UNIVERSIDADE FEDERAL DE PERNAMBUCO
CENTRO DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIAS DA COMPUTAÇÃO

ROGÉRIO LUIZ CARDOSO SILVA FILHO

ISOLATING VARIABLE EFFECTS IN SUPERVISED MACHINE LEARNING ILLUSTRATED IN EDUCATIONAL DATA MINING

Recife

2024

ROGÉRIO LUIZ CARDOSO SILVA FILHO

ISOLATING VARIABLE EFFECTS IN SUPERVISED MACHINE LEARNING ILLUSTRATED IN EDUCATIONAL DATA MINING

Tese de Doutorado apresentada ao Programa de Pós-graduação em Ciência da Computação do Centro de Informática da Universidade Federal de Pernambuco, como requisito parcial para obtenção do grau de Doutor em Ciência da Computação

Área de Concentração: Inteligência Computacional

Orientador (a): Paulo Jorge Leitão Adeodato

Coorientador (a): Kellyton dos Santos Brito

Recife

2024

Catalogação na fonte
Bibliotecária: Luiza de Oliveira/CRB 1316

S586i Silva Filho, Rogério Luiz Cardoso
Isolating variable effects in supervised machine learning illustrated in educational data mining / Rogério Luiz Cardoso Silva Filho – 2024.
131 f. il., tab., fig.

Orientador: Paulo Jorge Leitão Adeodato.
Coorientador: :Kellyton dos Santos Brito.
Tese (Doutorado) – Universidade Federal de Pernambuco. Centro de Informática. Programa de Pós-graduação em Ciência da Computação, Recife, 2024.
Inclui referências.

1. IA explicável. 2. Aprendizagem de máquina interpretável. 3. Explicadores globais. 4. Mineração de dados educacionais. 5. Importância de variáveis. I. Adeodato, Paulo Jorge Leitão. II. Brito, Kellyton dos Santos. III. Título.

006.31 CDD (23. ed.) UFPE - CCEN 2024 – 94

Rogério Luiz Cardoso Silva Filho

“Isolating Variable Effects in Supervised Machine Learning Illustrated in Educational Data Mining”

Tese de Doutorado apresentada ao Programa de Pós-Graduação em Ciência da Computação da Universidade Federal de Pernambuco, como requisito parcial para a obtenção do título de Doutor em Ciência da Computação. Área de Concentração: Inteligência Computacional.

Aprovada em: 18/04/2024.

_____

Orientador: Prof. Dr. Paulo Jorge Leitão Adeodato

BANCA EXAMINADORA

_____

Prof. Dr. Ricardo Bastos Cavalcante Prudêncio
Centro de Informática / UFPE

_____

Prof. Dr. Alex Sandro Gomes
Centro de Informática / UFPE

_____

Prof. Dr. Martin Carnoy
Stanford Graduate School of Education / Stanford University

_____

Profa. Dra. Karla Patricia Santos Oliveira Rodríguez Esquerre
Departamento de Engenharia Química / UFBA

_____

Prof. Dr. Bruno Campello de Souza
Departamento de Ciências Administrativas / UFPE

Dedico este trabalho aos meus pais, Rogério e Madalena, que excederam todos os seus limites para me dar a melhor educação, a minha amada esposa Alice, e ao meu querido filho Antônio, por todo amor e complacência.

ACKNOWLEDGEMENTS

Concluir um doutorado é como alcançar o cume de uma imensa montanha. A sensação de dever cumprido ao olhar para baixo e perceber toda a trilha percorrida, as paradas de descanso e tudo aquilo que se une para levar ao topo reflete bem a jornada de um doutorado. Ambos começam com um sonho, sereno, que vai se complicando e oferecendo inúmeras oportunidades de desistir, mas, ao final, fazem você refletir orgulhoso: "Como pude chegar até aqui?"

É neste momento que me encontro, e mesmo tendo passado grande parte desses últimos anos sozinho em frente a um computador, tenho convicção de que só pude chegar aqui por estar apoiado em muitos. Como em uma montanha, onde, apesar do esforço individual, o alcance do topo normalmente depende de guias, cozinheiros e carregadores.

Durante minha jornada acadêmica, tive o privilégio de contar com pessoas e instituições. Sou imensamente grato pelo apoio familiar de Maria Alice e Antônio, pela base construída por meus pais Rogério e Madalena, por minha avó Joana e tia Enedina (*in memoriam*) e pela irmandade de Ana Caroline e Daniel.

Agradeço ao professor Paulo Adeodato, um grande pensador que me guiou. Seus ensinamentos me acompanharão por toda a vida. Meu reconhecimento também ao professor Kellyton, um colega que se tornou coorientador. Aos amigos e a todos que torceram por mim, essa conquista também tem uma parte de vocês.

Também sou grato ao IFNMG. Agradeço o apoio da gestão e dos colegas de trabalho, que corajosamente puderam absorver minhas responsabilidades. Sou grato à Fundação Lemann por financiar meu intercâmbio e me permitir conhecer professores que muito contribuíram para minha formação, como Martin Carnoy e Eric Bettinger.

Minha gratidão também vai para todos os educadores - professores, bibliotecários, escritores, supervisores e funcionários - dos ambientes educacionais que frequentei ao longo da vida, que, direta ou indiretamente, me fizeram acreditar no caminho da educação formal.

Agora, é hora de descer, o que também exigirá certo esforço, enquanto os pensamentos sobre os próximos desafios começam a surgir. Da montanha, leva-se algumas fotos que se tornarão pequenas diante de todo o aprendizado acumulado na subida. Da mesma forma, espero que a minha nova formação vá muito além do título e que possa ser útil na construção de uma sociedade melhor e mais justa.

"Should philosophy guide experiments, or should experiments guide philosophy?"(LIU, 2016, p. 17)

ABSTRACT

This thesis investigates the application of Explainable Artificial Intelligence (XAI) in Supervised Machine Learning (SML) models. The motivation for this study stems from the development of Educational Data Mining (EDM), an area that frequently uses such models to analyze and extract insights from large datasets. A central issue of this work is the challenge of generating global explanations for SML, particularly in cases where data independence is not guaranteed. This is a recurring but still underexplored problem in EDM. Neglecting data interdependencies can lead to biased explanations, overestimating irrelevant variables or disproportionately assigning importance to predictors with similar relevance. To address these challenges, this work builds on Accumulated Local Effects (ALE), a recent method for post-hoc global explanation that visualizes the impact of features. ALE's pseudo-orthogonality property allows for isolating individual variable effects, distinguishing it from widely used methods in EDM such as partial dependence plots and Shapley-based explanations. In a preliminary stage, ALE techniques is compared to other existing ones by using a new methodology that evaluates how different these techniques approximate the true variable effects in various contexts of data dependency. In a preliminary stage, ALE techniques are compared to other existing ones using a new methodology that evaluates how well these techniques approximate the true variable effects in various contexts of data dependency. Furthermore, based on the ALE promising results of this stage, this work proposes new ALE-based scores to measure the impact of variables in SML. The scores are model-agnostic and can report both the magnitude and direction of the individual impact of features. The scores prove to be efficient in various scenarios when compared to existing metrics on synthetic and real-world datasets. Moreover, an empirical study using data from Brazilian secondary schools not only confirms the usefulness of the new scores in a real-world scenario but also extends the contributions of this thesis by identifying and offering new perspectives on the determinants of Brazilian school success over more than a decade.

Keywords: explaineble AI; interpretable ML; global explainers; EDM; feature importance; ALE.

RESUMO

Esta tese investiga a aplicação de Inteligência Artificial Explicável (IAE) em modelos de Aprendizagem de Máquina Supervisionada (AMS). A motivação para esse estudo decorre do desenvolvimento da Mineração de Dados Educacionais (MDE), uma área de estudo que frequentemente emprega tais modelos para analisar e extrair conhecimentos de vastos conjuntos de dados. Um aspecto central dessa tese é o desafio de gerar explicações globais para AMS, particularmente em situações onde a independência entre os dados não é garantida. Esta é uma problemática recorrente, mas ainda pouco explorada na MDE. A negligência das interdependências entre os dados pode levar a explicações enviesadas, valorização excessiva de variáveis irrelevantes ou atribuição desproporcional de importância a preditores de similar relevância. Para resolver estes desafios, a tese baseia-se em um método recente para a visualização do impacto das variáveis em modelos supervisionados, conhecido em inglês como Accumulated Local Effects (ALE), que se refere à distribuição acumulada de efeitos locais. A propriedade pseudo-ortogonal de ALE permite isolar os efeitos de variáveis individualmente, distinguindo-a de métodos amplamente usados em MDE, como os gráficos de dependência parcial e explicações baseadas em valores de Shapley. Em uma etapa inicial, as técnicas ALE são comparadas a outras existentes utilizando uma nova metodologia que avalia quão bem essas técnicas se aproximam do efeito real das variáveis nos modelos em vários contextos de dependência de dados. Além disso, com base nos resultados promissores dessa etapa, este trabalho propõe novos escores baseados em ALE para medir o impacto das variáveis em modelos de AMS. Esses escores são agnósticos a modelos e podem capturar tanto a magnitude quanto a direção do impacto individual das variáveis. Os escores demonstram eficiência em vários cenários quando comparados com as métricas existentes em conjuntos de dados sintéticos e reais. Além disso, um estudo empírico utilizando os dados das escolas secundárias brasileiras não apenas ratifica a utilidade dos novos escores em um cenário do mundo real, mas também estende as contribuições desta tese ao identificar e oferecer novas perspectivas sobre os determinantes do sucesso escolar brasileiro ao longo de mais de uma década.

Palavras-chaves: IA explicável; aprendizagem de máquina interpretável; explicadores globais; mineração de dados educacionais; importância de variáveis; ALE.

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS AND ACRONYMS

AAR	ALE Absolute Range
ABX	Absolute Difference Between Explanations
AI	Artificial Intelligence
ALE	Accumulated Local Effects
AUA	Average Uncentered ALE
AUC_ROC	Area Under Receiver Operating Characteristic Curve
cs_PFI	Conditional Permutation Feature Importance
cs_PFI	Conditional Subgroup Permutation Feature Importance
EDM	Educational Data Mining
IML	Interpretable ML
KDD	Knowledge Discovery in Databases
LSA	large-scale assessment
MDI	Mean Decrease in Impurity
ME	Marginal Effects
ML	Machine Learning
MUA	Maximum Uncentered ALE
NMSE	Normalized Mean Squared Error
NN	Neural Network
PD	Partial Dependence
PFI	Permutation Feature Importance
RF	Random Foresst
RMSD	Root Mean Square Deviation
SHAP	SHapley Additive Explanation
UAS	Uncentered ALE at a Specific Value
XAI	eXplainable Artificial Intelligence

CONTENTS

1 INTRODUCTION

This thesis introduces and implements global eXplainable Artificial Intelligence (XAI) techniques to enhance Knowledge Discovery in Databases (KDD), particularly within the educational domain. By integrating methodologies from both Machine Learning (ML) and Education—fields distinguished by their unique terminologies - this work relaxes the use of certain terms for clarity. Here, 'variable', 'feature', and 'predictor' are used interchangeably to represent individual, measurable attributes of the phenomena observed in the datasets. Similarly, 'label' 'target', 'dependent variable,' and 'outcome' denote the variables whose values the models aim to predict. Additionally, the term 'marginal effects' is utilized in two contexts. In econometrics, it indicates the incremental effect, while in statistics, it pertains to the probability distribution of a variable. Contextual clarifications are provided throughout to avoid misinterpretations.

1.1 CONTEXTUALIZATION

In modern society, data is a critical resource for guiding human decision-making processes. More recently, with the technological advances of the twentieth century, our capability to store, process, and analyze large volumes of data has put forward the data potential to enhance human activities (PROVOST; FAWCETT, 2013). As we transition into this new era characterized by data ubiquity, emergent paradigms in data analysis have arisen to meet the challenges and opportunities presented by this voluminous and complex data landscape.

In 2001, Breiman called for the use of an algorithm approach as a more accurate and informative alternative to the use of data to solve problems (BREIMAN, 2001b). The algorithmic modeling he refers to is ML, which, unlike traditional approaches which adjust data for a predefined model, learns empirically from data to estimate functions for making predictions on new data. According to Breiman, ML tools facilitate a move away from exclusive reliance on parametric models, adopting a more diverse set of tools. This approach could enable researchers to move beyond confirmatory research based on theory models and also allow them to derive new theories directly from data (MOLINA; GARIP, 2019).

In supervised ML (VAPNIK, 1999), models are iteratively optimized to minimize out-of-sample prediction error, a focus that diverges from disciplines more concerned with understanding the underlying data-generating processes (MULLAINATHAN; SPIESS, 2017). However,

(SHMUELI, 2010; ZHAO; HASTIE, 2021) argue that a model exhibiting both strong predictive performance and consistent assumptions closely approximates the underlying natural laws governing the data. This dual focus not only highlights the significance of ML

's predictive capabilities but also the critical importance of ensuring that models align with real-world phenomena. Further reinforcing this notion is the stance taken by (CAO, 2009), who stresses the importance of aligning data mining models with complex real-world challenges. Cao advocates for the integration of domain-specific knowledge throughout the entire KDD process, a strategy that promises to deliver more reliable and actionable insights.

Nonetheless, the emphasis on predictive performance in ML has prompted researchers to adopt increasingly complex models, often at the expense of interpretability. For instance, the coefficients in additive linear models or the rules derived from decision trees offer straightforward interpretability, explicitly mapping input features to model outputs (MOLNAR, 2023). In contrast, opaque models like neural networks and ensemble methods, though potentially superior in prediction, do not readily reveal the mechanisms relating input features to outcomes (LINARDATOS; PAPASTEFANOPOULOS; KOTSIANTIS, 2020). The complexity of these models poses significant challenges for interpretation within a KDD process, especially when seeking scientific insights and explanations for wrong decisions made, particularly before the Justice.

Within this context, and given the widespread adoption of ML in many tasks, the field of XAI has quickly become an important focus within the larger field of ML. XAI aims to explain the reasoning and decision-making processes of these models in a human-understandable manner (MILLER, 2019). These explanations are valuable not only for applications aimed at deriving insights from data but also for those whose primary objective is prediction. For instance, while categorizing a patient's health status in a hospital or predicting a student's likelihood of dropping out is beneficial, understanding the factors driving these predictions can significantly enhance the utility of the model by facilitating targeted interventions (RAZAVIAN et al., 2015; PELLAGATTI et al., 2021; BERENS et al., 2019). Furthermore, the transparency of ML models in sectors like criminal justice (WANG et al., 2023) and finance (BUSSMANN et al., 2021; CHEN et al., 2023) are increasingly mandated by legal and ethical considerations.

1.2 MOTIVATION AND SCOPE

The use of ML models in the educational domain, called Educational Data Mining (EDM) has gained considerable attention in recent literature (ROMERO; VENTURA, 2020). The EDM

encompasses a diverse array of applications, ranging from predicting student dropout rates (ARAQUE; ROLDÁN; SALGUERO, 2009; AGUIAR et al., 2015) to facilitating the creation of personalized learning paths (FANCSALI et al., 2018). Beyond predictive accuracy, the interpretability of these models can be critical for their responsible integration into educational settings.

In certain contexts, the accuracy of predictions may even take a back seat to the insights gained from model explanations. For instance, in the field of educational assessment, EDM has emerged as a potent tool for analyzing large-scale assessment (LSA) datasets. These datasets are invaluable for identifying key variables impacting educational systems, aiming to provide empirical evidence to inform discussions on educational policies (HERNÁNDEZ-TORRANO; COURTNEY, 2021). EDM allows for the extraction of knowledge from significant relationships within these extensive databases (GAMAZO; MARTÍNEZ-ABAD, 2020). Unlike traditional approaches that rely on theoretical distribution, EDM models are developed and validated using empirical data. This flexibility enables researchers to revisit and refine existing theoretical models (HUANG et al., 2003).

There are many discussions of what has consisted of a model explanation, and they can be delivered in many ways (GUIDOTTI et al., 2018). Feature-based explanations are among the most prevalent, focusing on identifying critical features that influence model output at either the example level (local) or the sample level (global). At the global level, these feature-based explanations are commonly conveyed through summary metrics reporting scores of the overall feature contribution or through plots detailing its different effects over the data sample (FILHO; BRITO; ADEODATO, 2023a).

Moreover, the explanation methods can be internal to the model (intrinsically) as the coefficients of linear regression and the path of a tree or by applying a second model that analyzes the initial one (post-hoc). Another criterion to classify these methods is related to their generalizability, whether they are model-specific or model-agnostic. While every intrinsic method is specific, all model-agnostic work is in a post-hoc framework (FILHO; ADEODATO; BRITO, 2021).

This thesis is principally concerned with the critical evaluation of the global explanations in the post-hoc framework, which has raised concerns due to two key challenges. The first challenge pertains to the predictive model itself: high predictive performance is not a sufficient indicator that the model has captured the true relationships in the data. Rather, it may be exploiting spurious correlations, thereby limiting the validity of any insights extracted from the model (MITTELSTADT; RUSSELL; WACHTER, 2019). The second challenge, which is the focus

of this thesis, lies in the explanation methods. Even if the model could accurately captures the true data relationships, the explanation methods may not effectively illustrate how the model actually works (RUDIN, 2019; MITTELSTADT; RUSSELL; WACHTER, 2019).

Incorporating domain knowledge and ensuring the model's structural integrity can help mitigate the first issue (FRYE et al., 2021; ZHAO; HASTIE, 2021). However, the problem of collinearity - where variables are interdependent - remains a significant issue for the second challenge (HOOKER; MENTCH; ZHOU, 2019). Collinearity becomes problematic when overlooked in the explanations. This issue is especially pronounced in methods that rely heavily on the structure of the model while neglecting the crucial interrelationships within the datasets.

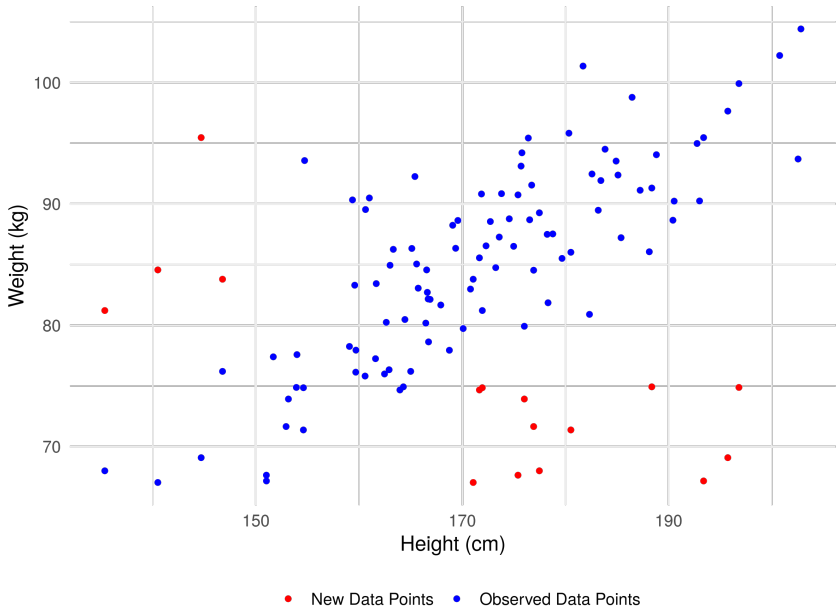
Recently, in the EDM field, many scholars have been relying on global explanations derived from XAI techniques that assume data independence in an attempt to extract knowledge from data. However, data independence might be a strong assumption for many EDM tasks, especially when using structural data based on personal and contextual information. This kind of data tends to be highly correlated with strong inter-feature dependencies. For example, in predicting student success based on LSA data, socioeconomic variables often have a significant influence on student achievement, along with other contextual variables such as demographics, school environment and process, parental education level, and access to educational resources (COLEMAN, 1968; ANDRADE; SOARES, 2008).

Standard supervised ML is widely recognized for its emphasis on performance. This focus can lead to learned functions that do not accurately reflect the true data-generating process behind student success (first challenge). Consequently, insights into this phenomenon are valuable only when the predictive function is thoroughly explained within the context of the data distribution used for training. An explainable approach that overlooks the data interrelationships may not yield reliable explanations (the second challenge). This issue becomes more critical when data dependencies naturally introduce biases in the explanations. For example, low relevant variables can be attributed a high relevance only due to co-dependency with a high relevant variable.

A common strategy of global post-hoc techniques involves modifying the value of a specific feature as a sign of its impact on the model (SCHOLBECK et al., 2020). Essentially, features are manipulated to generate new predictions. From a model-centric viewpoint, ignoring data relationships in these interventions can lead to misalignment with the actual data distribution. This may lead to the creation of unrealistic data points and, consequently, the potential for unusual predictions. For example, Figure 1 illustrates this situation in a dataset that controls

the weight and height of adults. If data dependencies are overlooked during interventions, scenarios like very low weight paired with very tall height might be erroneously generated (red points in the Figure), which are improbable or even physically implausible in reality. Predictions based on these unlikely data points can thus yield unreliable results, significantly reducing the practical value of the model's explanations. This issue is referred to as the "extrapolation" problem (MOLNAR et al., 2022; RUDIN, 2019). This can also pose a problem in educational datasets, where points outside the distribution are used to compute feature effects and inform decision-making.

Figure 1 – Illustration of the extrapolation problem. Blue dots are the observed data points. Red dots are new data points derived from interventions that do not align with the actual data distribution.



Source: self-provided

In light of these challenges, this thesis is motivated by the need for more rigorous methods to determine global feature contributions in EDM, especially when feature independence cannot be assumed. This work is significant not only for educational practitioners engaged in data-driven tasks, who would benefit from more reliable interpretations of ML models, but also for the ML research community. Researchers can generalize the methods and ideas presented here to advance the literature on XAI, fostering a more comprehensive adoption of ML systems.

1.3 PROBLEM STATEMENT

In the realm of supervised learning within EDM, consider the function $Y = f(X)$, where $X = (x_1, x_2, \ldots, x_n)$ represents a set of structured variables, M denotes the variance-covariance matrix and g a post-hoc XAI model. Assuming f is an effectively performing model and can be considered as a source of knowledge about an underlying phenomenon, insights about X can be gleaned from g by analyzing how f utilizes X to predict Y, especially when considering the dependencies encapsulated in M.

Moreover, in the context of global explanations, it is expected that the explanation contained in g represents the individual role of each component to which g refers. Specifically, the attributed relevance to $x_1 = g(x_1)$ should pertain solely to the role of x_1. Similarly, $g(x_1, x_2)$ should represent the combined effect of x_1 and x_2 exclusively. This property will ensure an interpretation of g akin to the coefficients of linear models, where the β values indicate the individual impact of each feature, assuming the model's conditions are met. Nevertheless, if M is not taken into account or is not correctly handled by g, biases may skew the interpretation of individual components of f resulting in mixed effects of features or even unrealistic feature effects representations.

A review of the literature on applying XAI in EDM primarily focusing on extracting knowledge from data, presented in Section 2, indicates that commonly used techniques for explaining opaque models may result in unrealistic explanations of X. This is largely due to a lack of constraints to address M in the computation of g. Specifically, the review highlights the use of Partial Dependence (PD) and SHapley Additive Explanation (SHAP) as the predominant tools for informing feature effects through plots.

The PD plots, introduced in 2001 by Friedman (FRIEDMAN, 2001), serve as one of those pioneering techniques for visualizing the effects of predictors. In recent years, SHAP (LUNDBERG; LEE, 2017) has gained widespread acceptance in both industrial and academic settings (BHATT et al., 2020) for delineating local and global effects of features. However, prevalent implementations of PD and SHAP overlook the matrix M. In other words, the feature relationships presented in M cannot be maintained during feature effect computations. This oversight can compromise the accurate interpretation of the individual relevance of components in f.

In EDM, one example of how this problem can arise involves variables related to socioeconomic factors. Socioeconomic status is often gauged using multiple correlated variables, such as parental education level, access to cultural resources, and availability of technological

devices. These variables frequently play a crucial role in many educational contexts. Variables that may not be inherently significant to an underlying predictive function but are correlated with those significant ones could be inappropriately emphasized if the correlation presented in M is extrapolated, introducing bias to the explanations of the individual role of features and their interactions on function M.

The Accumulated Local Effects (ALE) plots (APLEY; ZHU, 2020) were recently introduced as an alternative to elucidate feature effects. The primary motivation behind ALE is to address the extrapolation problem by ensuring a pseudo-orthogonality property. This property enables ALE to approximate orthogonality — where components operate independently from each other — thereby facilitating the isolation of individual explanations for each component within a predictive function. Consequently, ALE presents itself as an effective and straightforward alternative for explaining models where data exhibits significant dependence as is common in most EDM applications. However, to the best of our knowledge, it has not yet been employed.

Moreover, it's important to note that ALE has primarily been used to illustrate feature effects through plots of partial derivatives. These plots demonstrate the varying effects of a single variable across its value range, offering detailed insights. However, this level of detail may not always be practical. For example, when analyzing multiple variables simultaneously, such complexity can reduce human intelligibility, considering that an average person is capable of processing only a limited amount of information at once (MILLER, 1956). Furthermore, plots are not feasible to describe interactions between more than two features (APLEY; ZHU, 2020).

In the context of KDD's systematic processes, simpler metrics like scores offer a pragmatic approach. They can initially identify key variables, which then facilitates more in-depth exploration of the roles these features play. Scores are also useful for feature selection, providing a straightforward method for determining the most influential variables. Furthermore, scores are well-suited for comparative and trend analysis, aiding in the evaluation of the relevance of features across different educational systems and over time. This adaptability makes scores a valuable tool for broader analysis in educational settings, akin to the traditional coefficients in linear models, which are commonly used in education research.

