokach, 2018; Arrieta *et al.*, 2020). This is where XAI comes to the rescue. After training an ML algorithm, data scientists can append a second algorithm that summarizes and explains the black-box functionality to humans.

THE DIFFERENT TYPES OF XAI ALGORITHMS

So, how do we leverage XAI in complex, black-box models in educational contexts? If there are different XAI techniques, how do we choose the best one for any ML model

currently used in such scenarios? In short, the decision depends on the complexity and frequency of the answers required and the target audience.

Big picture or deep dive?

The first dimension to be analyzed is the level of interpretability. Consider a complex ML model built by data scientists. In several meetings and internal presentations, they reported that the model has high accuracy. When presented with a spreadsheet comparing the results, the predictions provided by the algorithm are sound. However, they cannot show a simple diagram showcasing, in broad terms, how the algorithm *works*.

For example, consider the diagram presented in Figure 3. It represents a deep neural network where each column represents a layer and each circle represents a node. A human cannot explain how it works or provide predictions just by looking at it. Each node has its own mathematical formula, and evaluating it during a meeting is often impossible. Consider the decision tree shown in Figure 4 as a comparison. It is easier to understand and does not require technical explanations of how it works. It can be presented in meetings with educators and other stakeholders, and humans can learn from it. This is an example of an XAI technique aimed at **global interpretability**. This technique is called *surrogate model generation* but can also be found with other names such as model distillation and model simplification (Samek; Müller, 2019; Arrieta *et al.*, 2020; Molnar, 2023).

In the same way that an abstract provides a brief understanding of the scope of an article, a surrogate provides a summarized explanation of a complex black-box ML model, such as a deep neural network. Although these techniques help understand the general patterns in data, they hide the nuances and particularities of the model when evaluating particular cases, since they favor the generalization and simplification of the model to be understood by humans.

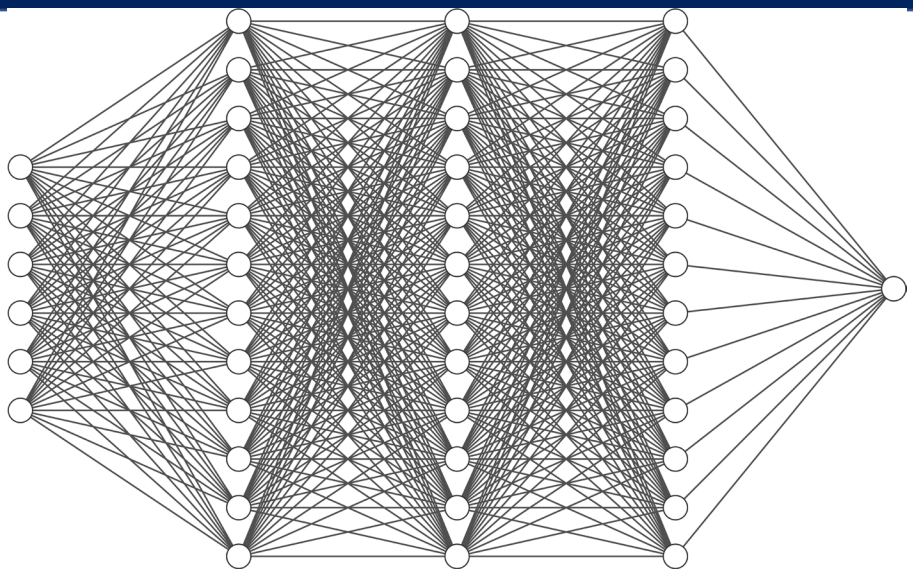

Figure 3: Example of a diagram of a deep neural network. Source: the author.

In contrast, **local interpretability** deals with a single individual or case scenario separately (Adadi; Berrada, 2018). Therefore, local interpretability techniques attempt to explain *why* the output of a trained ML model came to a specific prediction for a specific person or case.