Building upon the review presented in Chapter 2, it becomes evident that methods such as the Mean Decrease in Impurity (MDI) derived from tree-based models, along with the PFI and average SHAP values, are among the most prominent techniques for elucidating feature attribution through scores. The MDI assesses a feature's relevance by averaging the degree to which the feature is used as a split criterion during tree construction. Its variations PFI, which

was initially designed for random forests and further extended for other tree-based models, was also further conceptualized to be applied in a model-agnostic manner. The PFI takes the difference (or ratio) of model performance between the baseline model and the model when the feature is randomly permuted. The PFI is designed to provide a score for each feature based on how much difference replacing the feature with noise makes in predictive performance. PFI addresses some well-known limitations of MDI, such as its bias towards features with a high number of categories and continuous variables (LI et al., 2019). However, PFI, as the average SHAP, is a permutation-based method that computes explanations out-of-distribution, thus raising concerns about their applicability in assessing the importance of features in dependent datasets (STROBL et al., 2008; RUDIN, 2019; NEMBRINI, 2019; NICODEMUS, 2011).

Beyond the challenge of managing dependent data, ranking-based metrics, which illustrate the relative importance of variables in model performance such as PFI may not effectively illustrate the direct relationship between features and the target variable. This is a crucial aspect for EDM practitioners who seek to glean insights from data. It would be more advantageous if these scores also represent the individual contributions of features to predictions, rather than solely to performance. The scores should be able to be either positive or negative, indicating the specific nature of the relationship between the features and the target variable. Adopting this approach would align more closely with the interpretation of coefficients in traditional linear models.

Therefore, this thesis problem centers on the need for a robust method that can elucidate the isolated roles of each component in the predictive function. Given this challenge and ALE properties, the following question guides this research: Can ALE be incorporated into the range of XAI tools to be used in EDM to inform more accurate feature contribution either by plots and scores, even when data are not independent?

1.4 OBJECTIVE AND CONTRIBUTIONS

Addressing these inquiries, the main objective of this thesis is to assess ALE as a reliable alternative to explain the individual and isolated effects of features and their interactions in the supervised learning paradigm. Specifically, it focuses on ensuring robustness in the presence of dependent data, aiding in extracting knowledge from educational datasets.

This work aims to bridge this gap by not only critically evaluating the limitations of existing explanations but also introducing novelties that allow a more trustworthy adoption of ML in

EDM. Specifically, the following research questions address the core objectives of the study.

RQ1 - How do widely used feature effects techniques compare with ALE in accurately identifying true feature effects considering different inter-data dependencies?

By responding to **RQ1**, the thesis aims to raise empirical evidence about the robustness of ALE in recovering the role of features in supervised models under correlated data. While the properties of ALE have been previously delineated, primarily through mathematical and qualitative frameworks (APLEY; ZHU, 2020; MOLNAR, 2023), a notable gap remains in empirical quantitative analysis, particularly in evaluating how ALE strategies differ in explanations compared to commonly used techniques such as PD and SHAP. This gap not only underscores the need for a thorough comparative analysis of ALE with established methods, as suggested in (MOLNAR et al., 2022), but also highlights its potential for enhancing explanations in the field of EDM.

RQ2 - How effectively can score-based explanations derived from the ALE framework report individual and isolated attribution of the features in terms of their magnitude and direction compared to existing methods?

Addressing **RQ2**, this thesis aims to fill a gap in the area of score-based explanations, which is the most prevalent approach in EDM. By adopting the ALE framework, this work introduces new metrics that surpass the limitations of current methods, particularly in the context of correlated data. This advancement will facilitate knowledge discovery in educational data using supervised machine learning.

These research questions will be answered sequentially, aiming to provide **two main contributions**: **1) empirical evidence of ALE robustness compared with currently used methods in EDM** and **2) a new set of score-based metrics of feature effects size.**.

The first main contribution of this thesis is the evaluation of ALE against other widely used techniques in EDM, specifically in scenarios involving dependent data. This contribution fulfills the need for a thorough analysis of different strategies for managing data dependencies in the context of post-hoc global feature effects. Furthermore, it enhances the XAI literature by introducing a novel methodology for benchmarking feature effects. This methodology evaluates the robustness of various feature effect techniques in accurately representing the actual data-generating process. The design of this methodology involved the use of synthetic data, which enables a direct comparison between the known data-generating process and the outcomes provided by the explanation techniques.

The second major contribution of this thesis is a novel set of metrics developed to quantify

feature effect sizes. Building on prior research, which introduced scores summarizing graph-based techniques, (LONG; LONG, 1997; GREENWELL; BOEHMKE; MCCARTHY, 2018; LEE et al., 2023) this work introduces four innovative metrics inspired by the ALE framework. Each metric is designed to provide unique insights into the significance of features, offering diverse perspectives on their importance. These metrics are model-agnostic, suitable for a range of model types, and designed to reveal the extent and direction of feature effects, similar to the way traditional coefficients do in educational analysis. The effectiveness of these metrics has been tested, demonstrating their capacity to identify key variables and to isolate the effects of features, even among highly correlated variables. This validation was conducted using both synthetic and real-world datasets.

1.5 OUT OF SCOPE

Since this thesis encompasses a broad context, it is important to highlight a set of subjects that are outside the scope of this thesis:

- **Introduce new strategy to deal with the extrapolation issue**. The extrapolation problem is a well-known issue in the literature and is sometimes even considered a trade-off, where it is not possible to be fully adherent to the data and model simultaneously (LUNDBERG et al., 2020; CHEN et al., 2020). Rather than proposing novel strategies, this thesis focuses on discussing and expanding the understanding and application of existing strategies within this context. The primary focus is on extracting knowledge from data.

- **Guidelines for Explaining EDM Models**: This thesis advocates for the ALE framework as a favorable and appropriate alternative for globally elucidating the role of features in predictive models, especially within educational contexts dealing with correlated data. However, it does not intend to set rigid guidelines for explaining EDM models. As discussed in Chapter 2, explanations can vary both in form and function, tailored to suit the explainability requirements of the intended audience. The diversity of these explanations and their integration can substantially enhance knowledge discovery.

- **Compromise with causality**. While the primary aim is to provide dependable insights into the mechanisms generating the underlying data, it's important to acknowledge that traditional supervised models do not ensure an accurate reconstruction of the data-

generating process. Often, these models rely on spurious correlations for making predictions rather than establishing causal relationships. Consequently, this work does not delve into any causality-related issues.

1.6 SCIENTIFIC PRODUCTIONS

This section presents the publications that have arisen from this work. The research began before the focus of this thesis was formally established, consistently centered around applying supervised learning to extract knowledge from educational data. This exploration aimed to aid decision-making and support the formulation of educational policies. These endeavors have culminated in multiple publications, both in conference proceedings (CP) and peer-reviewed journals (JP).

Table 1 – Conferences

ID	Reference
CP1	Silva Filho, R.L.C. ; Adeodato, P.J.L. . Data Mining Solution for Assessing the Secondary School Students of Brazilian Federal Institutes. In: **8th Brazilian Conference on Intelligent Systems** (BRACIS), 2019, Salvador, p. 574
CP2	Silva Filho, R. L. C. ; Adeodato, P. J. L. ; dos Santos Brito, K. . Interpreting Classification Models Using Feature Importance Based on Marginal Local Effects. In: **10th Brazilian Conference on Intelligent Systems(BRACIS)**, 2021, São Paulo, p. 484

Table 2 – Journals

ID	Reference
JP1	Silva Filho, R. L. C., Brito, K., Adeodato, P. J. L. (2023). A data mining framework for reporting trends in the predictive contribution of factors related to educational achievement. **Expert Systems with Applications**, 221, 119729. $https://doi.org/10.1016/j.eswa.2023.119729$
JP2	Silva Filho, R. L. C., Brito, K., Adeodato, P. J. L. (2023). Leveraging Causal Reasoning in Educational Data Mining: An Analysis of Brazilian Secondary Education. **Applied Sciences,** 13(8), 5198. $https://doi.org/10.3390/app13085198$
JP3	Silva Filho, R. L. C., Brito, K., Adeodato, P. J. L. (2023). Beyond scores: A machine learning approach to comparing educational system effectiveness. **Plos One**, 13(8), 5198. $10.1371/journal.pone.0289260$
JP4	Silva Filho, Carnoy M. (2023). Trends in social class and race achievement gaps among secondary school graduates in Brazil. **Under revision: Large-scale Assessment in Education**, 2023.
JP5	Silva Filho, R. L. C. ; Adeodato, P. J. L. ; dos Santos Brito, K. . Measuring extrapolation: an comprehensive analysis of feature effects. **"Working Paper"**

Source: self-provided

1.7 OVERVIEW

This thesis is organized into a total of six chapters. In this chapter (Chapter 1), the motivations for carrying out this research were presented together with a brief overview of the objectives and research questions.

Chapter 2 provides an overview of the foundational areas and concepts critical to this thesis. It outlines the fields of XAI which are related to the gaps this thesis aims to fill. Furthermore, an overview of EDM and its intersection with XAI is provided, in addition to the formal definitions of techniques frequently cited throughout the thesis.

Chapters 3 and 4 present the main contributions of this thesis. Chapter 3 formally delineates how different strategies to handle the data relationships affect post-hoc XAI techniques. It further empirically demonstrates the robustness of ALE in recovering the true effects of features in comparison with other techniques on different data dependence scenarios.

Motivated by the results of Chapter 3, Chapter 4 subsequently introduces ALE-based metrics for assessing feature effect size. Each chapter begins with an introduction section that provides the necessary context and motivation, along with a brief review of related works that are specific to the chapter's individual contribution. This structure results in a smoother overlap between chapter introductions, which, while not ideal, is necessary to ensure clarity and

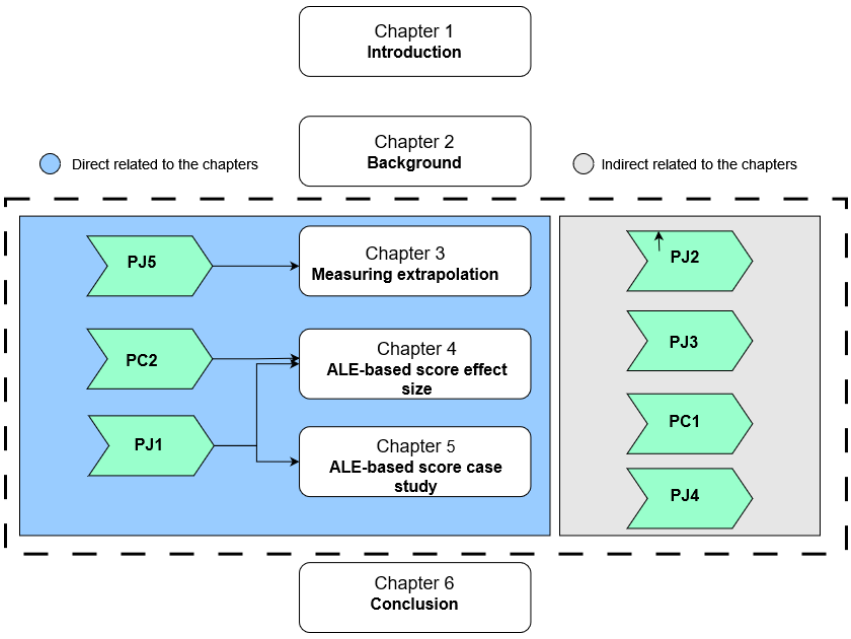
better position its contributions to the literature. The chapters also detail the methodology, experiments, and results of each contribution.

Chapter 5 showcases the usefulness of the ALE-based metrics developed in Chapter 4 to a practical case study. It presents a data-mining solution to investigate what and how variables have impacted Brazilian secondary school performance over a period of 11 years.

Finally, Chapter 6 presents the concluding remarks and discusses the main contributions of this thesis, and directions are outlined for possible future research.

Figure 2 depicts the overview of this thesis through a diagram. The diagram, delineated by a blue box, illustrates the direct correlation between specific chapters of the thesis and their corresponding scientific publications, highlighting the integration of these publications into the thesis discourse. In contrast, the gray box encapsulates additional publications that, while not directly tied to individual thesis chapters, fall within the broader scope of the thesis topic - extract knowledge from educational data using ML. Notably, all presented publications are authored by the thesis candidate as the first author during the Ph.D. period.

Figure 2 – Overview of Thesis Structure and Related Publications: Publications directly related to thesis chapters are shown in blue boxes, while additional publications within the thesis topic are in gray boxes. All works are first-authored



Source: self-provided

2 BACKGROUND AND LITERATURE REVIEW

2.1 SUPERVISED LEARNING

Supervised learning is a subfield of ML in which a model is trained on a labeled dataset to perform predictive tasks in a unseen dataset. The objective is to find a function $f : X \to Y$, where X is the feature space and Y is the output space, such that the function approximates the underlying mapping from input features to outputs as closely as possible.

Specifically, given a labeled dataset $D = (x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, where $x_i \in X$ and $y_i \in Y$, the supervised learning algorithm aims to minimize a loss function $L(f(X), Y)$ over D, defined as:

$$L(f) = \sum_{i=1}^{n} L(f(x_i), y_i)$$

Where, L measures the discrepancy between the predicted output $f(x_i)$ and the true label y_i. To enhance the model's generalization capabilities, f is evaluated on an statistically independent dataset, thereby mitigating the risk of optimistic empirical performance estimation.

2.1.1 Contrast with traditional statistical methods

Despite the models used in traditional statistics can be seen as one of the tools available in ML (HASTIE ROBERT TIBSHIRANI, 2014), there is a fundamental difference in how they estimate the weight of functions. To elucidate the distinctions between supervised learning and traditional statistical models, lets consider a linear model within the supervised learning framework, defined as:

$$y = \beta_0 + \beta_1 x + \epsilon \tag{2.1}$$

In contrast to the Ordinary Least Squares (OLS) approach prevalent in traditional statistical analyses, the objective in the context of supervised learning is also to optimize the coefficients β in a manner that minimizes out-of-sample error. Specifically, OLS optimizes β by minimizing only in-sample error, without explicit consideration for out-of-sample generalizability. The key divergence stems from supervised learning's strategic focus on balancing the bias-variance

trade-off, thereby allowing a certain level of bias (in-sample error) to mitigate excessive variance (out-of-sample error) (ATHEY; IMBENS, 2019).

The objective function for linear regression under supervised learning paradigm can be formalized as:

$$\sum_{i=1}^{n}(y_i - f(x_i))^2 + \lambda R(f) \tag{2.2}$$

In this equation, $(y_i - f(x_i))^2$, represents the in-sample error, while the regularization term $R(f)$ acts to prevent overfitting by constraining the model's complexity, thereby reducing the out-of-sample error. The chosen regularization parameter λ plays a crucial role in modulating the extent of this constraint, thereby influencing the model's generalization performance (HASTIE ROBERT TIBSHIRANI, 2014).

2.2 XAI

The wide use of Artificial Intelligence (AI) and ML has increasingly emphasized the importance of transparency and user comprehension of model behaviors, primarily under the terms Explainable AI (XAI) and Interpretable ML (IML). Despite the significant growth of this research area in recent years (ARYA et al., 2019), foundational works in the field can be traced back to the 1980s (FAGANT; SHORTLIFFE; BUCHANAN, 1980; BAREISS; PORTER; WIER, 1988). While some authors argue that XAI and IML can be conceptually distinct (WATSON, 2022), they are more commonly used interchangeably in the broader scientific literature as both terms share the objective of enhancing the transparency of ML models (MOLNAR, 2022)

In this thesis, the term "interpretability" will used to refer to its common dictionary meaning, while "explainability" will be specifically employed to describe the systematic extraction of knowledge about predictive models.

2.2.1 Inherent interpretable models

Inherent interpretable models, often referred to as intrinsically interpretable models, are those models distinguished by their transparent and easily understandable internal mechanics. These model provide explicit explanations of the relationships between input features and output predictions, facilitating a deeper understanding of their recommendations in the

decision-making processes.

Examples of such models include linear regression, decision trees, and induction rules. In the linear regression 2.1, for instance, each β coefficient quantifies the change in a dependent variable for a one-unit change in an independent variable, assuming all other variables are held constant. In other words, the model additive parametrization allows an isolated interpretation of the effects of individual features. Many other adaptions allow a linear model to capture more complex relationships (HASTIE ROBERT TIBSHIRANI, 2014), such as interactions and non-linearity. Nevertheless, in models that involve transformations of this linear predictor into other discrete outcomes, such as in the logit and probit models, the β interpretation is not straightforward and limited (MOOD, 2017; LONG; LONG, 1997).

The decision trees categorize outcomes based on decision rules at each node. On the other hand, inducing rules out of tree structure do not narrow the dimensional space as it occurs in the trees. These induced rules are clear statements in the natural language of how inputs lead to outputs in several different perspectives (FILHO; ADEODATO, 2019).

The primary advantage of using inherent interpretable models is their ease of interpretation, which is especially beneficial to high-stakes decisions (RUDIN, 2019). However, they are often outperformed by more complex models when it comes to predictive performance (LOYOLA-GONZALEZ, 2019). The simplicity that makes them easy to interpret can also be a drawback, as it might lead to the oversimplification of intricate relationships in the data. This can be a significant limitation when dealing with complex systems where multiple variables interact nonlinearly.

2.2.2 Post-hoc explainable techniques

When a second model is used to explain the first, it is categorized as a post-hoc explainable technique. Model-agnostic explanation model is any function g that approximates the original model f (LUNDBERG; LEE, 2017). While intrinsically interpretable models provide insights into predictions using their internal components, post-hoc techniques treat models as opaque, relying solely on their prediction function and data (MOLNAR, 2022).

Post-hoc techniques are typically model-agnostic and offer the flexibility to explain various types of models, including those that are transparent, in an effort to enhance existing explanations. For instance, in a logistic regression model, a model-agnostic technique can illustrate the individual feature effects across the entire range of the feature values, whereas the coefficients

only indicate the order of feature contributions. Similarly, understanding feature effects can complement the intrinsic rule-based explanations provided by a decision tree.

2.2.3 Measuring explainability

Explainability is a domain-specific notion and has a big criticism of the lack of formalization (RUDIN, 2019; WATSON, 2022; LIPTON, 2018). Explanations can take various forms, and there isn't a clear definition of what constitutes an explanation. Moreover, explanations can differ based on the type of input variables. For images, explanations are often visualized as heatmaps, whereas for text inputs, they typically involve highlighting text passages or emphasizing words (MOLNAR, 2022).

This thesis focuses on tabular data. For certain tasks within this domain, visual graphs by using plots may be the preferred form of explanation, while others might favor text or scores. This variation makes it challenging to find a widely accepted definition of explainability, even within this narrowed scope. Such diversity in explanations presents a challenge in defining quantifiable evidence for the field (DOSHI-VELEZ; KIM, 2017).

Unlike supervised learning, where much of the literature has advanced based on clear benchmarks of model performance, the sub-field of XAI or IML still faces vagueness in definitions. This is due to the challenge of measuring the trustworthiness of model explanations, as there is no ground truth for comparison in the real world which is only known by its observable data. Determining which explanation is superior is also difficult (ARYA et al., 2019), even for inherently interpretable models. For instance, we can't always say whether a decision tree path is more or less clear than a linear model's coefficients (MOLNAR, 2022).

Considering the variety of ways in which explanations can be derived, their evaluation depends on the intended purpose of use. For example, one can assess how effectively humans utilize explanations (EHSAN et al., 2021; WANG et al., 2019), or evaluate the explanatory function itself by measuring aspects such as size or sparsity (YANG; RUDIN; SELTZER, 2017; USTUN; RUDIN, 2016; CLAASSEN; MOOIJ; HESKES, 2013). Additionally, it is possible to quantify certain aspects related to explanations, such as evaluating the extent to which explanations predict model outputs (CHEN et al., 2022; LAKKARAJU; BACH; LESKOVEC, 2016). For benchmarking purposes, a common practice is the use of synthetic data with a known data-generating process. This approach facilitates the comparison of actual explainability with expected explanations in various contexts.

2.2.4 Who needs explanation?

The demand for ML explanations is pertinent across various sectors with the specific needs and objectives varying by domain and stakeholder diversity. Model explainability is not merely a desirable attribute but can be a crucial aspect for reasons ranging from model debugging to scientific exploration. This section delineates the roles of key stakeholders and the significance of explanations beyond performance within their respective domains.

Model creators, typically the developers of ML models, find interpretability crucial for debugging tasks (BHATT et al., 2020). It is important to know how the model relies on features to make predictions in order to fix unexpected behavior. For instance, the model creator might be interested in a model that makes decisions based on meaningful features rather than sensitive features in order to enhance generalizability or fairness. Such scrutiny cannot be achieved by only observing performance.

Operators, who use a model's outputs in their tasks, also require an understanding of the decision-making rationale. For instance, classifying a patient in a hospital into a particular health status should not be particularly helpful. It could be more useful to investigate the conditions that have contributed to this (RAZAVIAN et al., 2015), and this becomes even more crucial in the event of legal matters. Additionally, in the education domain, understanding why a student might drop out could be more valuable than just predicting it (PELLAGATTI et al., 2021; BERENS et al., 2019). This is because, as in medicine, the treatment depends on the probable cause.

The people who are subject of decision-making also have to get a kind of explanation. For instance, a loan approval model may recommend the rejection of an applicant based on specific financial variables. Understanding the rationale behind a decision empowers the applicant to make informed future choices or contest an unjust or biased decision. Regulatory examiners, often working in regulated industries, expect similar explanations. They are responsible for auditing ML models to ensure compliance with industry standards and ethical norms (CHEN et al., 2023; FLORES; BECHTEL; LOWENKAMP, 2016).

Finally, data analysts are increasingly utilizing ML models for tasks aimed at understanding data-generating processes in both industrial and scientific research contexts (FREIESLEBEN et al., 2022; FILHO; BRITO; ADEODATO, 2023a). These models often supplant traditional statistical methods due to their flexibility in handling large volumes of data without requiring prior domain knowledge. Although ML models can offer high predictive accuracy, they may lack explanatory

power, thereby impeding a comprehensive understanding of the phenomena under investigation. Incorporating interpretability can address this limitation by elucidating the relationships among variables, consequently facilitating hypothesis generation for subsequent research.

2.2.5 Explanations scope

In addition to the types of model explanation techniques and the nature of the data, explainers can be categorized based on their scope. Local explainers refer to individual predictions, while global explainers quantify the average behavior of a model. Specifically, explanations can be further categorized into feature effects, which are commonly expressed through graphs, and feature importance, which provides score-based summary measures of the overall contribution of features.

Feature effects describe how the impact of a feature varies across its value range, using a simplified function derived from f, $g : Xs \rightarrow Y$, being Xs a set of features to be explained with a size typically of 1 or 2 features. Examples of global feature effects are Marginal Effects (ME)(LONG; MUSTILLO, 2021; MIZE; DOAN; LONG, 2019), ALE plots (APLEY; ZHU, 2020), PD plots (FRIEDMAN, 2001) and SHAP (LUNDBERG; LEE, 2017).

Score-based explanations, often referred to as feature importance, essentially provide a ranking of features based on how much each one decreases the model's prediction error. The most commonly used methods are the tree-based MDI (BREIMAN, 2001a) and the model-agnostic PFI(FISHER; RUDIN; DOMINICI, 2018) and its variations (MOLNAR et al., 2023; STROBL et al., 2008). Additionally, there are score-based versions of some feature effects techniques such as PD (GREENWELL; BOEHMKE; MCCARTHY, 2018), ME (LONG; LONG, 1997) and SHAP (LEE et al., 2023).

2.2.6 Can XAI really obtain knowledge about the world?

Discussing the cability of XAI to extract knowledge from the world is essential, given that the core argument of this thesis hinges on XAI being an invaluable tool for deriving trustworthy insights. This view is in line with the growing trend among researchers towards more transparent AI and ML models, as a response to their increasing integration in society (ARYA et al., 2019). These researchers advocate that the empirical use of ML could pivot scientific research towards a theory-independent method, allowing data to convey its own

story without pre-existing hypotheses about the data, as noted in various studies (KITCHIN, 2014; ANDERSON, 2008; NAIMI; WESTREICH, 2014; ANDREWS, 2023; LIEBERSON; HORWICH, 2008).

While models known for their inherent transparency have faced minimal criticism, post-hoc techniques encounter more scrutiny despite their widespread use across various fields such as education (LEZHNINA; KISMIHÓK, 2022; MARTÍNEZ-ABAD; GAMAZO; RODRÍGUEZ-CONDE, 2020), healthcare (JAUHIAINEN et al., 2021; STIGLIC et al., 2020), social science (BERGER, 2023; BELLANTUONO et al., 2023), and sociology (LI et al., 2023; FAN et al., 2023).

The primary critique stems from the potential mismatch between what the opaque model is doing and what the post-hoc model attempts to explain (RUDIN, 2019; MULLAINATHAN; SPIESS, 2017; BABIC et al., 2021). On the other hand, (SULLIVAN, 2022) argues that gaining real-world knowledge with ML models is feasible as long as the link between model and phenomenon uncertainty can be assessed. Similarly, (CICHY; KAISER, 2019) and (ZEDNIK, 2021) suggest that XAI can aid in understanding the real world, but they remain vague about how the model and phenomenon are connected.

Through the lens of philosophy of science and epistemology, authors in (FLEISHER, 2022) draw parallels between XAI and the fundamental concepts of understanding. While there is some disagreement in the field, there is consensus that understanding is not an all-or-nothing state. Genuine understanding comes in degrees and can accommodate some inaccuracy and falsehood. In other words, understanding can still be valid even if the information or concepts it's based on are not entirely accurate. This ties into the concept of *idealization* in scientific models, which refers to the process of simplifying or abstracting certain aspects of a phenomenon or model to make it more tractable (JEBEILE; KENNEDY, 2015). Building on this, (FLEISHER, 2022) argues that XAI research has a solid foundation in science and promising avenues.