Local interpretability algorithms have more research and techniques available than global interpretability algorithms. A popular example is SHAP, which is based on game theory (Lundberg; Lee, 2017). This technique calculates values to determine the effects of removing a feature from a model using sampling approximations. These sampling approximations allow the calculation of these values without modifying already-deployed models within educational institutions. An example is shown in Figure 5 for a public database. Similar to a tug-of-war, it is possible to see each value's contribution towards a given prediction. Some values push toward a positive outcome, while others push towards a negative outcome. Each attribute contributes with different percentages towards a final result.

What do you want to see?

After considering whether local or global interpretability is more appropriate, the next step is to select the type of explainability that best aligns with the specific educational use case. In fact, there are six types of explanations that can be used for black-box ML

models: feature relevance explanations, local explanations, explanations by examples, explanations by simplification, visual explanations, and textual explanations (Adadi; Berrada, 2018; Arrieta *et al.*, 2020).

Feature relevance explanations will measure a feature's relevance, influence, or relative ranking over the predicted output of a black-box ML model (Adadi; Berrada, 2018). Therefore, these techniques can explain the main influential features of the whole model and the influential features per prediction individually. Data scientists can choose these techniques when education professionals need to understand and evaluate the most critical attributes of a black-box ML model. Sensitivity Analysis (Ancona *et al.*, 2019) and SHAP, as seen previously, are some of the most popular feature relevance explanation methods.

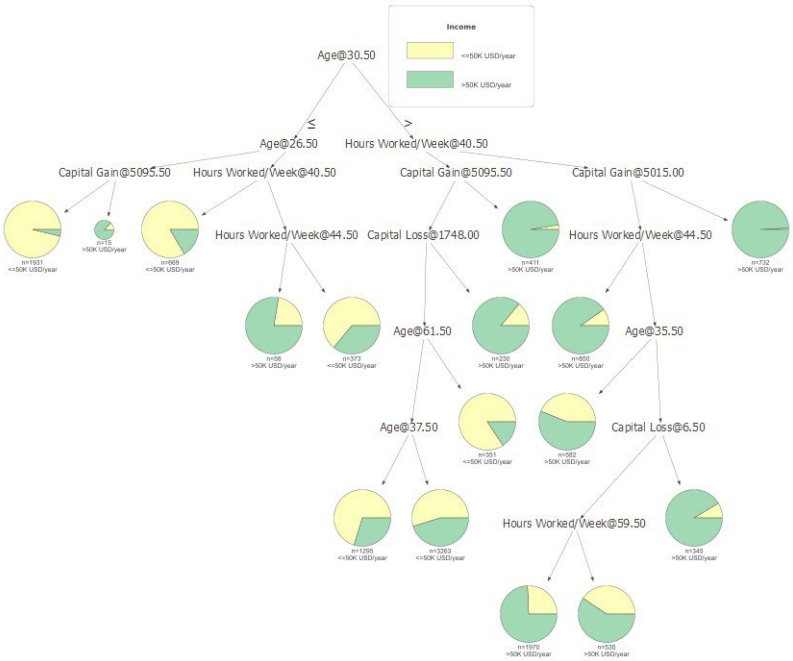

Figure 4: Example of a decision tree with a small number of nodes. Source: the author.

On the other hand, **local explanations** are methods specialized in explaining a small part of a larger ML model. Usually, these methods explain single predictions or a smaller group of predictions (Adadi; Berrada, 2018; Arrieta *et al.*, 2020). One of the most popular local explanation techniques is LIME (Ribeiro; Singh; Guestrin, 2016). Its goal is to provide locally faithful explanations instead of trying to explain or generalize the model globally.