Rudin and colleagues (RUDIN, 2019) advocate for the use of inherently interpretable models rather than combining opaque models with post-hoc techniques, especially in high-stakes decisions. Although this may initially seem like a criticism of post-hoc techniques, the critique centers around the inappropriate selection of models that are too complex without much performance improvement. Nevertheless, the use of inherent interpretable models does not prohibit the application of post-hoc methods. In fact, post-hoc methods are model-agnostic and can be applied to any model and can provide extra insights. As (MOLNAR, 2022) notes, inherent interpretable models should always be included in benchmarks.

In this context, it is posited that XAI, particularly via post-hoc techniques, has the potential to augment model interpretability. Such enhancement is achieved either by providing additional insights into inherently interpretable models or by shedding light on functions of otherwise opaque models. And while these insights may not perfectly mirror the target model, they can be crucial in developing new theories based on real-world data. These theories can then be explored further to advance science and knowledge. Therefore, results from ML explanations provide not an end goal, but the starting point for further analysis and conceptualization.

2.2.7 Model-agnostic global explainers

2.2.7.1 ME plots and scores

The Marginal Effects (ME), or analytical derivative, were initially defined in the traditional statistical literature. The marginal term here is derived from the econometrics discipline as the "additional" effect, which has a different mean from the rest of this thesis, where the marginal term is related to the probability distribution of an underlying feature.

The ME were established as an alternative to explain the coefficients of features in non-linear models, especially those entailing interactions that obscure the direct interpretation of coefficients. These more complex models lose their direct interpretation of coefficients, meaning that interpretation requires a first understanding of the details of the specified model(LEEPER, 2021; LONG; LONG, 1997). The ME are also useful to inform the variable contribution in the natural scale on Generalized Linear Models (GLM), which involve transformations of the linear predictor into other discrete outcomes, such as logistic regressions, where coefficients typically lack direct interpretability and do not align with the scale of interest.

The ME effects of a variable X_s are in the function of all other remaining variables X_c and represent for continuous variables the change in the probability when the X_s varies in small change, as defined:

$$ME(X_s) = \lim_{h \to 0} \frac{f(X_c|(X_s + h)) - f(X_c|X_s)}{h} \tag{2.3}$$

In practice, h is the value of X_s and the ME effects can be straightforwardly plotted over X_s . However, usually, summary measures are the main unit of interest, such as:

- Marginal Effect at the Mean (MEM) is simply the computation of the MEs around the

mean of the feature distribution. In practice, MEM is close to the AME if $f(X)$ is not too noisy.

- ME at the Representative Value (MER) is a simplification of MEM calculation for a value that could be an interesting operation point for the research domain. The marginal effect is calculated for each variable at a particular combination of X values. Thus, MER provides a means to understand and communicate model estimates at theoretically important combinations of feature values.

2.2.7.2 PD plots and scores

The Partial Dependence (PD) plots serve as a graphical representation that quantifies the effect of specific features on the predicted outcome within a supervised learning model while holding other variables constant (*ceteris paribus*). These plots offer insights into the average marginal contribution of a feature of interest X_s to the model's prediction, with the remaining features X_c held constant. By doing so, if the predictive model closes the real world, PDs allow a causal interpretation of the role of X_s in the model if data meets the independence assumption (ZHAO; HASTIE, 2021). The underlying function can be mathematically described as follows:

$$PD(X_s) = \frac{1}{n} \sum_{i=1}^{n} f(X_s = j, X_c) \tag{2.4}$$

where $f(X_s = j, X_c)$ represents the model's predicted output when the feature X_s is intervened upon to assume a specific value j, while the remaining features X_c are held their observed values in the dataset. The value j is drawn from the marginal distribution of X_s. To be plotted, j assumes values within a defined grid of the ordered X_s where 2.4 is computed. For categorical features, j assumes each category as a possible value.

A more specific method for estimating PD is utilizing Individual Conditional Expectation (ICE) curves. ICE curves (GOLDSTEIN et al., 2015) provide a distinct curve $PD(X_s)$ for each individual data point i in the sample. Essentially, the PD is computed as the average of these ICE curves. This granular decomposition facilitated by ICE allows for identifying potential interaction effects between X_s and the remaining features X_c at global level, which may not be observed when solely relying on PDs.

In an attempt to yield scores from PD, (GREENWELL; BOEHMKE; MCCARTHY, 2018) proposed a simple score considering that a feature's importance is inversely related to the flatness of its PD Plot; a flatter PD plot suggests lesser importance, while greater variation in the PD indicates higher significance. Ass PD ignores feature relationships, this PD-based score captures only the main effect of the feature and ignores potential feature interactions

2.2.7.3 ALE plots

The Accumulated Local Effects (ALE) technique was established as an additional alternative to illustrate the feature effects. Distinct from prior methods, ALE focuses on variations in predictions rather than the predictions themselves, thereby isolating individual feature effects. Also, ALE is computed by parts of the data in an attempt to keep adherent to the data relationships without extrapolating.

In the ALE framework, intervals are theoretically defined as in ME, using derivatives, but in practice, ALE employs a grid Z based on quantiles of the feature of interest X_s . This process involves computing the effect of X_s separately for each quantile z intervening on X_s twice: assuming the lower and upper quantile limits, while keeping all other variables, X_c, constant. The essence of ALE lies in adjusting X_s for all observations between these two bounds and calculating the change in the prediction function. This method seeks to capture the local effect (LE) of X_s within the confines of each quantile, effectively isolating its influence by comparing the outcomes when data interventions are applied at the quantile's lower and upper limits.

Assuming data independence and a linear effect of X_s within the quantile interval, the average differences in the predictions between the maximum and minimum interventions represent the isolated local effect of X_s, which is computed as:

$$LE(X_s, z) = \frac{1}{n}\sum_{i=1}^{n} f(X_s = X_s^{\mathsf{max(z)}}, X_c) - f(X_s = X_s^{\mathsf{min(z)}}, X_c) \tag{2.5}$$

For visualization, this *LE* is subsequently accumulated over the grid Z. Theoretically, this accumulation is accomplished by integrating the expectations over intervals defined by the derivatives of the variable of interest. In practice, however, it can be estimated via summation across grid Z as per Equation 2.5. Notably, this equation is also centered, ensuring that the

average of $ALE(X_s)$ is zero with respect to the marginal distribution of X_s.

$$\begin{aligned} ALE(X_s) &= \sum_{z \in Z, z} LE(X_s, z) \\ &= ALE(X_s, z) - \frac{1}{n} \sum_{i=1}^{n} LE(X_s, z) \end{aligned} \tag{2.6}$$

Estimating interaction effects requires a modification to the equations. For instance, in the case of two interactions (second-order), the intervals in 2.5 have to change to rectangular regions. Additionally, in 2.6, second-order ALE require double-centering concerning both variables involved in the interaction (see details in (APLEY; ZHU, 2020)). Importantly, ALE is not inherently suitable for analyzing categorical variables that lack ordinality.

2.2.8 ALE decomposition

In linear models, the predictive function is a sum of the components that can be treated individually, the intercept, and the weight of each feature included in the function (2.1). The same can be applied to any high-dimensional function that can be decomposed into a sum of components of increasing dimensionality. In the following equation from (MOLNAR; CASALICCHIO; BISCHL, 2019), the predictive function is expressed as a sum of the intercept, individual (first order) feature effects, and interactions (second and higher order) effects.

$$f(X) = \overbrace{f_0}^{\text{Intercept}} + \overbrace{\sum_{j=1}^{p} f_j(x_j)}^{\text{1st order effects}} + \overbrace{\sum_{j<k}^{p} f_{jk}(x_j, x_k)}^{\text{2nd order effects}} + \ldots + \overbrace{f_{1,\ldots,p}(x_1, \ldots, x_p)}^{\text{p-th order effect}} \tag{2.7}$$

Unlike other techniques, ALE allows the function decomposition as unique components (APLEY; ZHU, 2020). The ALE components are computed conditional on the values of intervals and over the marginal distribution of all other features. This orthogonality-like property- called pseudo-orthogonality by the ALE author, ensures that the main effects can indicate how each feature affects the prediction, independent of the values of the other feature. The interaction effect indicates the joint effect of the features, not considering the main effects of related features.

2.2.8.1 Global SHAP explanations

The aim of SHapley Additive ExPlanation (SHAP) is to clarify individual model predictions by quantifying the contribution of each feature via the Shapley Values (SV) (SHAPLEY; others, 1953). Initially designed for local interpretability, SHAP can be seamlessly adapted for global model explanation by aggregating individual feature contributions (LUNDBERG et al., 2020). The most used SHAP-based score to indicate feature contribution is the absolute average of SV for each feature. Owing to its robust theoretical foundation and extensive implementations as software libraries, SHAP has emerged as a predominant technique in both industrial applications and academic research (BHATT et al., 2020).

The SV derives from coalitional game theory, where each feature value assumes the role of a player in a cooperative game, and the model prediction represents the total value or payout of the coalition. The SV is designed to allocate this collective payout equitably among the contributing features. Specifically, the SV $\phi_v(i)$ for a player i with a characteristic function v is computed as follows:

$$\phi_v(i) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!\,(|N| - |S| - 1)!}{|N|!} \left(v\left(S \cup \{i\}\right) - v(S)\right) \tag{2.8}$$

Where the summation iterates over all possible coalitions S that exclude the player i, thereby calculating the average additional contribution of player i across all these coalitions. The term $|S|$ denotes the cardinality of coalition S while $|N|$ indicates the cardinality of the complete set of players N.

The fraction $\frac{|S|!(|N|-|S|-1)!}{|N|!}$ function as the weighting factor for each coalition S. It quantifies the number of ways to form S and then adds i relative to the total number of ways to form any coalition, including i.

Finally, the term$(v\left(S \cup \{i\}\right) - v(S))$ calculates the additional contribution of player i to coalition S.

When accounting for all possible coalitions, SHAP assumes feature independence and integrates over the marginal distribution, akin to PD Plots and other permutation-based explainability techniques. Consequently, this approach introduces the issue of extrapolation.

SHAP inherits axiomatic properties from SV, namely Efficiency, Symmetry, Dummy, and Additivity.

- **Efficiency**, also termed as local accuracy in the SHAP context (LUNDBERG; LEE, 2017),

stipulates that the sum of SV for all features must equate to the total predictive value generated by the coalition of all features.

- **Dummy** axiom pertains to features that do not affect the model's prediction; such features are allocated a zero SV, reflecting their lack of contribution.
- **Additivity** or **Linearity** property is particularly relevant in post-hoc interpretability settings. It posits that the total attribution of a feature is the summation of all SVs associated with that feature across different models or scenarios.

In addition to these inherited properties, SHAP introduces unique attributes:

- **Missingness** is designed to uphold the Efficiency property during the SHAP computation, especially when data may be incomplete or missing.
- **Consistency** ensures that the attribution of a feature changes in correlation with its SV. If a feature becomes more important, its attribution should increase correspondingly, and vice versa.

Consequently, SHAP can be represented as an additive feature attribution method.

$$g(z') = \phi_0 + \sum_{v=1}^{M} \phi'_v \tag{2.9}$$

Where g is the explanation model, $z' \in \{0,1\}^M$ is the number of simplified input features - a binary vector indicates the presence or absence of a given feature within the coalition S.

2.2.8.2 PFI scores

The Permutation Feature Importance (PFI) is a model-agnostic metric used to evaluate the contribution of each feature to the predictive power of a trained ML model, f. Given a feature matrix X and a target vector Y, the PFI for a particular feature is calculated by measuring the increase in a specified error measure $L(Y, f)$ when the values of that feature are randomly permuted.

Let $f : X \to Y$ be the trained model, where $X \in \mathbb{R}^{n \times p}$ is the feature matrix with n samples and p features, and Y is the target space. The error measure $L(Y, f)$ quantifies the discrepancy between the predicted and true target values. The PFI of a given feature x_i is defined as follows:

$$\text{PFI}(x_i) = E\left[L(Y, f(X)) - L\left(Y, f(x_{\text{-}i,\text{perm}})\right)\right] \tag{2.10}$$

Here, $x_{\text{-}i,\text{perm}}$ denotes the feature matrix X where the i-th feature column has been permuted randomly. The expectation $E[\cdot]$ is taken over multiple permutations to obtain a stable estimate.

A higher PFI value for a feature indicates a greater contribution to the model's predictive capability. Conversely, a low or negative PFI suggests that the feature may be irrelevant or even detrimental to the model's performance. Usually, the PFI values are normalized to be ranked. Typically, PFI values are normalized and sorted such that they sum to one, to facilitate comparative ranking among the features.

2.2.8.3 MDI scores

The Mean Decrease in Impurity (MDI) is a metric specifically designed for assessing feature importance in tree-based models like Random Forests and Gradient Boosting Trees. As PFI, MDI is a loss-based metric and measures the average reduction in impurity—typically Gini impurity, entropy, or mean squared error—that a feature brings about when used for splitting in the decision trees that constitute the model.

For a given feature x_i, its $\text{MDI}(x_i)$ is defined as:

$$\text{MDI}(x_i) = \frac{1}{T}\sum_{t=1}^{T} \Delta I(t, x_i) \tag{2.11}$$

where T is the total number of trees in the ensemble, and $\Delta I(t, x_i)$ is the reduction in impurity in tree t attributable to feature X_i.

The impurity reduction $\Delta I(t, x_i)$ for a specific tree t and feature x_i is given by:

$$\Delta I(t, x_i) = \sum_{n \in \text{Nodes}(t, x_i)} w_n \Delta I_n \tag{2.12}$$

where $\text{Nodes}(t, x_i)$ is the set of nodes that use x_i for splitting in tree t, w_n is the proportion of samples reaching node n, and ΔI_n is the impurity reduction achieved by the split at node n.

As in PFI, the MDI values are often normalized and sorted to provide a ranking of feature importances.

2.3 EDM

In recent times, a rapidly expanding body of ML literature has emerged, introducing a diverse range of new tools, including algorithms, data preprocessing techniques, frameworks, and model validation methods. These tools have been developed to provide support for empirical researchers who utilize data to address a variety of problems (ATHEY; IMBENS, 2019). When these tools are specifically tailored for use with educational data, they serve as the foundation of a growing research area known as Education Data Mining (EDM). EDM, as described by (ROMERO; VENTURA, 2020), represents an interdisciplinary field dedicated to the analysis of extensive and complex educational datasets, with the goal of building predictions and extracting actionable insights and knowledge to support decision-makers in the realm of education.

The application of EDM extends across multiple domains within the educational sector. Much of the research in this domain has been centered on data derived from learning management systems within specific educational institutions (FISCHER et al., 2020), mainly in universities (ROMERO; VENTURA, 2020). Research studies in EDM encompass various subfields, including the investigation of cognitive strategies (FANCSALI et al., 2018; MOUSSAVI; GOBERT; PEDRO, 2016), prediction of student dropout (CHATURAPRUEK et al., 2018; JAYAPRAKASH et al., 2014), and the development of intelligent tutoring systems (JIANG; PARDOS; WEI, 2019). A common thread in all these areas is the task of predicting student performance, a task that, despite significant enhancements facilitated by advanced ML algorithms, still requires further advancements in providing explanations for the underlying factors driving these predictions (YANG; WANG, 2021; KOVALEV; KOLODENKOVA; MUNTYAN, 2020).

The provision of such explanatory insights can be critical, as presenting probabilities may prove inadequate for enhancing educational systems. For example, in automating an admission system with an EDM solution, fully understanding the factors behind these probabilities can improve the admission process. It provides the committee with important information to increase fairness and transparency (ALGHAMDI et al., 2020; MAULANA et al., 2023). Also, other processes like loan grants can benefit from these explanations (MAULANA et al., 2023).

Additionally, there are EDM applications specifically designed to extract insights from data, making explanations a key objective. For example, EDM has been effective in processing large datasets generated by modern LSA tests in the field of Educational Assessment (LIU; RUIZ, 2008; FILHO et al., 2023; FILHO; BRITO; ADEODATO, 2023b; SAARELA et al., 2016). In these

cases, the main goal is to discover knowledge about educational systems, providing crucial insights that support discussions on educational policies and guide the generation of novel hypotheses for subsequent confirmatory work.

In the domain of supervised learning, regression and classification tasks are commonly employed techniques (ALDOWAH; AL-SAMARRAIE; FAUZY, 2019). These tasks utilize algorithms that are universally recognized across various data mining fields, including Support Vector Machines, Decision Trees and their variations, Logistic Regression, and Neural Networks. While there is no consensus on the single most utilized algorithm, tree-based algorithms—especially Random Forest—emerge as a clear preference (RASTROLLO-GUERRERO; GÓMEZ-PULIDO; DURÁN-DOMÍNGUEZ, 2020; KHAN; GHOSH, 2021; MARTÍNEZ-ABAD; GAMAZO; RODRÍGUEZ-CONDE, 2020; NAMOUN; ALSHANQITI, 2020). This preference may be attributed to their widespread implementation across ML tools and the straightforward manner in which they allow for the interpretation of feature importance via MDI scores

2.4 XAI ON EDM

Despite the critical relevance of explanations from predictive models, the literature on interpretability in EDM is relatively sparse. A recent review focusing on the prediction of student performance reveals that most studies have neglected to provide explanations when using non-transparent predictive models (CHITTI; CHITTI; JAYABALAN, 2020). This lack of attention to explanations was also documented in (LIVIERIS et al., 2023).

Among the studies that emphasizing, the approaches vary. Most studies adopt pre-existing off-the-shelf tools to provide explanations, while others incorporate these tools as integral components for intervention purposes (MU; ANDREAJETTEN; BRUNSKILL, 2020; AFZAAL et al., 2021). For local explanations, widely employed tools include SHAP (LIVIERIS et al., 2023; CHIU, 2020; DOEWES; PECHENIZKIY, 2020; OLIVEIRA et al., 2023) and LIME (LIVIERIS et al., 2023; MATETIC, 2019; ZABRISKIE et al., 2019; HASIB et al., 2022; CHEN et al., 2022). In contrast, score-based metrics emerge as the predominant choice for global explanations. These metrics serve the dual objectives of identifying significant features for interventions and facilitating feature selection. Among the most frequently utilized measures are the MDI of tree-based models (CORTEZ; SILVA, 2008; ASHRAF; ANWER; KHAN, ; ZHAO et al., 2020) and absolute average SHAP (HOQ; BRUSILOVSKY; AKRAM, 2023; ROHANI et al., 2023). Additionally, methods such as PD plots (HONG; KIM; HONG, 2022; MASCI; JOHNES; AGASISTI, 2018) and SHAP (CHIU, 2020;

HOQ; BRUSILOVSKY; AKRAM, 2023) are commonly employed to visually illustrate the influence of individual features.

To the best of our knowledge, the ALE technique discussed and explored in this thesis as an alternative for explaining global feature contributions when data independence cannot be assumed, has not yet been widely employed in EDM. Only in (RANGONE; PIZARRO; MONTEJANO, 2022), where the authors presented a general framework for auto-ML, was ALE considered as one of the explainers that could be utilized in the interpretation step. However, specific details or examples of its use were not provided.

2.4.1 Educational Assessment discipline

Educational Assessment refers to the discipline that aims to systematically evaluate student learning, skills, and performance to understand and improve educational outcomes. It encompasses a range of methods, but since the 1950s, quantitative analysis has been the standard procedure (MALLINSON; NOAH; ECKSTEIN, 1969). Moreover, a predominant data source is derived from the LSA tests.

The LSAs are standardized tests that collect, beyond student performance, much other information about the educational context in which the students are involved in. The LSAs thought standardizing psychometrics methodologies such as Item Response Theory built a robust process for assessing the capabilities of students to learn what they were supposed to learn. Also, the periodicity of these tests enables temporal comparisons by observing the different paths that collected information might take across educational systems over time, making them a topic of interest among educators, researchers, and policymakers (JOHANSSON, 2016; KAPLAN; HUANG, 2021).

Educational achievement is a multifactorial construct, influenced by a complex interplay of closely intertwined variables (ABAD; LÓPEZ, 2017). Within the EDM paradigm, a typical straightforward application of LSA data is a prediction function $f(X) = Y$, by which the contextual information $X = (x_1, x_2, ..., x_n)$ collected by the LSA questionnaires and complemented by other sources are mapped to the LSA score Y, a measure of educational achievement. This approach aligns with the concept of the education production function (BOWLES, 1970; SCHEERENS, 1991), where input contextual variables lead to educational outcomes.

In this framework, EDM can easily adapt to the large volume of data derived from modern LSAs to optimize scientific discovery and enhance the debate surrounding practices in the

field of education (GABRIEL; SIGNOLET; WESTWELL, 2018; GOMES; HIRATA; OLIVEIRA, 2020; MARTÍNEZ-ABAD; GAMAZO; RODRÍGUEZ-CONDE, 2020; LEZHNINA; KISMIHÓK, 2022; MARTÍNEZ-ABAD; GAMAZO; RODRÍGUEZ-CONDE, 2018). To identify and characterize the influence and interactions of factors related to educational achievement, high-performing supervised ML models and XAI can derive meaningful insights by understanding how f uses X to predict Y (CHEN; ZHANG; HU, 2021; DONG; HU, 2019; GOROSTIAGA; ROJO-ÁLVAREZ, 2016; HU; DONG; PENG, 2022; MARTÍNEZ-ABAD; GAMAZO; RODRÍGUEZ-CONDE, 2020)

Although inherent interpretable models are often used in this direction (ADEODATO, 2016; MARTÍNEZ-ABAD; GAMAZO; RODRÍGUEZ-CONDE, 2020) (GAMAZO; MARTÍNEZ-ABAD, 2020), many other scholars rely on explanations derived from opaque models. The tree-based algorithms such as random forest and gradient boosting are the most commonly used (GAMAZO; MARTÍNEZ-ABAD, 2020) algorithm, and many scholars have relied on their intrinsically derived feature importance (MAIA; BUENO; SATO, 2021; REBAI; YAHIA; ESSID, 2020; MASCI; JOHNES; AGASISTI, 2018; LEZHNINA; KISMIHÓK, 2022; CHANG; CHEN, 2018; MARTÍNEZ-ABAD, 2019; GABRIEL; SIGNOLET; WESTWELL, 2018; RODRIGUES et al., 2021; DEPREN; AŞKıN; ÖZ, 2017; HU; DONG; PENG, 2022; CHEN; ZHANG; HU, 2021). The preference for the use of scores to characterize the relevance of features is also aligned with traditional studies that rely on coefficients of additive models. However, there is a growing trend towards employing post-hoc feature effects techniques, such as PD plots and SHAP values, for a more detailed understanding of feature influences (LEZHNINA; KISMIHÓK, 2022; MASCI; JOHNES; AGASISTI, 2018; REBAI; YAHIA; ESSID, 2020; SCHILTZ et al., 2018).

3 MEASURING ERROR ON FEATURE EFFECTS

This chapter is dedicated to benchmarking the performance of various global explainable feature effect techniques under diverse data dependency scenarios, directly addressing RQ1. In addition to comparing these techniques, this work contributes to the existing literature by introducing a novel metric. This metric quantifies the degree of deviation of the explained feature effects from their true values. The introductory section establishes the significance and relevance of this chapter's contributions within the broader context of existing literature. Key concepts related to the contributions are defined, followed by a description of the benchmarking methodology employed. The chapter then presents and summarizes the experimental results, highlighting their role in addressing RQ1 and their impact on the field of XAI

> RQ1 - How do widely used feature effects techniques compare with ALE in accurately identifying true feature effects considering different inter-data dependencies?

3.1 INTRODUCTION

The global model explainability can be treated as a problem that entails the process of discerning, on average, how alterations in an input variable influence the model's predictions. In the case of linear models, the constant effects of features allow the attribution of feature contribution to be easily quantified using point estimates and variances of the model's parameters, which are essentially the estimated coefficient scores (HASTIE ROBERT TIBSHIRANI, 2014). Conversely, ML models, which may encapsulate non-linear relationships, necessitate a more nuanced understanding of the intricate associations between the variables of interest and the target variable. While scores are still paramount to reporting the relevance of features in ML for many tasks, more detailed visual representations to illustrate the behavior of the relationships between the feature of interest and the target can produce better communication about the whole feature behavior.

Several propositions have been made to alleviate the independence assumptions of some model feature effects explainers. These propositions often rely on the use of conditional instead of marginal distribution of features to avoid extrapolating the confinements of data relationships. Despite these advancements, the literature still lacks comprehensive analyses regarding

the extent to which different strategies to handle data relationships affect explanations of feature effects (MOLNAR et al., 2022). Prior works have predominantly relied on theoretical discourse or visual demonstrations (APLEY; ZHU, 2020; GKOLEMIS et al., 2022; GKOLEMIS et al., 2023; MANGALATHU et al., 2022; BAKHSHI; AHMED, 2021), akin to the illustration in Figures 3 and 4, to elucidate how explanations diverge when the actual data relationships are preserved during the computation of explanations and when not. Introducing a quantitative aspect would enhance the flexibility of this comparison framework, making it adaptable for future benchmarks.

In a quantitative comparison akin to the current study, (MOLNAR et al., 2023) evaluated model fidelity. The model fidelity concerns the difference between the predictions of the ML model and the explanation method. The authors capture the overall difference in the model prediction and the prediction of the partial function when employed by the PD plots and when deployed by ALE plots. These differences were next averaged across all data points and features. The author found similar results when comparing model fidelity to PD and ALE.

Other work (GKOLEMIS et al., 2022) compared the accuracy of ALE and a version of a more computationally efficient ALE to recover the known true feature effects. The authors use the Normalized Mean Squared Error (NMSE) to compare the computed partial functions and the true feature effects. The NMSE is based on the expected value of the computed functions divided by their variance.

Distinct from (MOLNAR et al., 2023), this study will assess the computed feature effects against the theoretically defined true feature effects in controlled experiments where the data generating functions are known. This is important since even a predictive model with explanations closely aligned to the model could, paradoxically, deviate significantly from the actual data-generating process (FISHER; RUDIN; DOMINICI, 2018; SLACK et al., 2020). Such a comparison offers a more robust methodology for determining whether the model's explanations are more consistent with the actual data or the model's intrinsic structure.

Moreover, different from both (MOLNAR et al., 2023) and (GKOLEMIS et al., 2022), this study will individually evaluate the entire range of the explained feature, potentially uncovering a more nuanced understanding of discrepancies between explainable models. Unlike NMSE from (GKOLEMIS et al., 2022), which disproportionately emphasizes larger errors, the Absolute Difference Between Explanations (ABX) proposed on this chapter, which is described bellow, uniformly accounts for all deviations across the feature's range, providing a more balanced evaluation and a more realistic assessment of a technique's accuracy. Finally, to the best of

our knowledge, this study is the first to incorporate SHAP global explanations into such a benchmark.

3.1.1 The extrapolation problem

One commonly utilized framework in the field of XAI is based on posthoc analysis and frequently involves the processes of sampling a subset of data, intervening on the data, getting model predictions using the fitted model, and subsequently aggregating them to quantify changes in outputs and produce model explanations (SCHOLBECK et al., 2020). While there are numerous variations in how each step is executed, a critical step within this framework is the intervention step.

Interventions involve altering the values of the features of interests on the *ceteris paribus* reasoning. When the interventions are beyond the confines of the actual conditional distribution, it becomes possible to generate unlikely data points. Such circumstances compel the model to make predictions in regions where it was not trained, potentially yielding unexpected results. Consequently, changes in the model's outputs can lead to unrealistic model explanations regarding the true data-generating process. This issue would not be a concern if the goal were to understand the function's behavior itself, but it certainly presents difficulties when the aim is to uncover the potential data-generating process.

XAI techniques that modify features based on their values concerning the entire dataset pose a significant challenge when explaining non-additive functions. These model-agnostic techniques inherently assume feature independence and intervene by using the marginal distribution of a feature, which can lead to misinterpretations when this intervention extrapolates outside of the training data's scope. The root of this issue lies not in the predictive models themselves but rather in the assumptions that these interpretive techniques make about the underlying data. In a hypothetical situation where it is possible to assume that the feature of interest is independent of others, extrapolation would not be a concern. Otherwise, the model's outputs could even be interpreted as the causal effect of the potential intervention of X on y (ZHAO; HASTIE, 2021). However, this is often not the case in real-world scenarios.

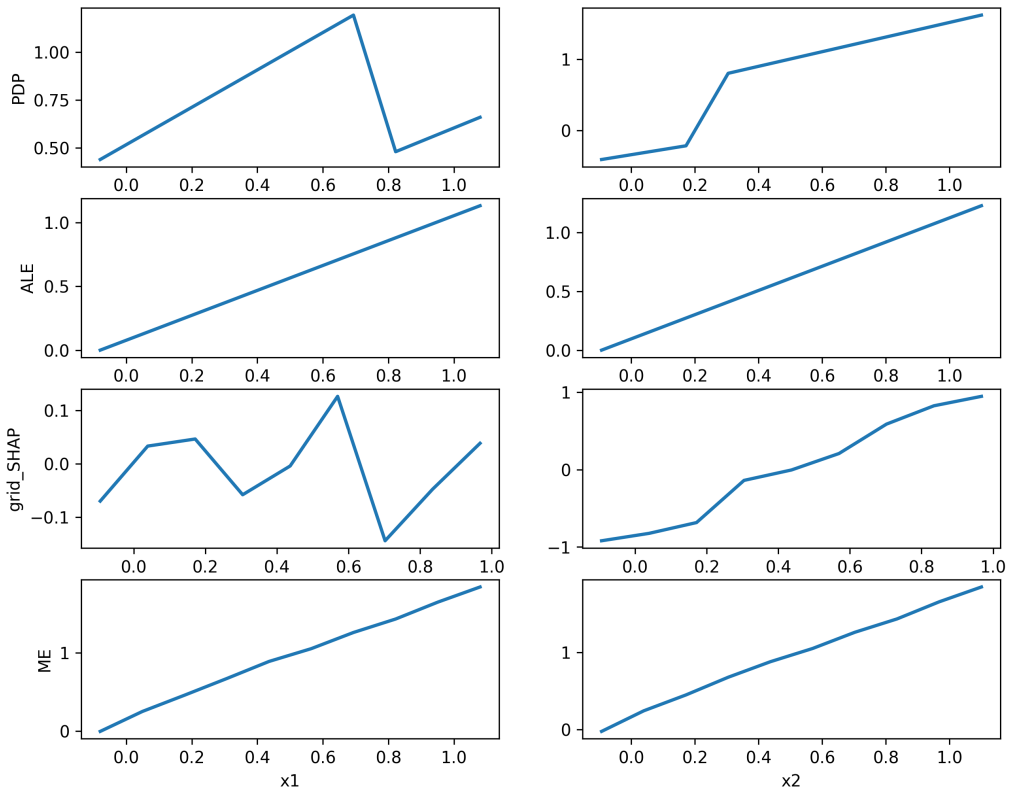
Based on (MOLNAR, 2023), Figure 3 illustrates a simulation of this issue. The figure provides explanations for the function f for each variable in $X = x_1, x_2$ regarding their predictive role in Y involving four different techniques: PD plots and SHAP, which tend to extrapolate by utilizing the marginal distribution of the feature of interest for intervention, and ALE and ME

plots, which do not.

To simulate the unexpected behavior of f when predicting outside of the training data envelope - thus violating conditional relationships accessing regions of the input space not covered by the training data - f was artificially constrained to make incorrect predictions within this region. Specifically, it was programmed to consistently predict a constant value in a region of the data distribution that does not exist, illustrating a potential problem when the function is used to predict beyond the bounds of its training data. The predictive function was adjusted to make predictions as follows:

$$\mathrm{f}(\mathrm{x}_1, x_2) = \begin{cases} 2 & \text{if } x_1 > 0.7 \text{ and } x_2 < 0.3 \\ x_1 + x_2 & \text{otherwise} \end{cases}$$

Figure 3 – Explanation of the customized linear regression model for predicting a constant in a region beyond the training data bounds. The dataset was generated using the function $f(x_1, x_2) = Y = x_1 + x_2$



Source: self-provided

Notably, PD plot and SHAP (for x_1) demonstrate greater sensitivity to the artificial bias introduced in the model. Conversely, ME and ALE plots remain robust in this same scenario,

consistently revealing the true linear effects of both x_1 and x_2 as defined in the function f. In other words, PD plot and closely align with the model (predict a constant in the region where $x_1 > 0.7$ and $x_2 < 0.3$) but move away from the data-generating function. While this isn't problematic per se, as the PD plot and SHAP approximate the model's behavior, they can pose challenges when they are used to explain the roles of highly correlated features due to their susceptibility to extrapolation.

3.1.2 Importance of interventional distribution

While preserving the conditional distribution seems to be the right way to explain the global behavior of dependent features, there is another important aspect: the use of interventional instead of observational expectation. Just keeping the conditional distribution without using an unrealistic combination of data points is not enough to recover the right effects of dependent variables. The need for interventional expectation in model-agnostic explainable techniques instead of observational has been already formally discussed in (JANZING; MINORICS; BLÖBAUM, 2020) as a recommendation to researchers who intend to extend the SHAP technique.

The concept of interventional expectation was first introduced in the causality literature (PEARL, 1993) aiming to elucidate the effect of manipulating a specific feature within a hypothetical scenario. In this context, the manipulation entails substituting the value ofx_1 for that feature, while holding the values of all other features constant. Specifically, given $X = x_1, x_2, ..., x_n$ a function $f(X)$ that predict Y, the interventional expectation regarding x_1 is defined by using the "do-operator" through $E[Y|do(X_1 = x_1)]$. The "do-operator" defined by Pearl allows researchers to formalize and analyze the causal effect of setting a variable X to a particular value x, effectively simulating an intervention in the system.

In the traditional statistical literature, the use of ME plots have long been proposed as one of the reliable way to interpret models in place of the coefficients (LONG; LONG, 1997). The ME corresponds to the average differences in the outcome when the features partially change from one specified value to another. The ME uses the observed conditional distribution and does not extrapolate the joint distribution present in the data. However, ME is not able to differentiate the effects of correlated variables as the changes in the outputs are computed while changing all dependent variables in tandem using the observed distribution.

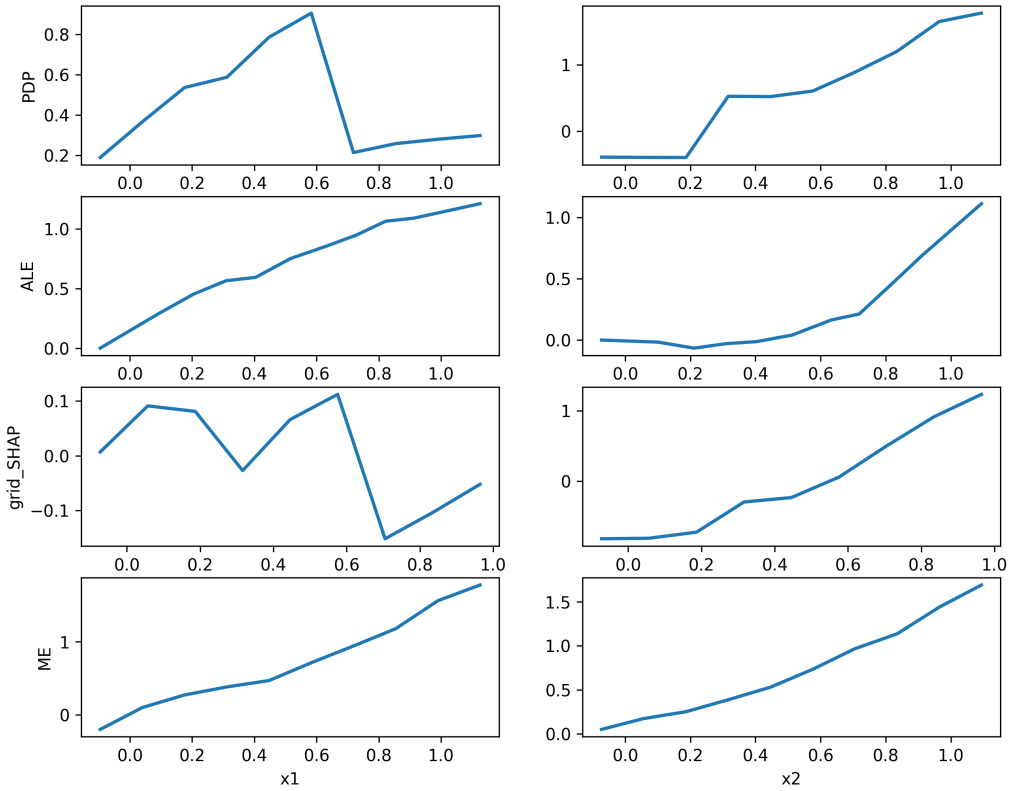
To illustrate how the use of observed distribution in a highly correlated scenario fails to recover the individual role of features, Figure 4 shows the same issue as Figure 3, where

the predictive function has unexpected results when predicting outside of the training data envelope. Now, the output Y is defined by $f(x) = x_1 + x_2^3$ and the data was fitted by the non-linear random forest algorithm, which might be able to detect the new cubic effect of x_2. In this case, differently from ALE plot that correctly traces a linear effect to x_1 and an exponential effect to x_2 , ME recovered the same effects for both variables.

The tension between the use of observational and interventional distribution has also been discussed around the implementation of SHAP. Some authors argue that the use of observed expectation can attribute importance to irrelevant features (JANZING; MINORICS; BLÖBAUM, 2020; SUNDARARAJAN; NAJMI, 2020) while using interventional can lead to extrapolation issues. In (CHEN et al., 2020), the authors suggest that there is no correct choice for this value function. Instead, the crux of the interpretation hinges on whether the aim is fidelity to the model or alignment with the data.

Within this context, ALE technique emerges as a suitable tool for global explanations of feature effects to real applications, where variables are not independent. ALE has a good balance of the tradeoff of being true to the model and true to the data, as it uses the interventional conditional expectation. In other words, ALE takes advantage of interventions to break variable dependencies while adhering to the data joint distribution when computing effects by parts of the data.

Figure 4 – Explanation of the customized random forest model for predicting a constant in a region beyond the training data bounds. The dataset was generated using the function $y = x_1 + x_2^3$



Source: self-provided

3.2 METHODS

This section present the methods used tom compare ALE plots with existing techniques to report global features explanations addressing RQ1. Initially, we introduce a novel metric designed to quantify the deviation of a technique's explanations from the actual effects when elucidating a feature within a dataset characterized by a known data-generating process. As benchmarks, we select the most widely utilized methods identified in the literature review presented in Section 2, specifically: ME, SHAP, and PD plots. Unlike ALE, ME, and PD, which provide global explanations, SHAP is primarily focused on local interpretations. However, SHAP explanations can be aggregated in various ways to derive global insights (LUNDBERG et al., 2020). For comparison purposes, a version comparable to the others, termed Grid_SHAP, was defined. The remaining techniques were implemented in accordance with their respective formal definitions (see Section 2.2.8.1) without centering them around the mean.

3.2.1 An alternative for global SHAP

The Grid_SHAP was defined as a partial-dependence-based alternative to ALE for straightforward comparison with other techniques. Grid_SHAP aggregates the SHAP values (SVs) within equally spaced intervals in terms of the data distribution. While the original SVs are centered around the mean, allowing the interpretation of the effects of a feature on a certain prediction for a specific instance from the base value, Grid_SHAP is uncentered, with the definition as follows:

$$V = \phi_0 + \phi_i(x) + E[\phi_i(x)] \tag{3.1}$$

where V is the uncentered SVs, x_i is the i variable of the dataset X, $\phi_i(x)$ the originally computed SVs for the feature x_i, ϕ_0 is the expected model output (baseline effect). $\phi_a dj$. And:

$$\bar{V_i}(q) = \frac{1}{|Q_{x_i}(q)|} \sum_{x \in Q_{x_i}(q)} V_i(x) \tag{3.2}$$

where, $\bar{V_i}(q)$ represents the calculated average Shapley value for feature i within quantile q. The expression $|Q_{x_i}(q)|$ indicates the size of quantile q for feature i, reflecting the number of data instances it comprises.

The $V_i(x)$ were computed using the *FastShap* package [1], a fast version of SHAP that uses monte carlo simulation to approximate SVs. This approach greatly facilitated the experimentation due to the high computational cost of computing the exact SVs in a model-agnostic manner.

3.2.2 Absolute Difference Between Explanations - ABX

Different from previous works, which qualitatively compare the robustness of global feature effects techniques, we take advantage of the theoretical effects of variables in synthetic data and compute the Absolute difference Between Explanations(ABX). The ABX is motivated due to the need for a measure that captures the difference in the actual feature effects and the computed explanations across all feature ranges. So, whether an underlying explanation technique is robust in some parts of the data but fails drastically in others, ABX would allow a fair comparison with others, which comes close during the whole feature range. ABX measures the absolute value of the area between the baseline explanation (theoretical) and the explained effects based on data without respect if the explainer overestimates or underestimates the feature effects. Formally, ABX is defined as:

$$ABX = \int_{min}^{max} |\phi(x) - \hat{\phi}(x)| dx \tag{3.3}$$

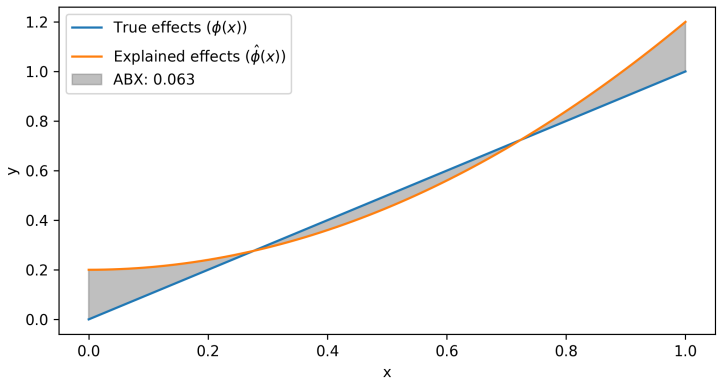
Visually and geometrically, the ABX statistic has a straightforward interpretation: it represents the absolute summation of the areas between the two curves over the feature explanation, as shown in Figure 5. Lower values of ABX signify a more robust technique, with zero being the minimum possible value, occurring when both functions completely overlap over the entire interval.

3.2.3 Experimental setup

Artificial data was used to provide a dataset where the true variable effects are known. Even in such a scenario, it is not guaranteed that the fitted function can recover the data-generating process, as discussed earlier in Section 2.2.6. To simplify the process for the predictive functions, simple scenarios were simulated. Specifically, four different datasets were created based on:

[1] https://cran.r-project.org/web/packages/fastshap/index.html

Figure 5 – A explanation plot which shows the theoretical true effects and the explained by a model-agnostic technique. The shaded regions between the two curves represents the ABX statistic



Source: self-provided

$y^i = f(X^i) = X_1^i \cdot X_2^i + \epsilon_i$, where $\epsilon_i \sim N(0, 0.01)$. The variables X_1 and X_2 are dependent from the same uniform distribution $N(0, 1)$ with the addition of a stochastic noise $N(0, 0.5)$.

- In the *independent* scenario y depends linearly only of X_1 and $f(X) = X_1 + \epsilon$
- In the *dependent linear scenario* y depends linearly of both X_1 and X_2 being $f(X) = X_1 + X_2\epsilon$
- In the *first dependent non-linear* scenario y depends linearly of X_1 and non-linearly of X_2 being $f(X) = X_1 + X_2^2\epsilon$
- In the *second dependent non-linear* scenario a more complex function was defined, and y depends .linearly on X_1 and hold a non-linear cubic polynomial relationship with X_2 being $y = x_1 + (x_2 - 0.9x_2^3) + \epsilon$

All experiments were conducted within the R programming environment, utilizing a 30-Monte Carlo simulation framework to ensure robust statistical analysis. As discussed in Chapter 2, one of the primary models selected for data fitting in EDM is the Random Foresst (RF), a choice motivated by its widespread acceptance and proven effectiveness in EDM tasks. In addition to RF, we explored another algorithmic class often used in EDM by incorporating a Neural Network (NN) model. Unlike the RF model, which is characterized by piecewise constant functions potentially leading to more noticeable changes in model output with variations in input variables, the NN model is based on differentiable functions, generally resulting in smoother transitions of output as input variables change. This contrast introduces greater di-

versity into our experimental design, allowing for a more comprehensive evaluation of post-hoc explanation techniques.

The NN was employed from the *nnet* package[2], and the RF from the *randomForest* package[3]. For each scenario, the number of sampled data points was varied with $N \in 200, 500, 1000$. The parameters for the NN algorithm—comprising ten nodes in the single hidden layer, a linear output activation function, and a regularization parameter of 0.0001—were determined to be approximately optimal through multiple iterations of 5-fold cross-validation for the first data scenario. The RF algorithm was executed with default parameters.

To calculate the ABX a numerical approximation was used. The approximation is based on the Trapezoidal Rule, implemented through the *pracma*[4] package in R. This approach involves linearly interpolating between data points to form an approximate representation of the curve. The area under this curve is then estimated by dividing it into trapezoidal segments and summing their respective areas.

Across all synthetic data scenarios, the average ABX from the 30-Monte Carlo process was adopted as the final measure. To investigate the influence of the number of quantiles on the results, metrics were computed by dividing the data into quantiles with $k = 10$ and $k = 50$, where k represents the number of equally distributed parts.

[2] <https://cran.r-project.org/web/packages/nnet/index.html>
[3] <https://cran.r-project.org/web/packages/randomForest/index.html>
[4] <https://cran.r-project.org/web/packages/pracma/index.html>

3.3 RESULTS

This chapter emphasized the empirical evaluation of the performance of the most used feature effects techniques: ALE, SHAP, ME, and PD plots. The primary objective is to discern the differences in the explanations provided by these techniques in terms of the global effects of variables compared to the true variable effects inherent in the data-generating process. The introduced ABX metric serves as the benchmark for this evaluation. Lower ABX values are preferable as they indicate a closer alignment of the explainable technique's output with the true variable effects along all the variable ranges.

Two distinct types of supervised models were considered: RFs and NNs, with hyperparameters optimized through a cross-validation process under various conditions. It is crucial to acknowledge that the conditions under which models are applied can inherently influence their outputs and, consequently, the interpretations derived from model explanation techniques. Nevertheless, we applied both models and explanation techniques consistently across these conditions to ensure that our evaluation remains unbiased. Moreover, in all tested conditions, both models demonstrated robust performance, with a Root Mean Square Deviation (RMSD) close to the standard deviation of the theoretical noise added to the target variable. This indicates that the models were effectively capturing underlying patterns in the data.

The results of the experiments are presented in Table 3 for x_1 and in Table 4 for x_2. Only values for k equal to 10 are presented, as there is no significant difference compared to the k equal to 50. Both models, NN and RF, performed well when measuring the RMSD. The RMSD values were found to be very close to the standard deviation of the artificial noise (0.1) added to the target variable. This suggests that, to a considerable extent, the models are effectively capturing the underlying relationship between the predictors $x1, x2$ and the response variable y .

Examining Table 3, it is evident that ALE technique outperforms the other techniques in all scenarios where data is dependent. In the hypothetical scenario of independent data - a condition that may diverge from real-world situations, where data typically exhibits some level of correlation - all techniques yield comparably satisfactory results for the RF model, with ME producing the best values. However, while ME most accurately captures the effects of variable 1 (x_1), it produces suboptimal results for variable 2 (x_2) (Table 4), which has no effect on y in this independent scenario. It is probable that some effect was assigned to the x_2 through x_1 even variables being independent of each other. This highlights the pitfalls of using the data

Table 3 – ABX statistic for the variable x_1

	ALE		grid_SHAP		ME		PD	
Scenario	NN	RF	NN	RF	NN	RF	NN	RF
independent								
N = 200	0.0784	0.0760	0.2613	0.0490	**0.0352**	**0.0383**	0.3297	0.0446
N = 500	0.0437	0.0702	0.1407	0.0351	**0.0242**	**0.0345**	0.1820	0.0346
N = 1000	0.0365	0.0605	0.1046	**0.0222**	**0.0212**	0.0274	0.1513	0.0274
linear								
N = 200	**0.0757**	**0.0818**	0.5398	0.5141	0.4904	0.4994	0.5250	0.5075
N = 500	**0.0465**	**0.0630**	0.5268	0.5052	0.4913	0.4973	0.5303	0.4961
N = 1000	**0.0471**	**0.0547**	0.5188	0.4944	0.4881	0.4792	0.4850	0.4994
non-linear 1								
N = 200	**0.0778**	**0.0735**	0.3918	0.3255	0.3241	0.3131	0.4757	0.3365
N = 500	**0.0482**	**0.0671**	0.3408	0.3344	0.3177	0.3154	0.3712	0.3300
N =1000	**0.0321**	**0.0500**	0.3748	0.3345	0.3201	0.3173	0.4411	0.3365
non-linear 2								
N =200	**0.0651**	**0.0860**	0.3657	0.2788	0.2695	0.2678	0.4064	0.2952
N = 500	**0.0552**	**0.0678**	0.3238	0.2777	0.2707	0.2685	0.3577	0.3002
N = 1000	**0.0410**	**0.0703**	0.2948	0.2818	0.2714	0.2720	0.3031	0.3116

Source: self-provided

joint distributions without interventions when explaining models under independent data.

In the same independent scenario, permutation-based techniques (PD and SHAP), which perform interventions (albeit at the cost of extrapolation), produce commendable results for the RF model but not for the NN model. Initially, this discrepancy may be attributed to the high flexibility of NNs (GRINSZTAJN; OYALLON; VAROQUAUX, 2022), which can potentially yield many functions consistent with the observed data but divergent from the true data-generating process. This flexibility can lead to a mismatch between explanations derived from the NN model and the actual data-generating process.

Upon closer examination, however, ALE and ME, which unlike ME plots and SHAP do not extrapolate the training data, exhibit comparably low ABX across both RF and NN models within the independent scenario. Consequently, the discrepancy from RF to NN in ME and PD outputs, indeed, suggests that the extrapolation can be problematic even in independent scenarios when explaining highly complex models such as NNs. NN may inadvertently yield unusual predictions outside of the training data, thus compromising their ability to capture

Table 4 – ABX statistic for the variable x_2

	ALE		grid_SHAP		ME		PD	
Scenario	NN	RF	NN	RF	NN	RF	NN	RF
independent								
N = 200	**0.0678**	**0.0729**	0.5576	0.5002	0.5031	0.5014	0.5886	0.5009
N = 500	**0.0476**	**0.0505**	0.5098	0.4926	0.4981	0.4927	0.5267	0.4933
N = 1000	**0.0278**	**0.0475**	0.5035	0.5002	0.5003	0.5019	0.5360	0.5005
linear								
N = 200	**0.0782**	**0.0726**	0.5351	0.5142	0.4962	0.5016	0.5036	0.5072
N = 500	**0.0471**	**0.0618**	0.5200	0.5042	0.4910	0.4910	0.5325	0.4987
N = 1000	**0.0505**	**0.0542**	0.5160	0.4948	0.4840	0.4814	0.4782	0.4983
non-linear 1								
N = 200	**0.0631**	**0.0885**	0.5166	0.4905	0.4837	0.4760	0.5885	0.5026
N = 500	**0.0367**	**0.0550**	0.5027	0.4947	0.4793	0.4805	0.5281	0.4957
N =1000	**0.0326**	**0.054**7	0.5118	0.4928	0.4809	0.4799	0.5905	0.5021
non-linear 2								
N =200	**0.0705**	**0.0732**	0.5435	0.5063	0.5111	0.5124	0.5747	0.5237
N = 500	**0.0480**	**0.0715**	0.5250	0.5056	0.5091	0.5070	0.5231	0.5264
N = 1000	**0.0449**	**0.0514**	0.5144	0.5088	0.5119	0.5125	0.5065	0.5397

Source: self-provided

the effect even from the independent features due to the model's complex behavior.