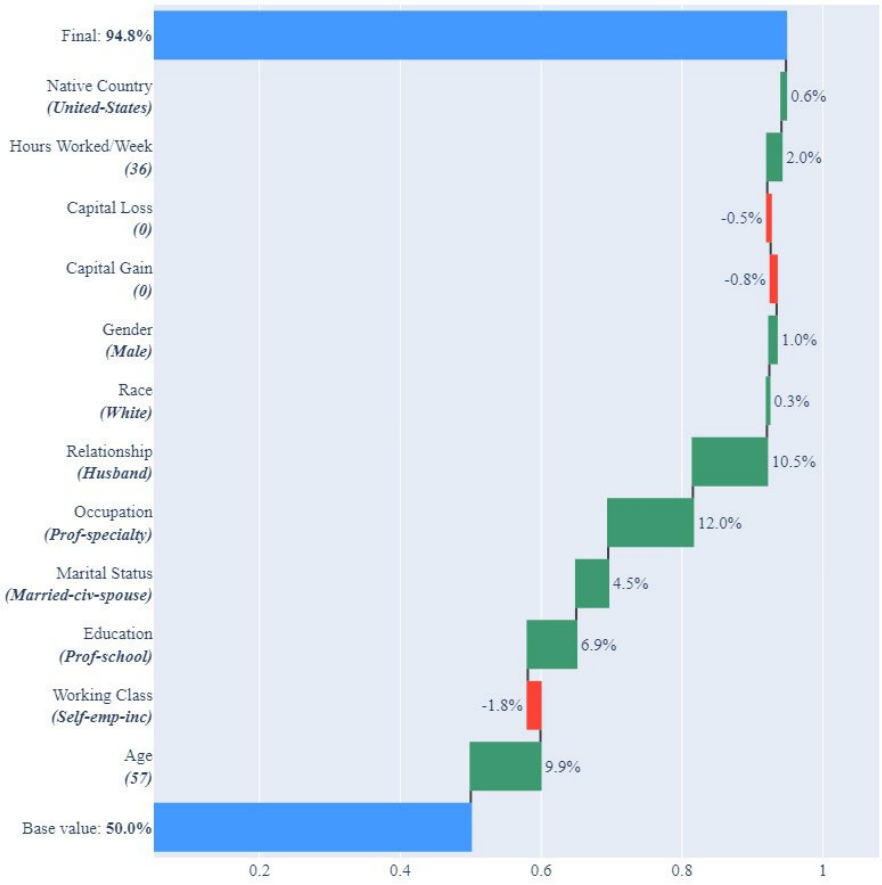

Figure 5: Example of the SHAP output for a trained black-box model. Source: the author.

The third type of explanation is an **explanation by example**. These explanations focus on displaying other examples close to the inputs informed by a trained ML model that, in turn, enable humans to understand the decisions taken by the ML model based on its inner relationships and correlations. Going back to the illustrative example of an ML algorithm built to predict student attrition, this technique will show the prediction (true or false) and some similar students in the past that led to the same outcome. With these examples, educators can understand how grounded the model is. Similarly, another type of explanation is *prototype generation*. Prototypes represent a learned concept humans can use to understand if the black-box ML model learned concepts correctly, thus avoiding the "Clever Hans" scenarios. For instance, what caused the algorithm to identify a student needing additional academic support? Was it solely due to a single low-test score, potentially a misunderstanding of instructions, or other factors that could be misleading?

Or is it a combination of factors, such as consistently low homework completion rates and limited engagement in classroom discussions?

Another example of explanations by example is the *counterfactual generation* (Wachter; Mittelstadt; Russell, 2018). Counterfactuals are modified individuals based on a single, original example. "What-if?" analyses fall into this category. Techniques such as DiCE (Mothilal; Sharma; Tan, 2020) try to identify the minimum changes required for a given individual to change its outcome. Let us explore how this technique can provide better-informed decisions, considering the illustrative example of a student attrition prediction algorithm once again. In this example, the ML algorithm identified a student at potential risk of attrition. Counterfactual generation techniques can then provide suggestions for minor adjustments that educators and other professionals could make to change that prediction. Therefore, the algorithm can provide suggestions such as additional support in the classroom, other extracurricular activities, or other proposals within the reach of the educational institution.

The fourth type is an **explanation by simplification**. These techniques aim to reduce the complexity of a black-box model by simplifying it by changing its architecture or modifying its parameters. The surrogate models explained earlier are an example of this (Molnar, 2023).

Finally, there are **visual explanations** and **text explanations**. Both techniques attempt to describe to humans what the algorithm "understands", using visual methods or textual descriptions. They can be combined with other techniques to show educators how an ML algorithm works. For example, SHAP is both a visual and a local explanation method (Lundberg; Lee, 2017; Arrieta *et al.*, 2020). Methods that rely on these resources are instrumental in presentations to educators and other decision-makers.

CONCLUSION

This paper highlights the value of responsibly utilizing XAI techniques to interpret and evaluate complex ML models in educational contexts. Schools, universities, and other educational institutions can leverage sophisticated ML algorithms to optimize decision-making. While these algorithms sometimes lack transparency, XAI methods can demystify their inner workings for non-data scientists, such as educators and other professionals.

Therefore, a crucial takeaway is that educators should only feel compelled to limit themselves to simpler, more transparent ML approaches if these can adequately address their objectives.

The second point is the flexibility of XAI techniques, allowing them to be tailored to different audiences within the educational community. Model surrogates can be invaluable for presentations to stakeholders, such as school board members or coordinators, providing a simplified explanation of how the model generally functions. Local explanations can aid teachers and administrators in working closely with data scientists to delve into specific cases, identifying how different factors contribute to individual student outcomes. Lastly, counterfactuals allow guidance counselors and support staff to explore potential strategies and resource allocation scenarios to improve outcomes for students needing additional support.

The third point emphasizes the importance of collaboration between educators and data scientists. Building an effective ML algorithm for educational applications should not be solely the responsibility of data scientists. Educators' expertise in Pedagogy and their understanding of the specific learning context are crucial for ensuring the algorithm addresses relevant challenges and avoids generating misleading or harmful insights. Conversely, the technical knowledge of data scientists is vital for building and fine-tuning the algorithm, preventing technical issues like data leakage and overfitting.

Integrating XAI into complex ML systems in education enhances transparency while maintaining high levels of accuracy. These techniques promote a continuous learning cycle between educators and AI, revealing previously hidden patterns and accelerating teachers' and administrators' understanding of factors influencing student outcomes. Additionally, XAI promotes the development of equitable algorithms, improving auditability and fostering trust in AI, which will become increasingly integral to educational systems in the years to come.

REFERENCES

BOOK:

DOMINGOS, P. **The Master Algorithm**: How the Quest for the Ultimate Learning Machine Will Remake Our World. London: Penguin Books Limited, 2015.

MOLNAR, C. **Interpretable Machine Learning**: A guide for making black box models explainable. Victoria: Leanpub, 2023.

RUSSELL, S. J.; NORVIG, P. **Artificial Intelligence**: A Modern Approach. Hoboken: Pearson, 2021. (Pearson series in artificial intelligence).

SAMEK, W. *et al.* **Explainable AI:** Interpreting, Explaining and Visualizing Deep Learning. Cham: Springer International Publishing, 2019. v. 11700. DOI: https://doi.org/10.1007/978-3-030-28954-6. Available at: https://link.springer.com/content/pdf/10.1007/978-3-030-28954-6.pdf. Accessed on: 28 Feb. 2024.

BOOK CHAPTER:

ANCONA, M. *et al.* Gradient-based attribution methods. *In:* ANCONA, M. *et al.* **Explainable AI:** Interpreting, Explaining and Visualizing Deep Learning. Cham: Springer International Publishing, 2019. p. 169-191. DOI: https://doi.org/10.1007/978-3-030-28954-6_9. Available at: https://link.springer.com/content/pdf/10.1007/978-3-030-28954-6.pdf. Accessed on: 28 Feb. 2024.