Conversely, step-wise algorithms, such as RF, tend to exhibit greater stability in their predictions owing to their construction from multiple decision trees, which individually handle variations in data in a more controlled manner.

The quantity of data points significantly affects the outcomes in the independent scenario. Generally, an increase in data points enhances the performance of all examined techniques within both algorithms. This trend is also observable in the dependent scenario for the ALE technique, which slightly improves ABX as the number of data points increases. Although a similar improvement can be observed for other techniques, it is not enough to decrease the ABX to levels comparable to those achieved by ALE technique.

The results for the ALE technique is even more favorable when considering the variable x_2 as shown in Table 4. In the independent scenario, all techniques failed to assign no effect to x_2 for both NN and RF models. In other scenarios, where the data are dependent, the ALE technique achieves better results, while the other techniques exhibit higher ABX values.

3.4 SUMMARY

This chapter explores the differences in explanations from various methods across different data dependencies. It presents a methodology designed to accurately measure the extent to which global explanatory methods can recover the true data-generating process. It presented a comprehensive benchmarking using artificial data, which embodies different generative processes across various scenarios to assess PD, ME, SHAP, and ALE plots. The methodology introduces ABX, a metric that measures the extent to which explained effects deviate from the theoretical feature effects.

The methodology has demonstrated that the ALE technique surpasses other techniques in feature explanation within datasets that resemble real-world conditions—namely, scenarios where variables are correlated. Specifically, in scenarios where the data-generating function is dependent on more than one variable (in that case x_1 and x_2) ALE achieves statistically superior results by closely approximating the true effects of features across their entire value range, in comparison to SHAP, ME, and PD plots. The experiments also show how the independence assumptions of explainers such as PD plots and SHAP can compromise the explanations of highly complex models like NNs, even in hypothetical scenarios where the data-generating function is truly independent.

The ABX metric, introduced in this chapter, provides a quantitative measure to quantify discrepancies in global feature effect explanations and establishes a foundation for future benchmarking efforts in the field of XAI.

This study highlights the importance of selecting the appropriate XAI technique based on the specific characteristics of the dataset in question. Specifically, ALE demonstrated paramount robustness in explaining feature effects when data is not independent. Therefore, providing empirical evidence that the techniques that either allow for extrapolation or do not use interventions can diminish the practical utility of their explanations.

4 ALE-BASED SCORE-EFFECTS SIZE

This Chapter presents novel metrics to highlight the relevance of features in supervised learning models, directly addressing the RQ2. Initially, the introduction section outlines the motivation and significance of the chapter's contribution. Additionally, it delineates pertinent literature to clearly demarcate the contributions within the existing research landscape. Subsequently, the chapter defines the scores formally and presents the experimental setup for validating them. The findings are then detailed, followed by a summary section that analyzes the main results and how they answer the RQ2.

> RQ2 - How effectively can score-based explanations derived from the ALE framework report individual and isolated attribution of the features in terms of their magnitude and direction compared to existing methods?

4.1 INTRODUCTION

While visualizations provide a more comprehensive understanding of feature effects, score-based explanations continue to be widely adopted in applied educational research (FILHO; BRITO; ADEODATO, 2023a). These explanations are particularly beneficial for feature selection processes, interpreting models with numerous features, and (WEI; LU; SONG, 2015; FILHO; ADEODATO; BRITO, 2021) describe interactions between more than two features (APLEY; ZHU, 2020)

In the educational domain, a standard score to interpret variable relevance is the coefficients of additive models. The coefficient represents the weight of each variable in the predictive function. Under inherent assumptions, the coefficients represent the extent and the direction of the role of the variable in the data-generating process. This enables educational practitioners to address questions regarding the average effect of variables within the predictive function.

The extrapolation problem was extensively discussed in Chapter 3 in the framework of feature effects. The extrapolation consists of using regions of the covariate space with little or no data and also is present in the context of score-based explanations. Specifically, within the context of PFI, the model's performance is evaluated in these data-sparse regions, which may produce scores that lack a strong link to the underlying data-generating process. Moreover, PFI often faces challenges in discerning the significance of individual features in datasets

characterized by substantial inter-variable dependencies. This complexity arises because the informational content of permuted features can persist via associations with other variables (STROBL et al., 2008; HOOKER; MENTCH; ZHOU, 2019)

In the SHAP framework, the absolute average is constantly used as a signal of feature relevance (SCAVUZZO et al., 2022). This strategy has also been adopted for other methods, such as PD plots (GREENWELL; BOEHMKE; MCCARTHY, 2018) and ME (LONG; MUSTILLO, 2021). As demonstrated in the previous chapter, all these techniques also have problems when data are dependent.

To address this issue, alternatives to PFI have been proposed. The principal advancement of these methods is to allow feature permutation within the conditional distribution instead of the marginal distribution.

In (CANDÈS et al., 2018; WATSON; WRIGHT, 2021), the researchers attempt to emulate the conditional distribution by leveraging 'knockoffs'—replicas of the original features that maintain the joint distribution while being independent of the outcome variable, conditioned on the other features. Nevertheless, the intricacy of this method lies its reliance on complex task - accurately creating knockoffs that faithfully mirror the dataset's joint distribution. Furthermore, the computational demands of this process are considerable. This computational cost also apply to PFI alternative methods that require retraining models (HOOKER; MENTCH; ZHOU, 2019; LEI et al., 2018; GREGORUTTI; MICHEL; SAINT-PIERRE, 2017).

Adopting simpler and more effective strategies to constrain the permutation process and avoid extrapolation, Conditional Permutation Feature Importance (cs_PFI), as initially proposed by (STROBL et al., 2008) and subsequently refined in (DEBEER; STROBL, 2020), utilizes the tree structures generated from random forest models to constrain the extrapolation problem. Similarly, (MOLNAR et al., 2023) computes the cs_PFI using an auxiliary decision tree to form subgroups for each feature, treating the feature of interest as the target variable. The tree splits turn the data points within leaves relatively independent with respect to other variables. The cs_PFI is computed within leaves and subsequently aggregated to produce an overall unbiased PFI.

This chapter aims to contribute to this suite of tools for assessing the robustness of score-based explanations in identify and isolate the relevance of features in supervised learning models, especially in situations where feature independence cannot be presumed. The new scores are motivated by the robustness of ALE technique under such conditions. The decomposition property of ALE, as outlined in Chapter 2 is particularly relevant in scenarios where

data is not independent. It aids in 1) ensuring that variables of low relevance are assigned minimal attribution, even if they are correlated with significant variables, and 2) distinguishing relevant variables regardless of their co-dependency.

The new scores are extracted from the ALE framework and can be computed for any model. Akin to cs_PFI, which is also model-agnostic, the new scores are computed across subgroups, but without the need to employing an auxiliary model to establish the subgroups. Additionally, the new scores can provide more information than traditional PFI-based scores. PFI-based techniques, which are ranking-based, focus on highlighting the relative importance of variables with respect to model performance—a perspective that does not fully align with the needs of educational practitioners who are interested in understanding the roles of features in the target variable. Typically, such an inquiry requires an examination of the relevance of features in model predictions rather than solely in performance, as has traditionally been conveyed by the coefficients of linear models.

It is expected that these new scores of feature effect size will be useful in practical scenarios where data are not independent, serving as an alternative to generating insights about the relevance of features and their interactions within the supervised learning paradigm. For comparison, a series of experiments with both artificial and real-world data will demonstrate how these scores can compete with and surpass the limitations of existing scores. Using the ALE framework to produce these scores also introduces certain limitations to the interpretation of the scores, which will be further discussed.

4.2 SCORES DEFINITION

In this section, we propose the ALE-based scores that allow the computation of 1) a more isolated individual effects of variables 2) to give different insights in the data-generating process even when data is not independent. Initially, this section presents metrics analogous to the coefficients in linear models, capable of reporting both the magnitude and direction of feature effects. Subsequently, it discusses a metric that only quantifies the overall relevance of features.

Formulating a single score to represent the effect size of a feature presents a complex challenge. In linear models, where the effects of features remain constant across their entire range, the slope obtained from the ALE method corresponds exactly with the linear model's coefficients. Utilizing a Bootstrap sampling approach in this scenario could potentially yield a confidence interval closely akin to that obtained through t-statistics, as illustrated in Table 5. However, this does not hold true for non-linear models. In such cases, assigning a single numerical value to express feature effects often fails to capture their intricate dimensions fully. Instead, a complete ALE plot is required to comprehensively capture the nuances of feature effects.

Nevertheless, considering ALE's decomposition property (see: Section 2.6), scores can meaningfully represent the individual effects of features if explicitly defined. For instance, ALE calculates isolated effects within specific data segments. When these segments are not too narrow or too broad and representative of the wider population, with minimal prediction noise, they can be particularly informative. In such cases, identifying the segment with the most pronounced feature effects can yield valuable insights into the degree to which a feature impacts the target variable. This is especially relevant for considering potential interventions in the current dataset.

The potential interventions are safeguarded due to ALE's implementation, which performs the do-operator (see: Section 3.1.2) for every data point of the segment. For instance, it can be insightful for the educational practitioner to understand that, regardless of individual differences, a potential educational policy (represented by one feature) may not be able to change the chances of any student success (target variable) more than a specified threshold. Similarly, finding out that an underlying policy tends to yield great results for a representative group of students can drive further research to determine who benefits most and who does not. Furthermore, the effect of a feature at a specific value can reveal its individual impact on

a particular group of interest of the researcher.

Building on this thought, four new metrics for feature effect size have been introduced. To ensure more reliable estimations, these metrics are designed to be computed on unseen data, distinct from the data on which the model was trained, as detailed in Algorithm 1:

Table 5 – Demonstration of the similarity of the coefficients and confident intervals of a linear regression and the slope computed from the ALE plot under a bootstrap sampling on artificial data

Feature	Slope	Lower CI	Upper CI	ALE_slope	Lower CI	Upper CI
0	44.018	44.015	44.021	44.021	44.015	44.027
1	-0.0025	-0.006	0.001	0.001	-0.006	0.007
2	-0.004	-0.007	-0.001	-0.000	-0.007	0.005
3	73.229	73.226	73.232	73.232	73.226	73.239
4	52.552	52.549	52.555	52.555	52.548	52.561
5	9.5595	9.556	9.563	9.563	9.557	9.568
6	63.286	63.283	63.289	63.289	63.283	63.295
7	13.519	13.516	13.522	13.522	13.516	13.528
8	69.562	69.559	69.565	69.565	69.559	69.572
9	40.185	40.182	40.188	40.188	40.182	40.195

Source: self-provided

Algorithm 1 Estimating ALE-based scores with optional uncertainty estimation

Require: model f, data $D = \{X_1, X_2, \ldots, X_k\}$, int j

1: Optional: Apply a sampling strategy (cross-validation or bootstrap) on D
2: Compute quantiles Q of order j using D for the target feature X_k, $q\grave{\imath}n\{1, ..., Q^k\}$
3: **for** $q \in Q^k$ **do**
4: $\quad \hat{X}_{max} := X$ do$(x_k := \max(q))$
5: $\quad \hat{X}_{min} := X$ do$(x_k := \min(q))$
6: $\quad LE_q(k) := f(\hat{X}_{max}) - f(\hat{X}_{min})$
7: **end for**
8: $ALE(k) = \sum_{q=1}^{Q^K} LE_q(k)$
9: Compute score based on a matrix $ALE(k)$ of size jXk

The ALE function is defined as a summation of local effects ascertained by the partial derivative. However, in practice, ALE implementations employ quantiles. The definition of quantiles already presents a challenge for accurately recovering feature effects under the visualization of ALE plots and further influences the precision and correctness of scores interpretation. Specifically, small quantiles may model noise, while larger quantiles may fail to uphold the inherent linearity assumptions of ALE within quantiles.

Therefore, for a more reliable interpretation of the scores, the quantiles (as defined by j in Algorithm 1, line 2) should be representative of the sample. This representation should be

meaningful within the domain context to reveal its summarized feature effects, while assuming linear feature effects within each quantile. In the domain of education, deciles and quantiles are common units of interest (CARNOY; ROSA; SIMõES, 2022; CARNOY et al., 2017)

Lastly, to account for the uncertainty inherent in the model and data, the metrics can be calculated using a cross-validation approach or a bootstrap sampling strategy. While Bootstrap allows for the construction of confidence intervals by the cost of high computational cost, cross-validation can produce generalized estimates with low computational effort. The ALE computation is relatively stable in absence of outliers, as the definition of quantiles tends to remain consistent. This stability is especially notable when the unique values of a feature are few or closely to the J definition. The greatest source of uncertainty typically lies within the model itself, as an algorithm may approximate different functions while delivering comparably effective performance. If outliers are present, appropriate handling would involve winsorizing the data distribution's tails or adjusting the quantile estimation method to minimize the influence of extreme values on predictions

In Line 9, in Algorithm 1, the process for summarizing the influence of a feature on model predictions is defined. Within this context, three metrics are first introduced, calibrated to the same scale as the model's output, offering insights into both the magnitude and direction of the feature's effect size. For clarity, these metrics are didactically compared to the traditional coefficients of linear models. Subsequently, a fourth metric, based on a ranking system, is introduced.

Consider: j represents the feature of interest k the quantile for feature j. $ALE_{j(k)}$ represents the uncentered Accumulated Local Effect for a specific feature j up to the k-th quantile. $z_{1j(k)}$ is the lower bound of the k-th quantile for feature j. $z_{2j(k)}$ is the upper bound of the k-th quantile for feature j. n is a specific value of interest of j

1) Maximum Uncentered ALE (MUA): the MUA metric urges actionability and extracts the maximum change in the predictions (y) that a feature of interest may derive. The maximum is related to a specific data interval and may be positive or negative, requiring a previous absolute comparison to achieve the highest value.

$$\text{MUA}(j) = \max_{k} ALE_{j(k)} \tag{4.1}$$

2) Uncentered ALE at a Specific Value (UAS): the UAS is an arbitrary choice of a specific value n of the feature of interest. In practice, it requires discovering into which quantile

this value falls and returns the accumulated uncentered ALE up this quantile. This metric may be helpful when the analyst has sufficient domain knowledge or is verifying hypotheses.

$$\text{UAS}(j, n) = ALE_{j(k)} \,|\, z_{1j(k)} \leq n \leq z_{2j(k)} \tag{4.2}$$

3) Average Uncentered ALE (AUA): the AUA represents the first-order moment of the distribution and offers insights into the general magnitude and direction of feature effects across specified intervals. The AUA captures the average isolated impact of a feature of interest on the predictions over these intervals. A lower AUA value does not necessarily indicate an insignificant variable; rather, it may reflect a variable with a lower average effects.

$$\text{AUA}(j) = \frac{1}{K} \sum_{k=1}^{K} ALE_{j(k)} \tag{4.3}$$

4.2.1 An analogy to the coefficients of linear regression

This section draws a simplified analogy between the coefficients of linear regression and the newly introduced metrics (MUA, AUA, and UAS). This analogy is instrumental as the coefficients represent completely different measures. However, it will be important to demonstrate how the new metrics can support research inquiries in the field of education.

In linear regression, the coefficients of a variable quantify its overall effect size. When a variable's effect is not constant across its distribution, interaction terms are introduced. These terms account for the variable's differential impact across specific segments of its distribution, such as quartiles. Consider the following model:

$$\text{Y} = \beta_0 + \beta_1 A + \beta_2 (A \times D_3) + \epsilon \tag{4.4}$$

Here, β_1 represents the baseline effect of variable A, while β_2 (the coefficient of the interaction term $A \times D_3$) captures the additional effect of A in the third quartile regarding the remains. In the case of a monotonically increasing effect of A on Y, the sum $\beta_1 + \beta_2$ aligns with the MUA. Conversely, if A's effect is monotonically decreasing, the MUA corresponds to β_1 added to the coefficient of the first quartile's interaction term, which is not included in Equation 4.4.

MUA thus represents the maximal average effect of variable A across its quartiles. As the metrics is an uncentered version of ALE framework, the average effect of variables will always

be included, which makes them more informative to the actual dataset. This is particularly important as the function is unknown and the interpretation relies only on a unique number. The AUA represents the mean of all such summations between β_1 and the coefficients of the interaction terms for each quartile. Meanwhile, the informs the same β_1 added to the interaction term of the quartile that encompasses a chosen specific value.

ALE Absolute Range (AAR): the AAR is a non-directional metric, distinguishing it from those previously mentioned. It quantifies the feature contribution by measuring the total absolute range from the minimum to the maximum effect. The AAR can be scaled from 0 to 1, providing a ranking-based alternative. This scaling enhances generalizability, facilitating broader comparisons of a specific feature's AAR across various datasets and models. A lower range indicates less reliance on the feature by the model.

$$\text{AAR}(j) = \left| \max_k ALE_{j(k)} - \min_k ALE_{j(k)} \right| \tag{4.5}$$

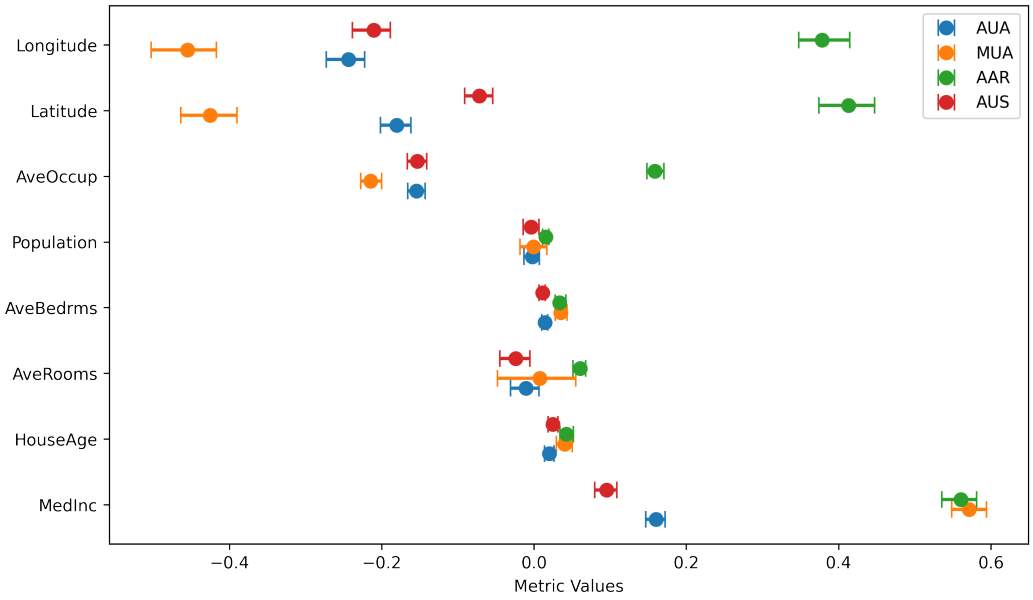
4.2.2 Building confidence intervals

Confidence intervals are crucial for inference and analyzing the reliability of estimates. They provide an estimate of the range within which the true metric values are likely to fall, given a specified confidence level. To construct these intervals in the context of ALE-based metrics, strategies such as data-only bootstrap or full-bootstrap can be used. The data-only bootstrap accounts solely for data uncertainty, whereas the full-bootstrap method considers uncertainties arising from both data and the model training. This method involves repeatedly resampling the dataset (and retraining the ML model for the full bootstrap), designated for explanation, with replacement and recalculating the metrics for each sample, thereby capturing data estimates variability.

Nevertheles, bootstrap sampling can be computationally demanding, particularly for large datasets or complex models. As an efficient alternative to capture robust estimates, k-fold cross-validation can be employed, though it does not inherently provide confidence intervals. In this method, the entire dataset is utilized for both training and explaining the model. The data is divided into $'k'$ subsets. The model is trained $'k'$ times, using a different fold as the explanation set each time, while the remaining data is used for training. This ensures that every data point contributes to both training and feature explanations.

Figures 6 and 7illustrate the use of both data-only and full-bootstrapping strategies for all metrics, including their confidence intervals at a 5% significance level. The was defined using the median for each feature. The AAR was not normalized. These metrics are derived from predictions of house prices using the openly available California Housing dataset[1] , employing the random forest regression algorithm with its default parameters as implemented in scikit-learn[2]. Prior to modeling, data preprocessing was conducted to manage outliers, setting bounds at the 97.5^{th} and 2.5^{th} percentiles to mitigate the impact of extreme values. This dataset was chosen for demonstrating uncertainty computation because it is open, simple, and extensively used in the MLliterature, which can facilitate future comparisons

Figure 6 – Illustration of the confidence interval construction at a 5% significance level through a data and model (full) bootstrap over 100 iterations



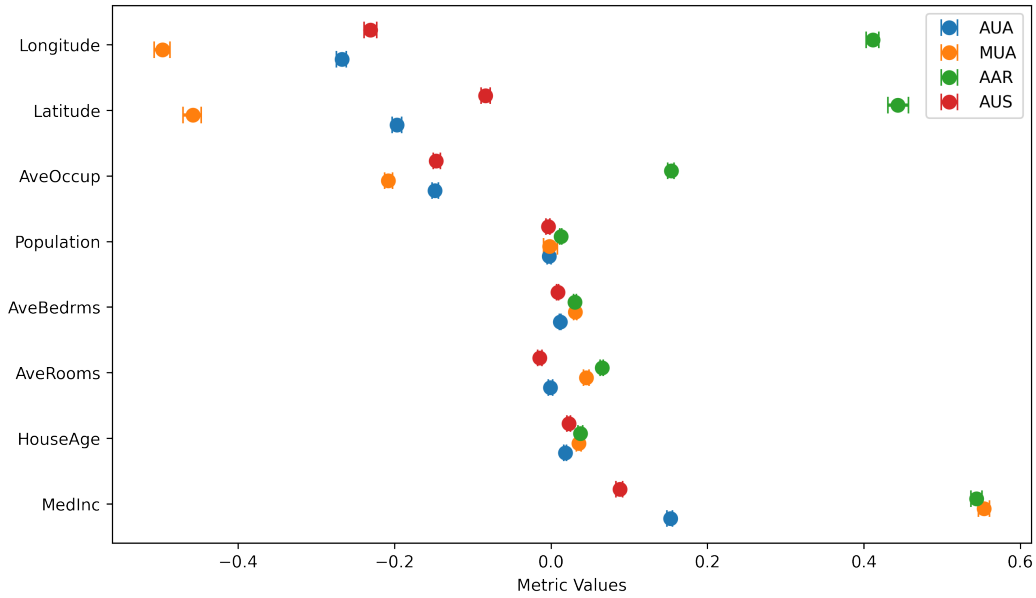
Source: self-provided.

4.2.3 Interaction feature effects size

Although ALE framework can be defined for any interaction order effects (APLEY; ZHU, 2020), the visualization through plots is typically reliable only up to second-order effects. This is

[1] <https://scikit-learn.org/stable/modules/generated/sklearn.datasets.fetch_california_housing.html>
[2] https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html

Figure 7 – Illustration of the confidence interval construction at a 5% significance level through a data-only bootstrap over 100 iterations



Source: self-provided

because higher-order effects cannot be effectively represented in plots. By employing a scoring system, the interaction of n-order effects can be made more readable; however, interpretability decreases and utility becomes more limited as the value of n increases. Following ALE definition (APLEY; ZHU, 2020), the computation of scores essentially follows the algorithm 1. Notably, instead of one-dimensional quantiles, n-effects analysis requires n-dimensional partitions, and an adjustment for the right interpretation of the interaction additional effect is necessary.

For a given model function f, the analysis typically begins with partitioning each feature into quantiles or intervals. In the context of second-order effects, a two-dimensional grid is created by forming the Cartesian product of the quantiles of both features (APLEY; ZHU, 2020). This approach is then extended to n dimensions for higher-order effects involving more features. The essence of this analysis lies in how the model's response varies across these multidimensional partitions.

In the end, it is necessary to subtract the main effects of both features from each partition's local effects. Thus, the interpretation of the interaction is essentially the uncentered additional effect resulting from the interaction of both features.

4.3 METHODS

This section details the methodology employed to assess the efficacy of the proposed metrics in addressing RQ2. The evaluation was defined with two primary objectives: 1) to verify whether variables of low relevance are attributed minimal significance, even when they are highly correlated with relevant variables, and 2) to ascertain if the proposed metrics can effectively isolate the effects of relevant variables that are also correlated with each other.

To systematically achieve these objectives, the evaluation process was structured into two phases. First, synthetic data was simulated with the desired dependencies of the features and a qualitative comparison is then carried out to examine how various methods emphasize variables in different scenarios. Following this, evaluation metrics are introduced for a quantitative analysis. Using simulated data allows for a more precise demonstration of the metrics' behavior than analyzing a real dataset, where additional interdependencies among features might skew the results. Although it is not possible to ensure that the fitted model will recover exactly the simulated data-generating process, it is expected to come close by keeping the setup simple and focusing on a small set of variables. Additionally, the same model will be explained for different techniques, which will allow a comparison of differences in the technique's behavior even if the models fit data differently from the expected.