BONIFRO, F. *et al.* Student Dropout Prediction. *In:* BITTENCOURT, I. I. *et al.* (Ed.). **Artificial Intelligence in Education**. Cham: Springer International Publishing, 2020. p. 129-140. DOI: https://doi.org/10.1007/978-3-030-52237-7_11. Available at: https://link.springer.com/content/pdf/10.1007/978-3-030-52237-7.pdf. Accessed on: 28 Feb. 2024.

MULLARKEY, M. T. *et al.* Citizen data scientist: A design science research method for the conduct of data science projects. *In:* TULU, B.; DJAMASBI, S.; LEROY, G. (Ed.). **Extending the Boundaries of Design Science Theory and Practice**. Cham: Springer International Publishing, 2019. p. 191-205. DOI: https://doi.org/10.1007/978-3-030-19504-5_13.

PATIL, T.; BHAVSAR, A. K. Data science team roles and need of data science: A review of different cases. *In:* KOTECHA, K. *et al.* (Ed.). **Data Science and Intelligent Applications**. Singapore: Springer Singapore, 2021. p. 13-22. DOI: https://doi.org/10.1007/978-981-15-4474-3_2. Available at: https://link.springer.com/content/pdf/10.1007/978-981-15-4474-3.pdf. Accessed on: 28 Feb. 2024.

SAMEK, W.; MÜLLER, K. R. Towards explainable artificial intelligence. *In:* SAMEK, W.; MÜLLER, K. R. (Orgs.). **Explainable AI**: Interpreting, Explaining and Visualizing Deep Learning. Cham: Springer International Publishing, 2019. p. 5-22. DOI: https://doi.org/10.1007/978-3-030-28954-6_1. Available at: https://link.springer.com/content/pdf/10.1007/978-3-030-28954-6.pdf. Accessed on: 28 Feb. 2024.

SINGH, P. Model deployment and challenges. *In*: SINGH, P. (Org.). **Machine Learning Models to Production**: With Flask, Streamlit, Docker, and Kubernetes on Google Cloud Platform. Berkeley: Apress, 2021. p. 55-66. DOI: https://doi.org/10.1007/978-1-4842-6546-8_2.

ARTICLE IN A PERIODICAL (MAGAZINE OR JOURNAL):

ADADI, A.; BERRADA, M. Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). **IEEE Access**, v. 6, p. 52138-52160, 2018. DOI: https://doi.org/10.1109/ACCESS.2018.2870052. Available at: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8466590. Accessed on: 28 Feb. 2024.

ARRIETA, A. B. *et al.* Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. **Information Fusion**, Elsevier, v. 58, p. 82-115, June 2020. DOI: https://doi.org/10.1016/j.inffus.2019.12.012.

BEATTIE, I. R.; THIELE, M. Connecting in class? College class size and inequality in academic social capital. **The Journal of Higher Education**, Routledge, v. 87, n. 3, p. 332-362, 2016. DOI: https://doi.org/10.1080/00221546.2016.11777405.

BEER, C.; LAWSON, C. The problem of student attrition in higher education: An alternative perspective. **Journal of Further and Higher Education**, Routledge, v. 41, n. 6, p. 773-784, 2017. DOI: https://doi.org/10.1080/0309877X.2016.1177171.

BERNACKI, M. L.; WALKINGTON, C. The role of situational interest in personalized learning. **Journal of Educational Psychology**, v. 110, n. 6, p. 864-881, 2018. DOI: https://psycnet.apa.org/doi/10.1037/edu0000250.

BUCKLEY, P.; LEE, P. The impact of extra-curricular activity on the student experience. **Active Learning in Higher Education**, v. 22, n. 1, p. 37-48, 2021. DOI: https://doi.org/10.1177/1469787418808988.

CALUDE, C. S.; LONGO, G. The Deluge of Spurious Correlations in Big Data. **Foundations of Science**, v. 22, n. 3, p. 595-612, 2017. DOI: https://doi.org/10.1007/s10699-016-9489-4.

CHIU, T. K. F. The Impact of Generative AI (GenAI) on Practices, Policies and Research Direction in Education: A Case of ChatGPT and Midjourney. **Interactive Learning Environments**, Routledge, v. 0, n. 0, p. 1-17, 2023. DOI: https://doi.org/10.1080/10494820.2023.2253861. Available at: https://www.tandfonline.com/doi/full/10.1080/10494820.2023.2253861#d1e328. Accessed on: 28 Feb. 2024.

FABIAN, K.; TOPPING, K. J.; BARRON, I. G. Mobile technology and mathematics: effects on students' attitudes, engagement, and achievement. **Journal of Computers in Education**, v. 3, n. 1, p. 77-104, 2016. DOI: https://doi.org/10.1007/s40692-015-0048-8.

GOLDBERG, P. *et al.* Attentive or not? toward a machine learning approach to assessing students' visible engagement in classroom instruction. **Educational Psychology Review**, v. 33, n. 1, p. 27-49, 2021. DOI: https://doi.org/10.1007/s10648-019-09514-z. Available at: https://link.springer.com/content/pdf/10.1007/s10648-019-09514-z.pdf. Accessed on: 28 Feb. 2024.

GRIFFIN, G. The 'work-work balance' in higher education: between over-work, falling short and the pleasures of multiplicity. **Studies in Higher Education**, Routledge, v. 47, n. 11, p. 2190-2203, 2022. DOI: https://doi.org/10.1080/03075079.2021.2020750. Available at: https://www.tandfonline.com/doi/full/10.1080/03075079.2021.2020750#d1e109. Accessed on: 28 Feb. 2024.

GUNNING, D.; AHA, D. DARPA's explainable artificial intelligence (XAI) program. **AI Magazine**, v. 40, n. 2, p. 44-58, 2019. DOI: https://doi.org/10.1609/aimag.v40i2.2850. Available at: https://onlinelibrary.wiley.com/doi/epdf/10.1609/aimag.v40i2.2850. Accessed on: 28 Feb. 2024.

GUNNING, D. *et al.* XAI - Explainable Artificial Intelligence. **Science Robotics**, v. 4, n. 37, 2019. DOI: https://doi.org/10.1126/scirobotics.aay7120.

HAVERINEN-SHAUGHNESSY, U.; SHAUGHNESSY, R. J. Effects of classroom ventilation rate and temperature on students' test scores. **PLOS ONE**, Public Library of Science, v. 10, n. 8, p. 1-14, 08 2015. DOI: https://doi.org/10.1371/journal.pone.0136165. Available at: https://journals.plos.org/plosone/article/file?id=10.1371/journal.pone.0136165&type=printable. Accessed on: 28 Feb. 2024.

KORKMAZ, C.; CORREIA, A. P. A review of research on machine learning in educational technology. **Educational Media International**, Routledge, v. 56, n. 3, p. 250-267, 2019. DOI: https://doi.org/10.1080/09523987.2019.1669875.

KORSAKIENE, R. *et al.* Factors driving turnover and retention of information technology professionals. **Journal of Business Economics and Management,** Taylor & Francis, v. 16, n. 1, p. 1-17, 2015. DOI: https://doi.org/10.3846/16111699.2015.984492. Available at: https://journals.vilniustech.lt/index.php/JBEM/article/view/2698/2209. Accessed on: 28 Feb. 2024.

LUAN, H.; TSAI, C. C. A review of using machine learning approaches for precision education. **Educational Technology & Society**, International Forum of Educational Technology & Society, v. 24, n. 1, p. 250–266, 2021. Available at: https://www.jstor.org/stable/26977871. Accessed on: 28 Feb. 2024.

L'HEUREUX, A. *et al.* Machine learning with big data: Challenges and approaches. **IEEE Access**, v. 5, p. 7776-7797, 2017. DOI: https://doi.org/10.1109/ACCESS.2017.2696365. Available at: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7906512. Accessed on: 01 Mar. 2024.