In the second phase, the proposed metrics are evaluated using two publicly accessible real-world data sets from the medical and educational domains. The first data set exhibits high interdependence among features with numerous variable pairs demonstrating a Pearson correlation coefficient exceeding 0.8. The second data set pertains to the educational domain, which is the primary focus of this thesis. For this dataset, domain-specific knowledge is leveraged to elucidate disparities between highlighted feature contributions.

The baseline metrics employed for comparison include MDI from tree-based models, the model-agnostic version of PFI, the Average SHAP, and the conditional version of PFI cs_PFI. The MDI scores were directly extracted using the scikit-learn library (PEDREGOSA et al., 2011). Average SHAP were computed via the TreeExplainer package [3]. The TreeExlainer (TreeSHAP) was proposed by (LUNDBERG et al., 2020) and is confined to certain tree-based algorithms. From the same author of the SHAP, The TreeSHAP offers computational efficiency over the original SHAP formulation and is also more robust to the extrapolation issue making it a better baseline

[3] <https://shap.readthedocs.io/en/latest/example_notebooks/tabular_examples/tree_based_models/UnderstandingTreeSHAPforSimpleModels.html>

alternative than the original SHAP. The PFI was implemented as defined in (FISHER; RUDIN; DOMINICI, 2018), utilizing mean squared error for regression and accuracy for classification as the performance metric. The cs_PFI was implemented in accordance with (MOLNAR et al., 2023). The cs_PFI approach requires the utilization of an auxiliary decision tree to partition the feature space prior to feature permutation. The tree has a maximum depth of 2 and the minimum number of observations in each leaf as 10% of the dataset. For both PFI and cs_PFI, the random feature permutation was executed five times, and the average was adopted as the final score.

In addressing the constraints associated with MDI and TreeSHAP, which are solely applicable to tree-based algorithms, the RF algorithm was chosen. The RF is prevalent in EDM(MARTÍNEZ-ABAD; GAMAZO; RODRÍGUEZ-CONDE, 2020) and has also demonstrated optimal performance in many tasks using tabular data (GRINSZTAJN; OYALLON; VAROQUAUX, 2022). The ALE-based score AUA was chosen for qualitative benchmarking, as it offers a more intuitive measure and aligns closely with the interpretation needs of educational practitioners interested in the main effects of variables on predictions. For the quantitative analysis, the normalized AAR was employed due to its stronger adherence to other ranking-based techniques. Regarding SHAP, the common average was utilized in the qualitative analysis, and the absolute average SHAP for the quantitative analysis.

4.3.1 Evaluation metrics

To address RQ2, the primary objective is to define metrics that can isolate the effects of individual variables in dependent data. It is essential to ensure that irrelevant variables are not inaccurately considered important due to their correlation with significant variables. For this purpose, the metric *PropTrueVar* is introduced. It calculates the proportion of the total relevance attributed to all variables that is assigned to the irrelevant variables. Technically *PropTrueVar* is defined by:

$$PropTrueVar = 1 - \frac{\text{score\_important\_variable}}{\sum(\text{all\_scores})} \tag{4.6}$$

Additionally, not separating the effects of different features can lead to incorrect conclusions as it cannot be possible to identify from which variable the effects come, particularly when several correlated variables similarly affect the target variable. To assess such cases, the

EquiTrueVars metric is defined. It measures the fairness in attributing importance to correlated variables that are theoretically equally relevant to the target variable. Fairness is quantified by the standard deviation (σ of the attributed scores for the important variables. Low $sigma$ values indicate that the explanation correctly assigned similar relevance to the equally important variables, while high values suggest a lack of fairness in importance attribution. The *EquiTrueVars* is defined as follows:

$$EquiTrueVars = 1 - \sigma(\text{score\_important\_variables}) \tag{4.7}$$

Lastly, the *EquiPropTrueVars* metric is introduced to measure both aspects. It evaluates the combined property by first assessing the attribution given to important variables (using *EquiTrueVars*) and then penalizing methods that tend to overemphasize irrelevant variables (based on *PropTrueVar*).

$$EquiPropTrueVars = EquiTrueVars - (PropTrueVar - 1) \tag{4.8}$$

4.3.2 Assessing scores for feature selection

Feature selection stands as a critical aspect of ML, garnering increased focus with the advent of high-dimensional datasets from diverse fields. While not directly tied to the central research questions of this thesis, the relevance of feature selection remains significant in ML, particularly in the context of enhancing model performance while reducing model complexity. This section is dedicated to exploring how ALE-based scores can offer a viable alternative in this direction. The focus here shifts from detailing the relationship between features and the target variable to evaluating the contribution of features to model performance. The AAR, which measures the total effect range from minimum to maximum, emerges as a potential indicator of feature importance in this regard.

Given that an exhaustive comparison of feature selection methods falls outside the scope of this thesis, for the sake of simplicity, this study will narrow its focus to EDM and MDI from tree-based models, particularly due to MDI's widespread application in feature selection tasks. Additionally, evaluating the method's impact on the performance and complexity of RF models introduces some bias in favor of MDI. This inherent bias arises because MDI is computed as part of the training process of the Random Forest model, which is also used to fit the data.

Therefore, if AAR demonstrates comparable performance to MDI in this context, it would be a substantial indication of AAR's effectiveness as a feature selection tool. The MDI was extracted from the RandomForest model of the Sci-kit learn library [4].

The initial experiment utilizes another open-source UCI dataset, specifically an educational dataset (REALINHO et al., 2022) containing 36 independent variables. This experiment highlights the divergences between the MDI and AAR methods in their potential for feature reduction.

Initially, irrelevant variables (close to zero, "< 0.0001") are excluded to evaluate model performance. Subsequently, the model's efficacy is evaluated using the top 10 features as identified by both MDI and AAR metrics. These metrics are derived from the same RF algorithm applied to a subset (30%) of the dataset, while the remaining portion (70%) is used to assess model performance in predicting student dropout (enrolled or graduated). The AAR and MDI values were normalized such that their sum equals one. This normalization ensures a standard scale for comparison.

To illustrate feature selection in high-dimensional settings when the number of features is large (p) relative to the number of observations (n), this study used two datasets from OpenML (VANSCHOREN et al., 2014). Specifically, dataset id=312 with n=2407 and p = 299 and dataset id=1485 with n= 2600, p= 500), both previously employed in feature selection research. Here, varying thresholds of feature irrelevance are analyzed concerning both model complexity (number of features) and performance.

4.3.3 Synthetic data

Simple scenarios were set where is easier to identify differences in the metrics behavior. Especially, considering a set of variables X drawn from the same normal distribution $N(0, 1)$ with values ranging between zero and 1 and ε following $N(0, 0.1^2)$ as well as a dependence between variables pairs M plus noise $N(0, 0.05^2)$ (ex.:$x_1 = x_2 + \varepsilon$), diverse scenarios were simulated to explore different dependency structures. Each scenario has 1000 data points and Y is a function of X with different dependence pairs M:

Scenario 1: X is matrix of size 5, Y depends only on x_1: $Y = x_1 + \varepsilon$ and M maps a strong dependence between x_1, x_2, being the all the other variables independent

Scenario 2: X is matrix of size 5, Y depends only on x_1: $Y = x_1 + \varepsilon$ and M maps a strong dependence between x_1, x_2; x_2, x_3; x_4, x_1, being only x_5 independent

[4] https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html

Scenario 3: X is matrix of size 5, Y depends on x_1 and x_2: $Y = x_1 + x_2 + \varepsilon$ and M maps a strong dependence between x_1, x_2 ;x_2, x_3; x_4, x_1, being only x_5 independent

Scenario 4: X is matrix of size 5 Y depends on x_1: $Y = x_1$ and M maps the Pearson correlation p being $0.5 \leq p < 1.0$ among all variables.

Scenario 5: X is matrix of size 5 Y depends equally on x_1, x_2 and x_3: $Y = x_1 + x_2 + x_3$ and M is defined by $coor$ being $0.5 \leq coor < 1.0$. The variable $coor$ controls that extent of the correlation among variables through: $X_1 = corr \times X_2 + corr \times \varepsilon$ and $X_3 = corr \times X_1 + corr \times \varepsilon$

Scenario 6: X is a matrix of size 10 . The target variable Y is equally influenced by a set of five relevant variables, denoted as $G_1 = \{x_1 : x_5\}$ and has a weak relation with variables from $G_2 = \{x_6 : x_10\}$. The G_1 are generated with intercorrelation controlled by the parameter $corr_intra$. The remaining five variables are correlated with G_1 through the $corr_inter$. Both $corr_intra$ and $corr_inter$ will vary from 0.1 to 0.8.

4.3.4 Real-world data

To demonstrate the practical use of the new metrics on real-world data, two more openly available scientific datasets from the repository were used[5]. The first is the Breast Cancer dataset (DUA; GRAFF, 2017), well known in the ML community for its high interdependence between variables. The second is an educational dataset (YILMAZ; SEKEROGLU, 2020), which is particularly relevant as it comes from the domain that motivates this thesis. In these experiments, the focus is gaining a qualitative understanding of how differently the proposed and baseline metrics isolated and distinguished the effects of variables.

The breast cancer data included benign and malignant cell samples from 369 patients, 212 with cancer, and 157 with fibrocystic breast masses. Each sample contained thirty features which were used by the model to predict the type of cancer. The educational data consisted of the performance of 145 higher education students, 57 with scores above or equal to 4 and 88 below, described in 31 variables linked to socioeconomics, demographic, and student behavior. At all, 30 features (all features less "STUDENT ID", "COURSE ID) were used to predict student grade. The grade target variable was transformed to a classification task and grades below 4, receive 0 otherwise 1.

[5] <https://archive.ics.uci.edu/datasets>

4.3.5 Experimental setup

In the first phase of the experiments, which involved synthetic data, regression and classification functions from Scenarios 1 to 3 were tested and evaluated qualitatively. For the classification, all $Y > 0$ was set to 1, otherwise 0. From Scenarios 4 to 6 the entire dataset was used for model fitting, and the defined evaluation metrics were extracted across the same 30-Monte Carlo replicates.

For both datasets, the algorithms were applied to a 5-fold cross-validation setting in a binary classification task. Thus, both algorithms were applied to the same folds with scikit-learn default parameters, and the mean was adopted as a feature contribution for each metric. It's important to note that all metrics were computed using identical fold partitions, ensuring consistency in the evaluation process.

4.4 RESULTS

This section presents the results of experiments conducted to validate the proposed scores to extract insights about the data-generating process. Initially, the findings from synthetic data are displayed, followed by two examples using real-world data. Ultimately, additional experiments underscore the potential of the proposal in the context of the feature selection task.

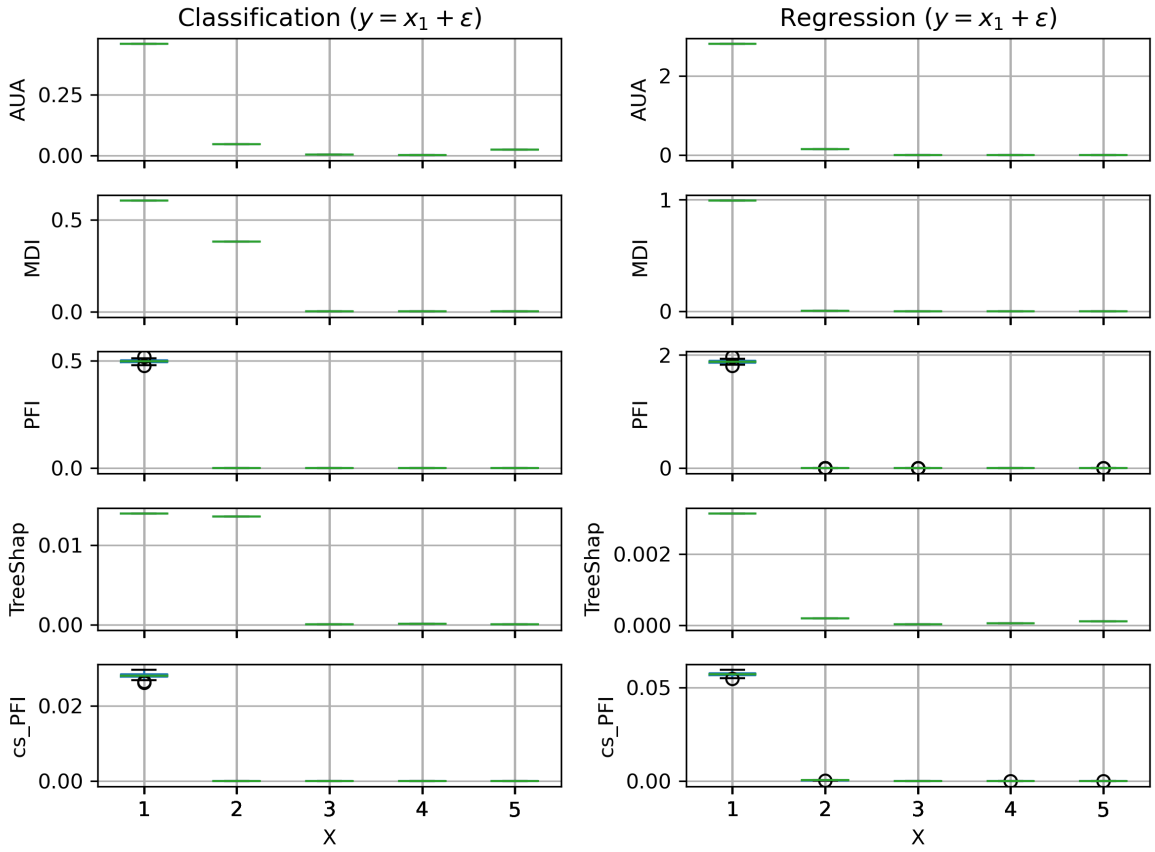
4.4.1 Synthetic data

4.4.1.1 Qualitative analysis

In Figure 8, the results for **Scenario 1** are depicted. In this simplified setup, all methods accurately identify x_1 as the key variable. However, the strong correlation between x_1 and x_2 leads ThreeShap and MDI to incorrectly assign substantial relevance to x_2 in the classification model. In contrast, AUA and permutation-based methods (PFI and cs_PFI) exhibit better results. On the other hand, for the regression model, all metrics had the expected result highlighting x_1 by far as the most important variable assign almost zero relevance to the other variables. It is worth noting that the permutation-based models display some variability in the scores, a consequence of their random sampling process. Increasing the number of iterations could reduce this variability while also increasing computational load.

With similar results, Figure 9 shows to the **Scenario 2** how TreeShap and MDI give too much relevance for the other correlated feature in the classification task, even though y depends only on x_1. The MDI and cs_PFI correctly assign minimal relevance to the uninformative but dependent variable x_2, whereas AUA attributes slightly higher, yet still very low relevance to this variable. This is expected, as AUA computes differences in predictions while PFI and cs_PFI differences in performance. As only one variable determines outcomes, permuting any other feature does not change performance, although it may introduce some prediction noise. Nevertheless, this noise does not critically affect the recognition of x_1 as the unique key variable to the second scenario. This pattern is evident in the regression task, where all techniques yield satisfactory results. However, techniques that measure changes in predictions, such as TreeShap and AUA, do account for some noise in the irrelevant variables. The TreeShap and MDI also erroneously attribute so much relevance to the independent and irrelevant variable

Figure 8 – Scenario 1 - Comparison of the ALE-based metric (AUA) with the baseline for the classification and regression task using the random forest model fitted over 100 Monte Carlo replicates.

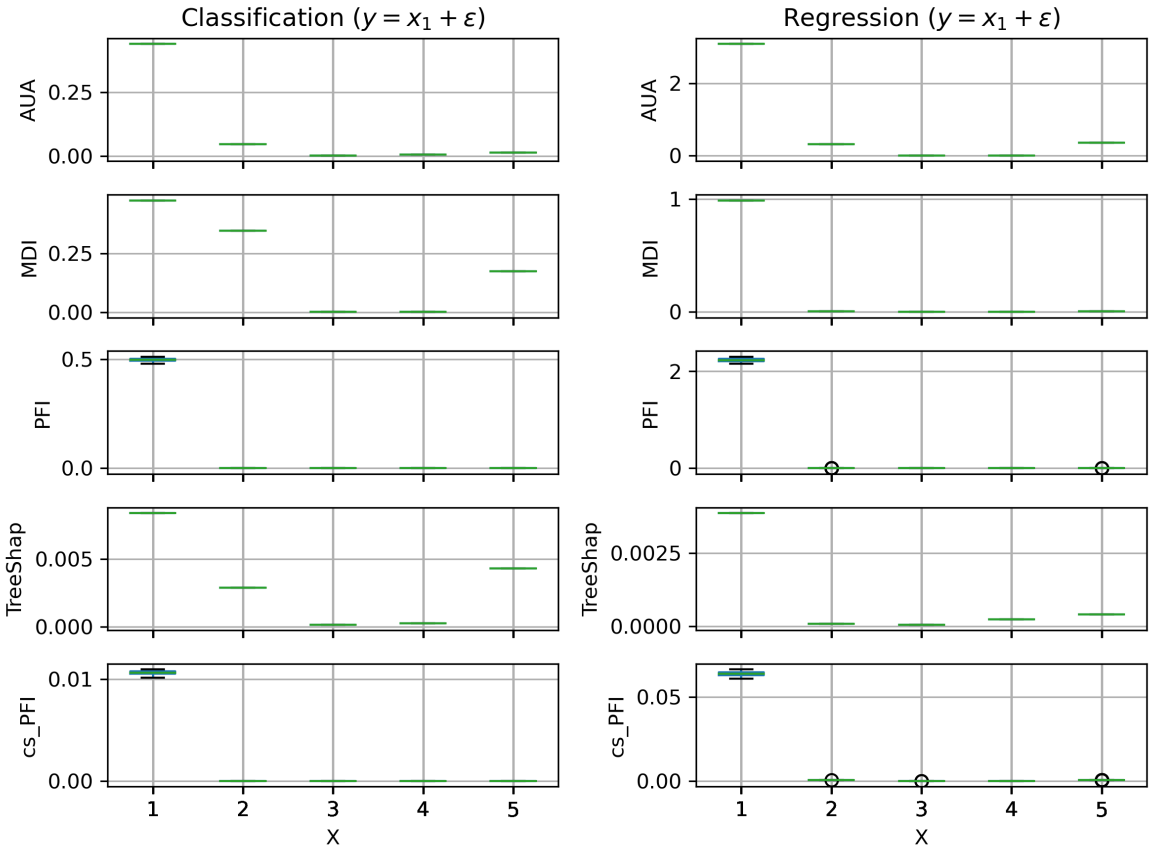


Source: self-provided

x_5 in the classification task.

Figure 10 displays the results for **Scenario 3**. Unlike the second scenario, where a single variable determines the outcome, the target variable y in the third scenario is influenced by both x_1 and x_2. Given that x_1 is dependent on x_2, PFI and cs_PFI encounter difficulties in accurately assessing the relevance of these variables, erroneously favoring one over the other in both classification and regression tasks to PFI and only in the regression task for cs_PFI. Conversely, TreeSHAP, MDI, and AUA yield more accurate feature relevance rankings. However, in the classification task, both AUA and MDI incorrectly attribute some importance to the feature x_5, which, although highly correlated with x_1, has no direct influence on y. The error is more critical for MDI, which assigns approximately half the relevance to x_5 as it does to x_1, whereas AUA erroneously assigns a notably lower relevance.

Figure 9 – Scenario 2- Comparison of the ALE-based metric (AUA) with the baseline for the classification and regression task using the random forest model fitted over 100 Monte Carlo replicates.



Source: self-provided

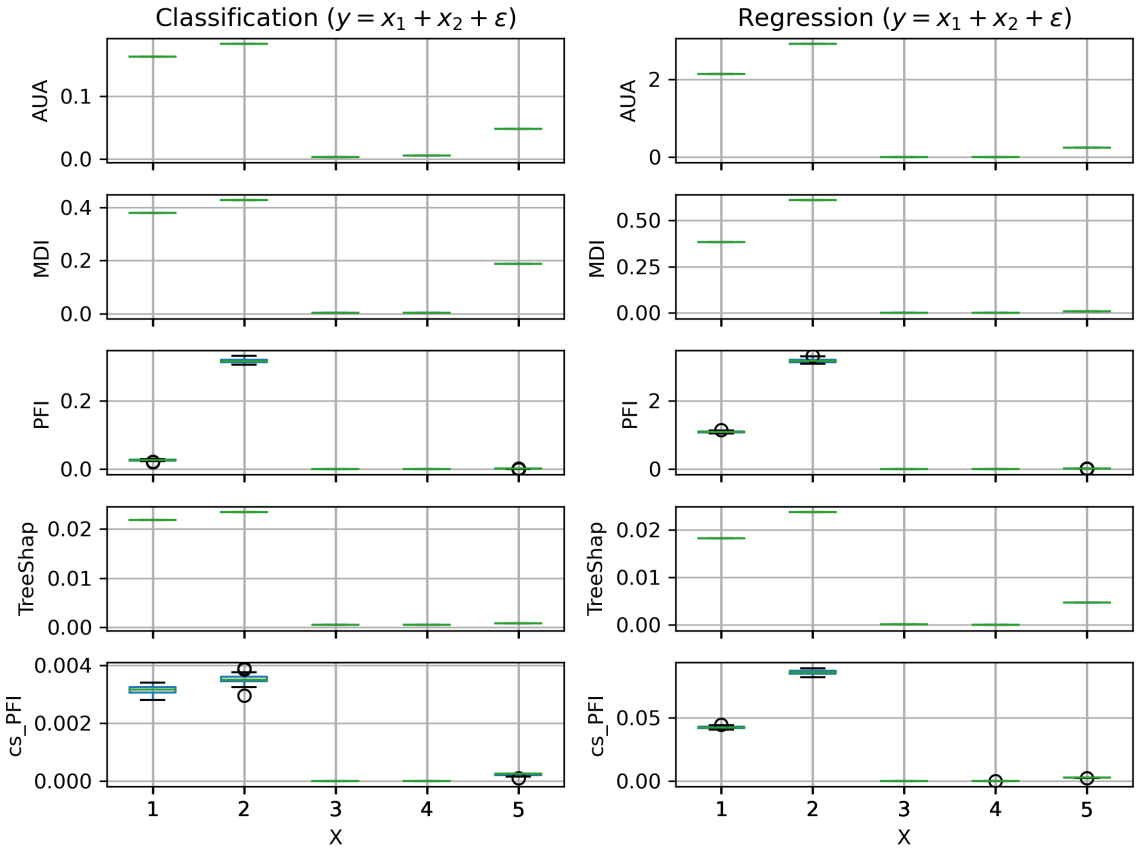
4.4.1.2 Quantitative analysis

Figure 11 shows how the performance of methods changes to the *PropTrueVar* as the Pearson correlation between variables escalates from 0.1 to 0.8 as defined in **Scenario 4**.

A *PropTrueVar* value approaching 1 signifies an effective technique in assigning higher relevance to the true explanatory variable, whereas a value substantially deviating from 1 indicates a less reliable method. Figure 11 corroborates previous findings, revealing a more robustness of PFI, cs_PFI and the ALE-based (AAR) across interactions, while MDI and TreeSHAP fail to identity and isolate the effects of the unique relevant variable x_1 as the level o colinearity increase. Also, TreeSHAP shows a high variability in this simple setting, probably due to its strategy to handle the conditional expectation based on the branch of trees.

Figure 12 illustrates the method's performance on *EquiTrueVar* using **Scenario 5**. PFI achieves the worst results due to its difficulty in highlighting two important variables when

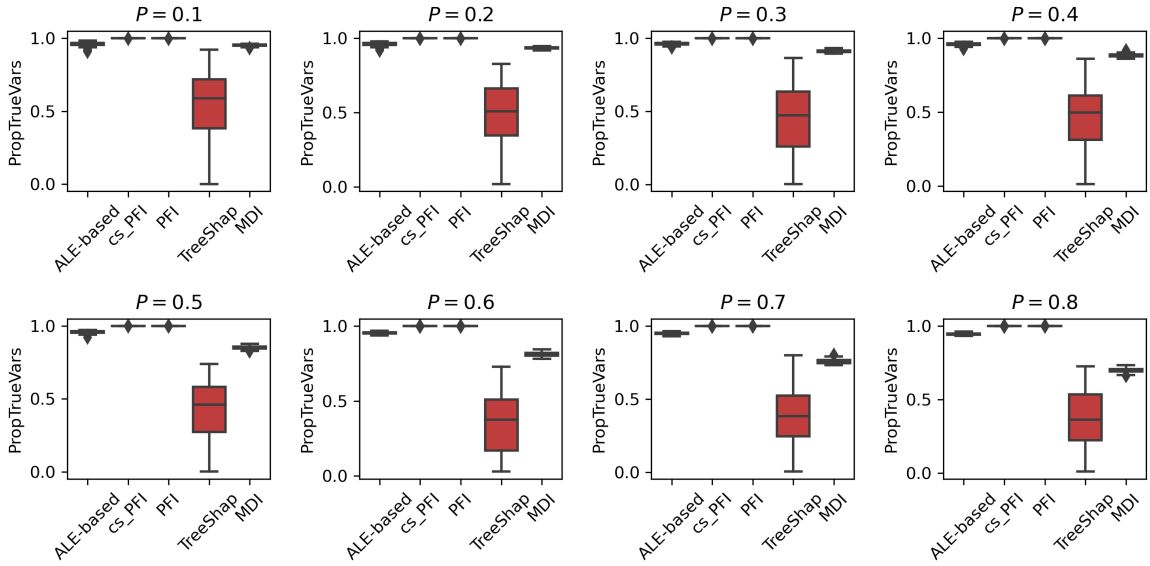
Figure 10 – Scenario 3 - Comparison of the ALE-based metric (AUA) with the baseline for the classification and regression task using the random forest model fitted over 100 Monte Carlo replicates.