MUNIR, S.; ZAHEER, M. The role of extra-curricular activities in increasing student engagement. **Asian Association of Open Universities Journal**, v. 16, n. 3, p. 241-254, 2021. DOI: https://doi.org/10.1108/AAOUJ-08-2021-0080. Available at: https://www.emerald.com/insight/content/doi/10.1108/AAOUJ-08-2021-0080/full/pdf?title=the-role-of-extra-curricular-activities-in-increasing-student-engagement. Accessed on: 01 Mar. 2024.

NABIZADEH, A. H. *et al.* Learning path personalization and recommendation methods: A survey of the state-of-the-art. **Expert Systems with Applications**, v. 159, p. 113596, 2020. DOI: https://doi.org/10.1016/j.eswa.2020.113596.

NAH, F. F. *et al.* Generative AI and ChatGPT: Applications, challenges, and AI-human collaboration. **Journal of Information Technology Case and Application Research**, Routledge, v. 25, n. 3, p. 277-304, 2023. DOI: https://doi.org/10.1080/15228053.2023.2233814. Available at: https://www.tandfonline.com/doi/epdf/10.1080/15228053.2023.2233814?needAccess=true. Accessed on: 01 Mar. 2024.

PESSACH, D.; SHMUELI, E. A review on fairness in machine learning. **ACM Computing Surveys**, Association for Computing Machinery, New York, NY, USA, v. 55, n. 3, feb. 2022. DOI: https://doi.org/10.1145/3494672. Available at: https://dl.acm.org/doi/pdf/10.1145/3494672. Accessed on: 01 Mar. 2024.

SAGI, O.; ROKACH, L. Ensemble learning: A survey. **Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery**, Wiley Online Library, v. 8, n. 4, p. e1249, 2018. DOI: https://doi.org/10.1002/widm.1249.

SANTALLY, M. I.; SENTENI, A. Effectiveness of personalised learning paths on students learning experiences in an e-learning environment. **European Journal of Open, Distance and E-learning**, ERIC, v. 16, n. 1, p. 36-52, 2013. Available at: https://eric.ed.gov/?id=EJ1017459. Accessed on: 01 Mar. 2024.

SCHRÖER, C.; KRUSE, F.; GÓMEZ, J. M. A systematic literature review on applying CRISP-DM process model. **Procedia Computer Science**, v. 181, p. 526-534, 2021. DOI: https://doi.org/10.1016/j.procs.2021.01.199. Available at: https://www.sciencedirect.com/science/article/pii/S1877050921002416/pdf?md5=34d7ce6ba598594c11c16770ac53a4e8&pid=1-s2.0-S1877050921002416-main.pdf. Accessed on: 01 Mar. 2024.

SHWARTZ-ZIV, R.; ARMON, A. Tabular data: Deep learning is not all you need. **Information Fusion**, v. 81, p. 84-90, 2022. DOI: https://doi.org/10.1016/j.inffus.2021.11.011.

VIAENE, S. Data scientists aren't domain experts. **IT Professional**, v. 15, n. 6, p. 12-17, 2013. DOI: https://doi.org/10.1109/MITP.2013.93.

WACHTER, S.; MITTELSTADT, B.; RUSSELL, C. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. **Harvard Journal of Law &**

Technology, v. 31, n. 2, p. 841, 2018. DOI: https://doi.org/10.2139/ssrn.3063289. Available at: https://jolt.law.harvard.edu/assets/articlePDFs/v31/Counterfactual-Explanations-without-Opening-the-Black-Box-Sandra-Wachter-et-al.pdf. Accessed on: 01 Mar. 2024.

YOUSAFZAI, B. K.; HAYAT, M.; AFZAL, S. Application of machine learning and data mining in predicting the performance of intermediate and secondary education level student. **Education and Information Technologies**, v. 25, n. 6, p. 4677-4697, 2020. DOI: https://doi.org/10.1007/s10639-020-10189-1. Available at: https://link.springer.com/content/pdf/10.1007/s10639-020-10189-1.pdf. Accessed on: 01 Mar. 2024.

ZHANG, K.; ASLAN, A. B. AI technologies for education: Recent research & future directions. **Computers and Education: Artificial Intelligence**, v. 2, p. 100025, 2021. DOI: https://doi.org/10.1016/j.caeai.2021.100025. Available at: https://www.sciencedirect.com/science/article/pii/S2666920X21000199/pdfft?md5=6f79cd1edcc954755ba3b4feec270613&pid=1-s2.0-S2666920X21000199-main.pdf. Accessed on: 01 Mar. 2024.

ZHOU, L. *et al.* Machine learning on big data: Opportunities and challenges. **Neurocomputing**, v. 237, p. 350-361, May 2017. DOI: https://doi.org/10.1016/j.neucom.2017.01.026.

PUBLICATION OF PROCEEDINGS OF SCIENTIFIC EVENTS

BINNS, R. On the apparent conflict between individual and group fairness. *In:* 2020 CONFERENCE ON FAIRNESS, ACCOUNTABILITY, AND TRANSPARENCY, 20., January 27-30, 2020, Barcelona. **FAT*'20: Proceedings** [...]. Association for Computing Machinery, 2020. p. 514-524. ISBN 9781450369367. DOI: https://doi.org/10.1145/3351095.3372864. Available at: https://dl.acm.org/doi/pdf/10.1145/3351095.3372864. Accessed on: 01 Mar. 2024.

FRIEDLER, S. A. *et al.* A comparative study of fairness-enhancing interventions in machine learning. *In:* CONFERENCE ON FAIRNESS, ACCOUNTABILITY, AND TRANSPARENCY, 19., January 29-31, 2019, Atlanta. **FAT*'19: Proceedings** [...]. New York: Association for Computing Machinery, 2019. p. 329-338. DOI: https://doi.org/10.1145/3287560.3287589. Available at: https://dl.acm.org/doi/pdf/10.1145/3287560.3287589. Accessed on: 01 Mar. 2024.

LUNDBERG, S. M.; LEE, S. A unified approach to interpreting model predictions. *In:* INTERNATIONAL CONFERENCE ON NEURAL INFORMATION PROCESSING SYSTEMS, 31., December 4-9, 2017, Long Beach. **NIPS'17: Proceedings** [...]. Long Beach: Association for Computing Machinery, 2017. p. 4768-4777. DOI: https://doi.org/10.5555/3295222.3295230. Available at: https://dl.acm.org/doi/pdf/10.5555/3295222.3295230. Accessed on: 01 Mar. 2024.

MOTHILAL, R. K.; SHARMA, A.; TAN, C. Explaining machine learning classifiers through diverse counterfactual explanations. *In:* 2020 CONFERENCE ON FAIRNESS,

ACCOUNTABILITY, AND TRANSPARENCY, 20., January 27-30, 2020, Barcelona. **FAT*' 20: Proceedings** [...]. Barcelona: Association for Computing Machinery, 2020. p. 607-617. DOI: https://doi.org/10.1145/3351095.3372850. Available at: https://dl.acm.org/doi/pdf/10.1145/3351095.3372850. Accessed on: 01 Mar. 2024.

RIBEIRO, M. T.; SINGH, S.; GUESTRIN, C. "Why should I trust you?": Explaining the predictions of any classifier. *In:* ACM SIGKDD INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING, 22., June 12-17, 2016, San Diego. **Proceedings** [...]. San Diego: Association for Computing Machinery, 2016. p. 1135–1144. DOI: https://doi.org/10.18653/v1/N16-3020. Available at: https://aclanthology.org/N16-3020.pdf. Accessed on: 01 Mar. 2024.

NOTE ON AUTHORSHIP

Wellington: conceptualization, methodology and writing – original draft.

Eduardo: supervision and project administration.

Recebido em: 04/03/2024
Parecer em: 01/04/2024
Aprovado em: 24/05/2024