Source: self-provided

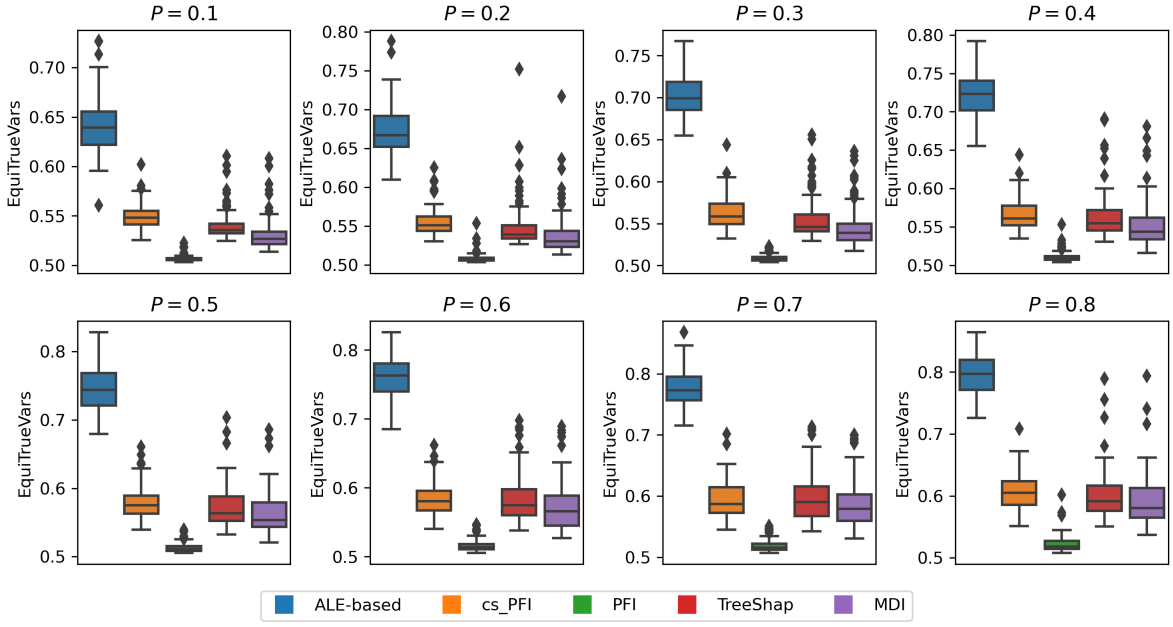
they are not independent. The ALE-based score shows great performance at all measured levels of feature dependence. The superiority of the ALE-based method is also evident in **Scenario 6** when assessing the *EquiPropTrueVars*. The *EquiPropTrueVars* measures a scenario in which the correlation between relevant G_1 and irrelevant variables G_2 increases, alongside the correlation within G_1 variables. Figure 13 illustrates the superior performance of the ALE-based approach (AAR) across almost all values of intra- and inter-correlations. The performance with respect to *EquiPropTrueVars* diminishes when the intra-dependence within G_1 becomes critically high, while the dependence between G_1 and G_2 appears to have a lesser impact on the proposed metric. However, in nearly all configurations, the ALE method achieves the best results.

Figure 11 – Scenario 4 - The performance for the metric PropTrueVar across different levels of correlation between the true, relevant variable, and irrelevant variables. PropTrueVar measures the proportion of total relevance attributed to all variables that are assigned for the true relevant variable. The PFI, cs_PFI, and ALE-based scores achieve better results when the correlation is high (after 0.6).



Source: self-provided

Figure 12 – Scenario 5- The performance for the metric *EquiTrueVar* across different levels of correlation between the two true relevant variables. The ALE-based scores achieve better results.

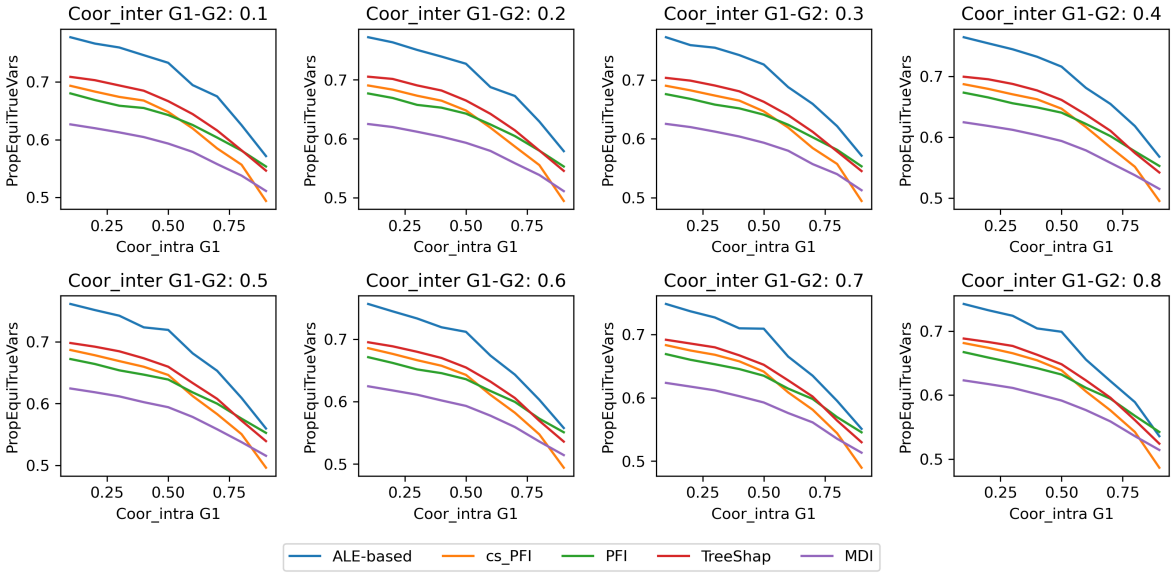


Source: self-provided

4.4.2 Real-world data

The results for the Breast Cancer dataset are depicted in Figure 14. The red bars represent the top 10 features as determined by each metric. Intriguingly, the strong interdependence

Figure 13 – Scenario 6- The performance for the metric *EquiPropTrueVar* across different levels of correlation between relevant variables and irrelevant variables as well as within the relevant variables. The ALE-based scores achieve better results for almost all settings.



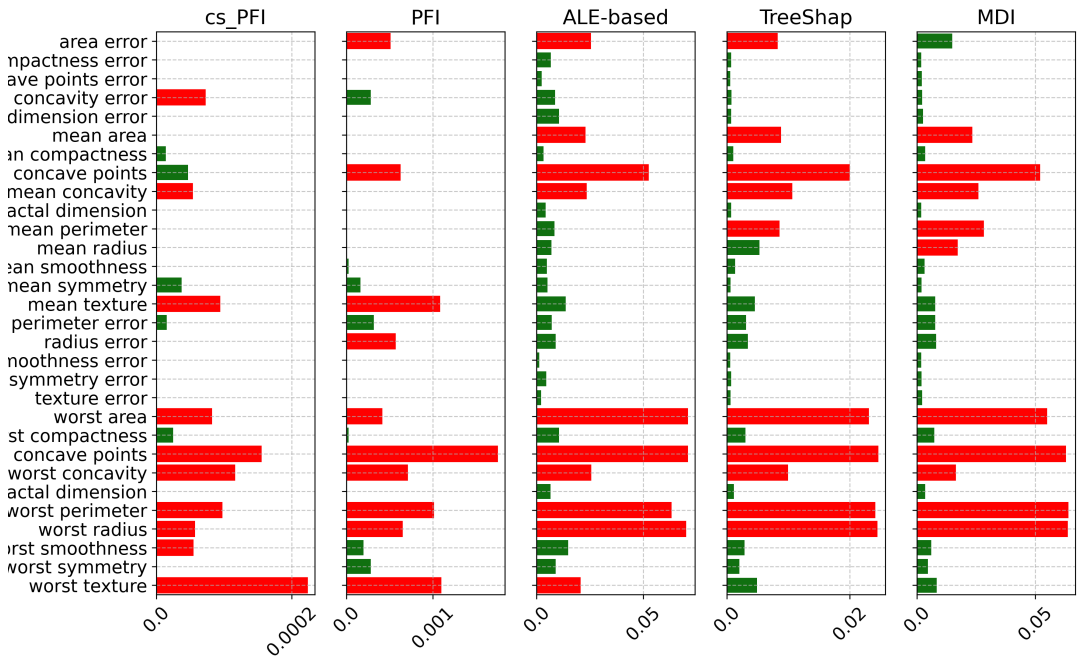
Source: self-provided

among many variables in this dataset highlights the limitations of permutation-based metrics such as cs_PFI and PFI, which depend on differences in model error for their calculations. Looking at the absolute value of feature effects size, these metrics suggest that almost none of the features are relevant, as evidenced by the assignment of values close to zero. There are numerous variables with zero effect, and no variable changes the model accuracy by more than $2x10^{-}3$ in cs_PFI and $2x10^{-}2$ in PFI.

In the previous experiment, cs_PFI and PFI metrics performed well for PropTrueVar, where only a single variable was relevant. However, in the current scenario, they cannot distinguish the impact of individual variables due to the high interdependence among variables, which is visually evident in Figure 15 where dark colors represent strong correlations as measured by the Pearson coefficient.

In contrast to PFI and cs_PFI, the AAR, MDI, and TreeSHAP methods consistently identify many relevant features despite some discrepancies. They revealed a similar set of features within the top 10. All metrics were organized in a ranking-based setting. For the glsALE-based score, the glsAAR was used, and for TreeSHAP, the absolute average method was employed.

Figure 14 – Cancer data - Red bars represent the top 10 features determined by each metric.



Source: self-provided

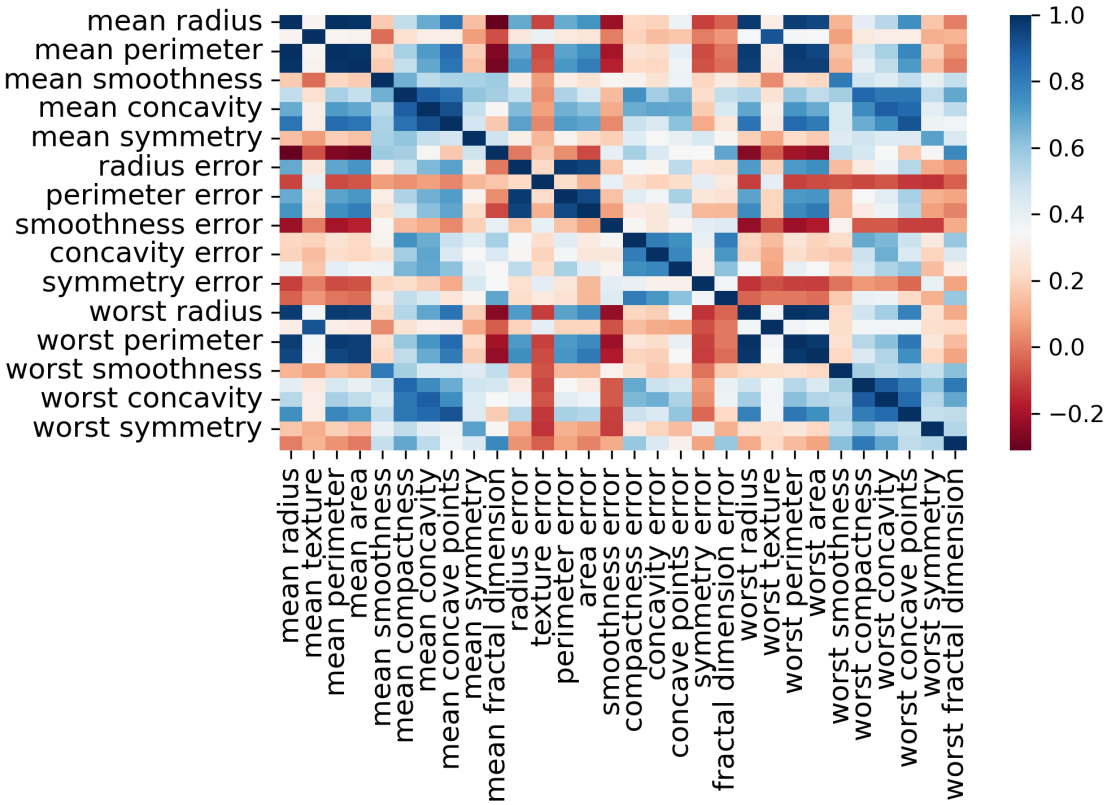
4.4.2.1 An inconsistency of TreeSHAP

In the Educational dataset examined, Figure 16 illustrates that the MDI exhibits greater density in its outputs compared to other metrics. Specifically, cs_PFI and PFI are the most sparse and exhibit high symmetry. Both metrics underscored features 29 (cumulative grade point average in the last semester) and 8 (salary) as paramount. Moreover, feature 29 emerged as a significant variable across multiple metrics. However, a discrepancy was observed between AUA and TreeSHAP regarding the direction of this feature contribution. AUA indicated a negative influence for feature 29, counter to domain knowledge, while TreeSHAP suggested a positive influence. To further scrutinize this divergence, we employed the original version of SHAP (in red in Figure 16), which corroborated AUA's findings.

Furthermore, an additional logistic regression model further substantiated this negative effect with a statistically significant coefficient of -0.99. PD plots for both logistic regression and random forest models also corroborated this negative direction, raising evidence of the negative effect of feature 29 to this dataset.

Upon closer analysis of the cross-validation process, it was observed that TreeShap assigns both positive and negative effects across iterations, a phenomenon not seen with other

Figure 15 – Cancer data correlation map. Dark colors represent strong correlations.



Source: self-provided

techniques. These changes in sign, depending on the fold, introduce noise into the final signal of relevance and diminish the overall relevance of the feature due to positive and negative cancellations.

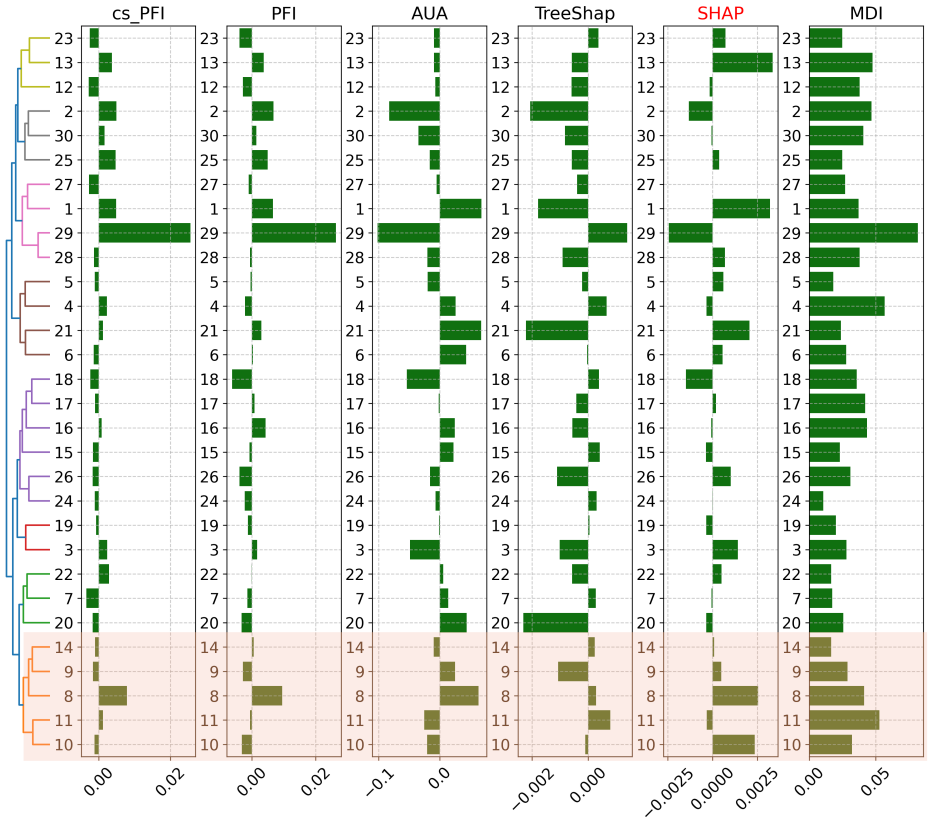
The inconsistency observed in TreeSHAP might be related to its strategy of holding conditional expectations to avoid data extrapolation, as previously documented (FILHO; BRITO; ADEODATO, 2023a). In a controlled experiment, the authors evaluated multiple algorithms using different metrics of feature contribution on a dataset where two features were identically correlated with the target variable and shared the same joint distribution. Despite these conditions, the two features received significantly different average feature contributions, but only in the TreeSHAP metric. This divergence arises from the different conditional expectations used to compute the effects of both features, even though they have exactly the same distribution. The conditional distribution used to compute feature effects depends on the specific terminal branches where the features are located within the decision trees. In this experiment, the small sample size and the repeated stochastic nature of random forest modeling likely contribute to this inconsistency across the 5-fold cross-validation process.

4.4.2.2 The robustness of the ALE-based score

In Figure 16, a dendrogram computed using the scipy package[6] identifies a set of features highly correlated with Salary (feature 8), a variable widely acknowledged as a determinant of educational performance (COLEMAN, 1968; COLEMAN, 2019). This correlated group comprises features such as *Transportation to the University* (9), *Accommodation Type in Cyprus* (10), *Mother's Education* (11), and *Parental Status* (14). Upon closer inspection of this cluster (highlighted in orange in Figure 16), it becomes evident that various techniques may accentuate correlated features, thereby introducing potential false positives. The MDI metric distributes relevance broadly across the entire group, with *Mother's Education* (11) emerging as the most significant. Conversely, other metrics, with the exception of TreeSHAP, align with domain knowledge and prioritize *Salary* (8). Among these, PFI, cs_PFI, and AUA appear to be more robust in isolating inter-feature dependencies within the group. TreeSHAP, however, probably erroneously attributes nearly equal importance to *Accommodation Type in Cyprus* (10) as it does to *Salary* (8), despite the former should not be directly related to educational performance.

[6] <https://scipy.org/>

Figure 16 – Feature attribution to the educational data. The orange area highlights a cluster of features highly correlated with the Salary, as illustrated in the left dendrogram. The dendrogram was computed using the scipy package with the complete linkage option.



Source: self-provided

4.4.3 Feature selection

In assessing the efficacy of ALE-based scores for feature selection, the AAR proved effective, reducing model complexity without significantly impacting model performance, akin to the baseline MDI. The initial experiment, utilizing educational data, underscored the benefits of AAR, achieving a 28% reduction in the number of features with only a slight impact on model performance. In contrast, despite the fact that MDI had been built during random forest training and can have the full information about how the algorithm used features, it was less effective in reducing feature count, deeming only one variable as irrelevant (< 0.001). Notably, when comparing the performance of models using the top 10 variables identified by both scores, the results were consistent despite a disagreement in 4 variables within the top

10. The Table 6 presents these findings, showcasing outcomes for the random forest algorithm.

Table 6 – Comparison between MDI from RF and the ALE-based score AAR for feature selection in an openly educational dataset to predict student drop-outs.

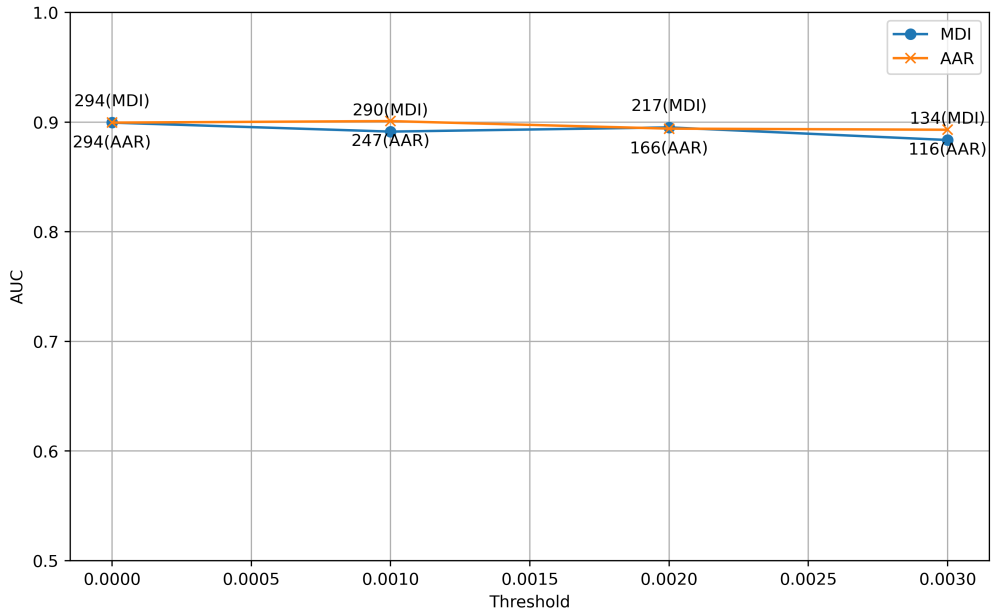
Model	AUC_ROC	Number of features
Random Forest (All Features)	0.91	36
Random Forest (MDI Relevant Features)	0.91	35
Random Forest (AAR Relevant Features)	0.90	25
Random Forest (MDI Top 10 features)	0.88	10
Random Forest (AAR Top 10 Features)	0.88	10

Source: self-provided

In the second series of experiments with OpenML datasets, the AAR method effectively reduced the number of features. For dataset id=312, shown in Figure 18, reducing the number of features did not significantly boost the model's performance, as the Area Under Receiver Operating Characteristic Curve (AUC_ROC) (FAWCETT, 2006) remained stable for both AAR and MDI methods. However, AAR was more successful in cutting down the number of features. In the initial threshold, AAR reduced the number of features by 16%, compared to MDI's reduction of 13%. At the subsequent evaluated thresholds, AAR achieved reductions of 44% and 60%, while MDI achieved reductions of 26% and 54%, respectively.

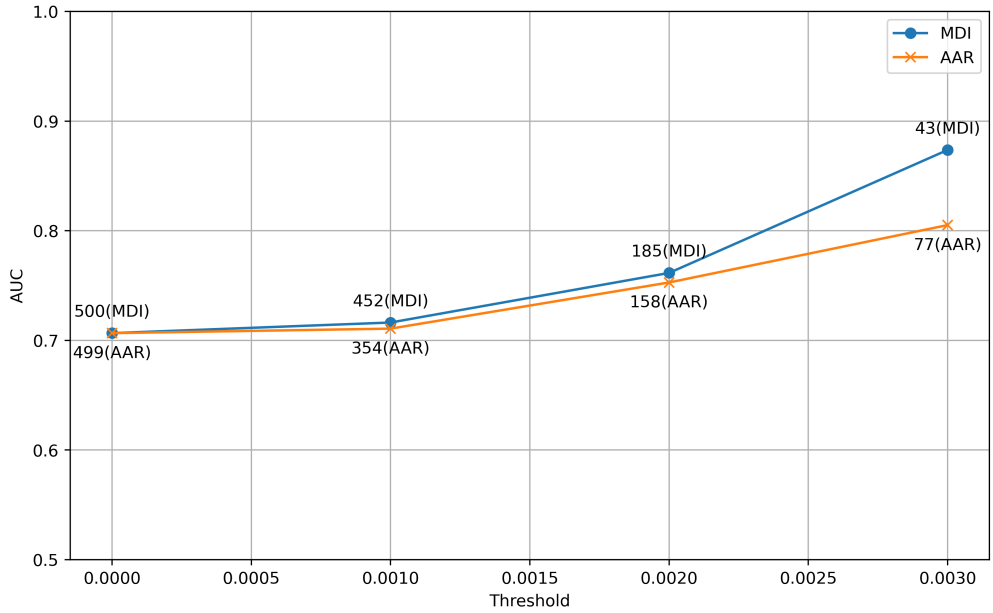
With dataset id=1485, removing features led to better model performance. This aligns with earlier results, showing AAR's efficiency in simplifying the model without compromising its effectiveness, particularly up to a threshold of 0.002. At a threshold of 0.001, AAR managed to reduce 20% more features than MDI. Notably, at a threshold of 0.03, MDI performed better, possibly due to its direct connection to the RF model used for fitting the data. To investigate this further, the RF model was replaced with a logistic regression model in the second step. The results are presented in Figure 19. Using the logistic model, AAR demonstrated even stronger results up to a 0.002 threshold, while MDI continued to achieve greater feature reduction at the 0.003 threshold. However, as expected, the difference in performance at this threshold is now negligible.

Figure 17 – Comparative analysis of the ALE-based AAR score and MDI from RF in reducing model complexity by minimizing the number of features. The dataset is from OpenML (id=312) in a classification task. The number represents the number of remaining features on the dataset for each metric.



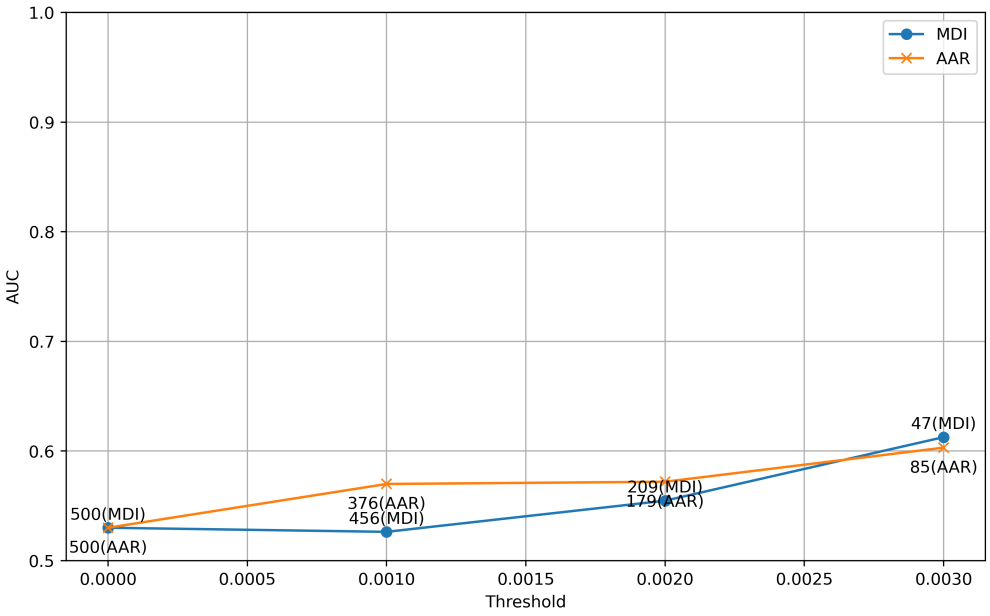
Source: self-provided

Figure 18 – Comparative analysis of the ALE-based AAR score and MDI from RF in reducing model complexity by minimizing the number of features. The dataset is from OpenML (id=1485) in a classification task. The number represents the number of remaining features on the dataset for each metric.



Source: self-provided

Figure 19 – Comparative analysis of the ALE-based AAR score and MDI from RF in reducing a logistic regression model complexity by minimizing the number of Features. The dataset is from OpenML (id=1485) in a classification task. The numbers represent the number of remaining features on the dataset for each metric.



Source: self-provided

4.5 SUMMARY

This chapter introduces four novel metrics - MUA, UAS, AUA, and AAR - derived from ALE framework that quantify the significance of features within predictive models. Each metric provides a clear definition that captures a specific aspect of feature relevance. Although primarily driven by the limitations of existing metrics in EDM, which frequently overlook the contributions of features in datasets with interrelated dependent variables, these new metrics also enhance the overall interpretability of supervised learning models. The experimental evaluation demonstrates the enhanced robustness of the ALE-based framework in such contexts when compared to established methodologies. Specifically, a first round of experiments illustrates the ability of the new metrics to effectively identify critical features in synthetic datasets engineered with varying generating functions and degrees of feature interdependence. These new metrics either surpass or match the performance of existing baseline measures. Specifically, the ALE-based metric generally yielded better results compared to SHAP and MDI, as they do not attribute significance to features that are irrelevant yet highly correlated with another relevant feature. Furthermore, they also outperformed PFI and cs_PFI in identifying multiple important features that are also correlated among them. When evaluating both properties, ALE metrics achieve better results.

We focus on RF and NN. Both models demonstrated good performance, with an close to the standard deviation of the theoretical noise added to the target variable, indicating that the models were capable of detecting underlying patterns

Further, empirical evidence from real-world datasets corroborates the findings from synthetic experiments, showcasing the ALE-based metrics' capacity to discern the main effects of variables. In a qualitative experiment utilizing an educational dataset and domain expertise to assess the expected relevance of key variables, ALE-based scores yielded better results than the baseline. They accurately identified and isolated the relevance of variables without attributing significance to potentially irrelevant features solely due to their correlation with relevant ones.

Moreover, the chapter discusses the potential pitfalls of using SHAP-based scores, particularly when computed through the computationally efficient TreeShap method for tree-based algorithms. It can lead to misleading interpretations due to its constraints to avoid extrapolation based on the tree structure of the explained model.

Finally, the potential of the proposal for feature selection and dimensional reduction tasks was demonstrated using openly real-world datasets, including one from the educational domain

and two high-dimensional datasets which are commonly used in feature selection benchmarking. The experiments showcased the ALE framework as a reliable, model-agnostic alternative for this purpose, capable of reducing model complexity with minimal performance degradation comparable to the baseline, sometimes achieving significantly better results. Specifically, the ALE-based score achieves better or comparable results when contrasted with the widely used feature importance of random forest (MDI) in various scenarios. The ALE could remove irrelevant variables while increasing or keeping the model performance even in scenarios inherently biased in favor of the baseline.

5 CASE STUDY - BRAZILIAN SECONDARY EDUCATION ASSESSMENT

Building on the insights from the previous chapters, which demonstrated the utility and reliability of ALE in clarifying the roles of features in models performing well amidst non-independent data, this chapter illustrate a case study employing the proposed ALE-based metrics. These metrics are applied to assess the impact of various features on student performance, as well as to trace the evolution of these influences over time.

The adoption of scores offers a more apt alternative compared to other explanation types, especially in the context of analyzing feature effects over an extended period within complex multivariate scenarios (FILHO; BRITO; ADEODATO, 2023a). Through this reporting approach, educational practitioners can gain a clearer understanding of how numerous features fluctuate over time, thereby improving the interpretation of models to support data-driven decision-making.

Scores provide a succinct, overarching measure of the model's output, distilling the essence of complex relationships into a single, interpretable metric. This approach is particularly advantageous when the primary goal is to gain an initial, high-level understanding of the model's behavior across multiple dimensions. While scores may not offer the nuanced details of feature-specific effects, they serve as an effective starting point from a human-centric perspective for further, in-depth analysis. Specifically, in this study case, using a unique score to represent the overall contribution of features facilitates a more transparent and comprehensive understanding of the role of many features over time.

5.1 INTRODUCTION

A prominent application of educational data mining is in exploring LSA. The advent of ML as an alternative to traditional statistical models, which have been prevalent in policy-oriented research since Coleman's 1968 study (COLEMAN, 1968), marks a significant shift in this domain.

The modernization of LSAs has notably improved data collection regarding educational system performance (HERNÁNDEZ-TORRANO; COURTNEY, 2021). Beyond performance metrics, LSAs gather information on the education system, including students' socio-demographic and school characteristics. A prime example is the Programme for International Student Assessment

(PISA), which offers a global perspective on secondary education learning outcomes (Varkey Foundation, 2018). Following international efforts, national LSAs have also played a significant role in the evaluation and improvement of their educational systems (JOHANSSON, 2016).

The availability of vast, structured educational data has spurred researchers' interest in more flexible methods that overcome the limitations of traditional statistical techniques. These limitations include constraints in handling a large number of variables without prior assumptions about the data (MARTÍNEZ-ABAD; GAMAZO; RODRÍGUEZ-CONDE, 2018; MASCI; JOHNES; AGASISTI, 2018). These studies often utilize the supervised learning paradigm to predict LSA achievement (output) based on contextual variables from LSA questionnaires (inputs). Characterizing this input-output mapping enables the identification of sources for educational policies and practices associated with academic achievement. This characterization has been expressed either through the models themselves (GOMES; JELIHOVSCHI, 2020; FILHO; ADEODATO, 2019) or through post-hoc explanation techniques (GABRIEL; SIGNOLET; WESTWELL, 2018; SCHILTZ et al., 2018).

This chapter contributes to such research, aiming to identify and track key features related to student performance. The primary goal is to demonstrate the applicability of ALE-based metrics introduced in the previous chapter. Additionally, this chapter offers new contributions: 1) It defines a new model-agnostic process for reporting trends in the predictive contribution of features in repeated cross-sectional data, and 2) it presents a case study within the context of the Brazilian secondary education system, providing new insights for educational literature.

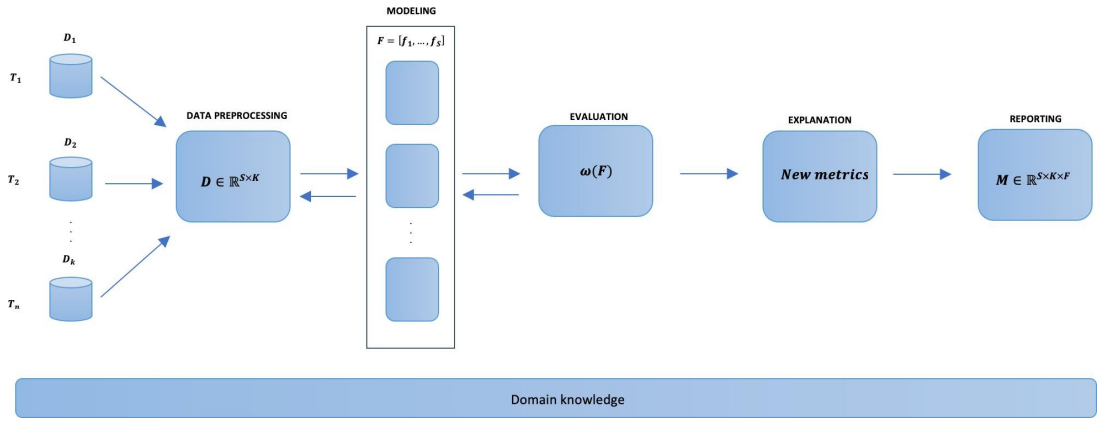
5.2 METHODOLOGY

To effectively report the trends in how contextual features predict LSA outcomes, the educational production function framework (BOWLES, 1970) and repeated cross-sectional analysis (BUCK; ERMISCH; JENKINS, 1995) were utilized.

To effectively report the trends in how contextual features predict LSA outcomes, this chapter utilizes the educational production function framework (BOWLES, 1970) and conducted a repeated cross-sectional analysis (BUCK; ERMISCH; JENKINS, 1995). This analysis was guided by the well-established Cross-Industry Standard Process for Data Mining (CRISP-DM) (CHAPMAN et al., 1999)and complemented by the Domain-Driven Data Mining (D^3M) approach (CAO; LIN; CHENGQI, 2005; CAO, 2009). This process involves multiple phases integral to the. Here, domain knowledge plays a pivotal role throughout, ensuring that the developed models are not

only statistically sound with a good predictive performance, but also actionable and relevant in real-world applications.

Figure 20 – The case-study methodology



source: self-provided

Figure 20 outlines the case-study methodology. Specifically, consider a dataset $D = [d_1, \ldots, d_k]$ with contextual features $X = [x_1, \ldots, x_S] \in \mathbb{R}^{S \times K}$, where K represents the number of time steps (LSA waves), and S is the number of features. The data preprocessing step involves standardizing D across K to ensure uniform meaning and measurement for X. In the modeling step, a compact set of predictive functions $F = [f_1, \ldots, f_S]$, where each $f(X) = Y$ and $Y = 0, 1 \in \mathbb{R}^K$, is constructed and applied sequentially. Let $x_{i,k}$ denote the input feature i at time K, and $X_{i,:} \in \mathbb{R}^K$ be an independent time vector of feature i. The most effective models in F during the evaluation step, for each feature vector at time K ($X_{:,k} \in \mathbb{R}^S$), will yield a corresponding vector of scores in the explanation step. The reporting step culminates in a score matrix M of dimensions $S \times K \times F$, offering valuable insights into the dataset. Incorporating diverse model types in F can enhance understanding of the data generating process.

In general, the comparison of feature relevance from different models, even with the same metric, may be inappropriate and requires caution (FISHER; RUDIN; DOMINICI, 2018). Important variables for one well-performing model may be unimportant for another model. On the other hand, this practice may help to make the analysis even more insightful if combined with domain knowledge as defined in (FILHO; BRITO; ADEODATO, 2023a) as "Rashomon" set analysis.

5.2.1 Background and data source

The National Secondary School Exam (ENEM) was conceived to assess the quality of Brazilian secondary schools based on a student evaluation of the test. In 2009, it was reframed to the Item Response Theory, thereby making it comparable over time. The ENEM was established as the mechanism for student admission to higher education. Hence, the ENEM has become a reliable, rich data source regarding the Brazilian secondary system. The ENEM microdata contains student socio-economic-cultural information and their grades achieved in the test. Together with the national school census (CE), which details the conditions of Brazilian schools, from physical infrastructure to faculty information, they build a robust, extensive database of Brazilian secondary education. Both databases were publicly available on the INEP website[1]. The period covered is from 2009 to 2019. The dataset refers to over 40 million students in thousands of schools across the country. However, only students in the last year of public secondary education were considered.

As the school "ID" is the primary key in combining the ENEM and school census datasets, all students who did not attend schools that were identified to be in the survey across years were removed. This led to a large decrease of about 80% of the dataset. Additionally, the following criteria defined the scope.

1. Students were not included if they were not in the last year of municipal or state public secondary schools..
2. Students were not included if they did not follow a regular curriculum
3. As a double-check, students not in the most probable age range meeting criteria 1 and 2 (17-19 years old) were also eliminated.
4. Only schools with ten or more students were selected..
5. To ensure that all schools had at least a minimum infrastructure to function, schools with no electric energy, sanitation, or piped water were excluded.

[1] https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/microdados

5.3 PREPROCESSING

The CE and ENEM datasets have suffered several changes over time. As an illustration, there were 293 variables in the ENEM questionnaire in 2009, while in the following year, 2010, only 57. Added to this difference in the number of variables collected, there were also changes related to the representation of the variables, such as 1) features were binary for some years and categorized by quantity for others; 2) categories were represented by numbers in some years and by strings in others; and 3) in the case of some variables, categorical features were transformed to binary features.

It was important to standardize the data to overcome these issues and to allow the comparison of the findings over the years. First, only variables presented in all waves were used. Next, the data were standardized in regard to content and meaning. A variable with less information was used as a reference for mapping the others. For instance, if a variable was binary in one year and multiple categorical in others, the binary version was adopted for all years. The income features were normalized using a contemporary minimum wage. The variables related to the use of technological devices were individually treated. For example, before 2019, the available information on technology devices at school was measured by just one variable (student's computer), in 2019, the questions also asked about notebooks and tablets - these were turned into a single indicator. Missing values for all variables were analyzed separately, since there were not many of them, and were given the model value. Alternatively, the mean of the non-missing values was used for those that did not have a clear explanation. The chosen variable to indicate the outcomes was the arithmetic mean of the students' test scores in all areas of knowledge covered in the test. To reduce the influence of outliers, all numerical variables were normalized for each year separately, using the α-winsorized values of the distribution ($alpha/2 = 0.025$ at each tail) as their minimum and maximum.

5.3.1 New features

Some features frequently brought to the fore in discussions on the quality of secondary Education (OCDE, 2013), especially those related to the faculty, are not initially present in the databases. Nevertheless, some may be derived from the information available on datasets and four new features were created: a) Faculty appropriate training (measuring the ratio of

teachers with the appropriate background for the subject they teach)[2] b) Number of jobs (the average number of schools where teachers work) c) Faculty pedagogical training (those in the faculty with pedagogical training), d) Faculty <DOMAIN> (four derived features indicating the ratio of faculty teachers per each knowledge covered by ENEM), e) Faculty work overload (the ratio of teachers per number of subjects covered in school), and f) Faculty education (weighted average of teacher educational level, Ph.D. – higher weight, Bs. lower weight). The source information for creating the variables was available in the CE datasets, which is able to identify all classes and subjects assigned to teachers for each school together with their backgrounds. To verify whether the teacher has the correct background in order to create the Faculty appropriate training index, an auxiliary table released by INEP was used. Lastly, forty-one (41) input features compound the final dataset, as listed in Appendix B, as long as their descriptive statistics.

5.3.2 Evaluating comparability over time

A cross-validation scheme assessed whether data followed a uniform distribution over time and were statistically comparable. This simple experiment evaluated the performance of the model when data corresponding to one year was omitted from the estimation procedure. Therefore, each year was used to compute the model performance while the others were used to train. The average standard deviation of the AUC was 0.02 of the RF, indicating that the datasets are stable over time, thereby enabling a good generalization.

5.3.3 Experimental setting

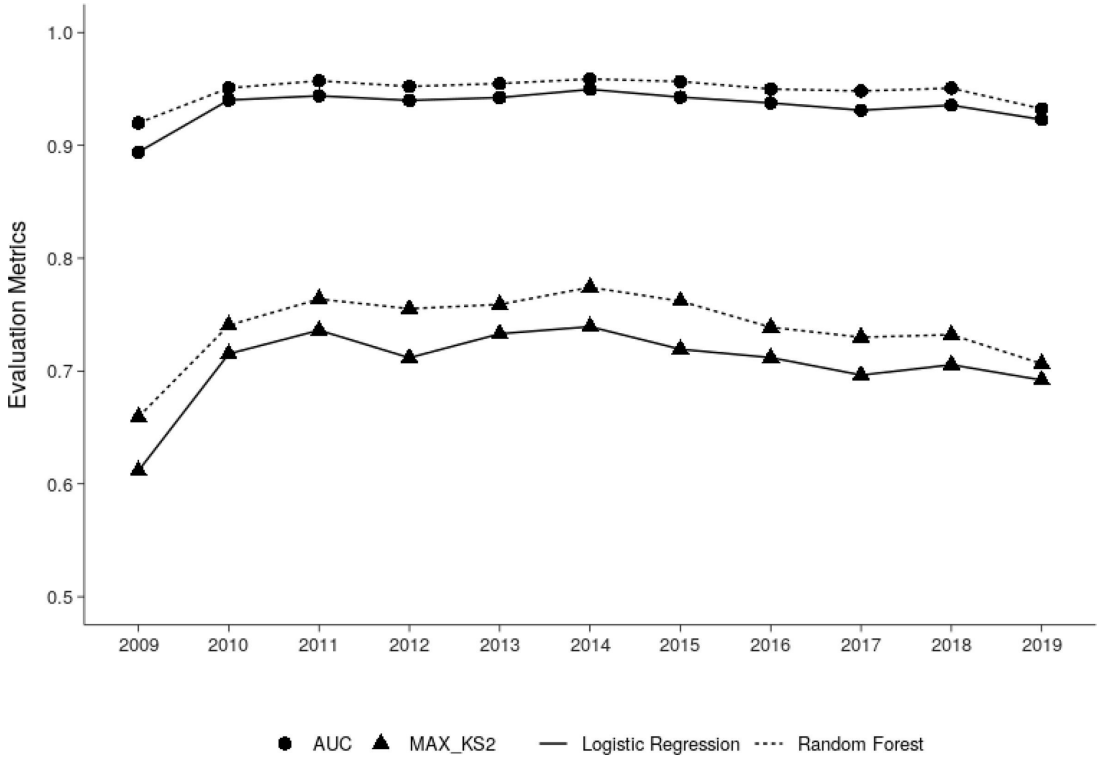
To facilitate comparison, the and RF algorithms were employed. The is simple and additive, while the RF may account for interactions without assuming any prior distribution for the data. It is expected that the combined analysis of explanations derived from these different algorithms, if well-performed, may be insightful for knowledge extraction. The performance of the models, as well as the feature contribution, was assessed by means of k-fold cross-validation (k = 10) for each LSA data cycle. All preprocessing and data analysis were performed with *Python 3.6*, using the *scikit-learn* library with default parameters.

[2] For each subject the weight "1" was assigned if its teachers had graduation in the relevant area and "0.5" if they did not. Also, the index was normalized by 13, the considered total of mandatory subjects in the Brazilian educational secondary curriculum (see: http://basenacionalcomum.mec.gov.br/historico/)

5.4 RESULTS

This section presents the results of Brazilian case study. With regard to evaluating the models, Figure 21 presents the AUC_ROC and the maximum distance between Kolmogorov-Smirnov curves(KS2_max)(KOLMOGOROV, 1933). Both metrics achieved good results, with RF outperforming LR with a slight difference in both metrics over the whole period (1% for AUC_ROC and 4% for KS2_Max in the period).

Figure 21 – AUC and KS2_max of logistic regression and random forest models



Source: self-provided

Figure 22 presents the feature contribution of both models by the maximum uncentered ALE (MUA). This metric indicates the maximum and isolated influence given the trained data of an underlying feature in predicting school achievement. Both models similarly highlight the well-known determinants of school performance during the whole period, such as *income (per capita)*, *race*, *mother's education*, *father's education*, and *students'age* regarding size and direction. This ratifies the most substantial influence of these features on school performance, as has been well-established in the education literature (CALDAS; BANKSTON, 1997; COLEMAN, 1968; COLEMAN, 2019). It also confirms previous Brazilian studies (CARNOY; ROSA; SIMõES, 2022) that have used a somewhat different methodology. The *per capita* income significantly

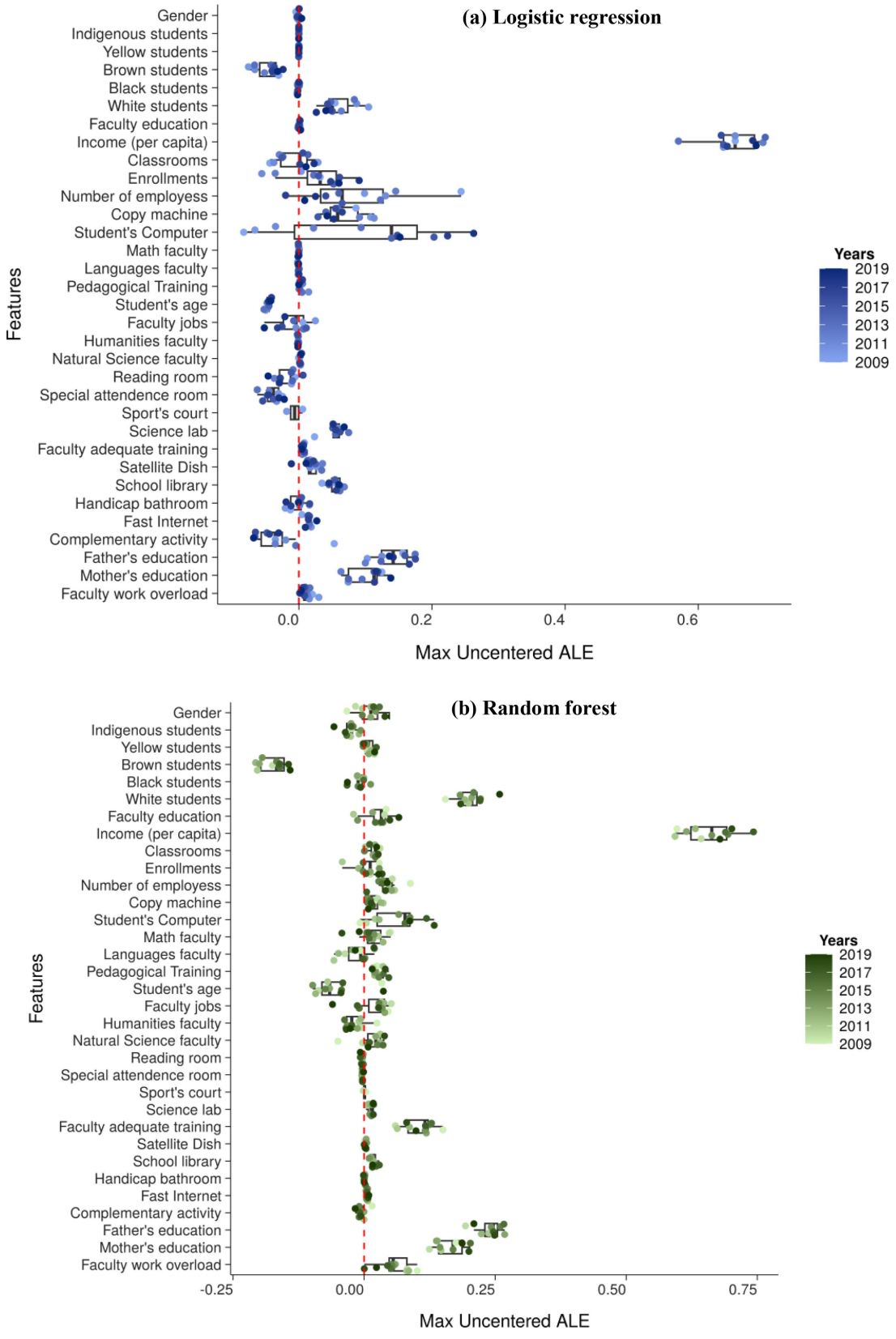
influences the likelihood of a school being categorized in the third quartile, with an approximate factor of 0.6. This is roughly three times greater than the impact of parents' education, which holds an estimated factor close to 0.2, over the entire period under study. *Race* has also been significantly related to achievement. While brown students are linked to the school in the lower quartile, white students are linked to higher achievement. Moreover, the feature that indicates the number of computers available to students (*Student's computer*) seems to be a favorable policy with an upward trend (darker points are far from line zero) in classifying schools in the higher quartile, especially in the LR models.

5.4.1 Faculty features

By using the combined analysis, it becomes insightful to explore why certain variables were highlighted in one model but not in another. A LR model, when devoid of explicit interactions, essentially operates as an additive model, capable of emphasizing only the main linear effects of a feature. This appears to be the case for *Faculty education* in Figure 22. Contrary to existing educational studies (CALDAS; BANKSTON, 1997; DARLING-HAMMOND, 2000), *Faculty education* lacks significant explanatory power in LR (Figure 22) models. The *Faculty education* index, which includes the proportion of teachers with Bs., Specialization, Master's, and Ph.D. degrees, has increased (from 0.13 in 2009 to 0.18 in 2019) but may not exert uniform impact across all schools. This variation could stem from the differential impact of the index in conjunction with other model features that the RF model could slightly capture. This variation might result from inherent relationships or the diverse efforts of various states in enhancing teacher education levels across the country (FILHO; BRITO; ADEODATO, 2023b). Although there is no feature explicitly representing states, the non-linear effect captured by the RF could be indicated through the proxy behavior of other features.

Yet in the faculty context, the similar positive size of *Faculty work overload* (equal to 1 when all teachers teach just one subject) and *Faculty adequate training* (equal to 1 when all teachers have an appropriate background in the subjects they teach) in both models suggest that teaching more than one subject is not a problem if they have the appropriate training. On the other hand, the importance of the number of schools where teachers work (*Faculty jobs*) decreases in both models, passing from positive to negative effects. This behavior conforms to earlier qualitative results (BARBOSA, 2013; SOUZA; OLIVEIRA; NASCIMENTO, 2020).

Figure 22 – Feature effects size measured by the MUA= from LR (a) and RF (b) in classifying Brazilian secondary schools using the ENEM score as a performance metric from 2009 to 2019

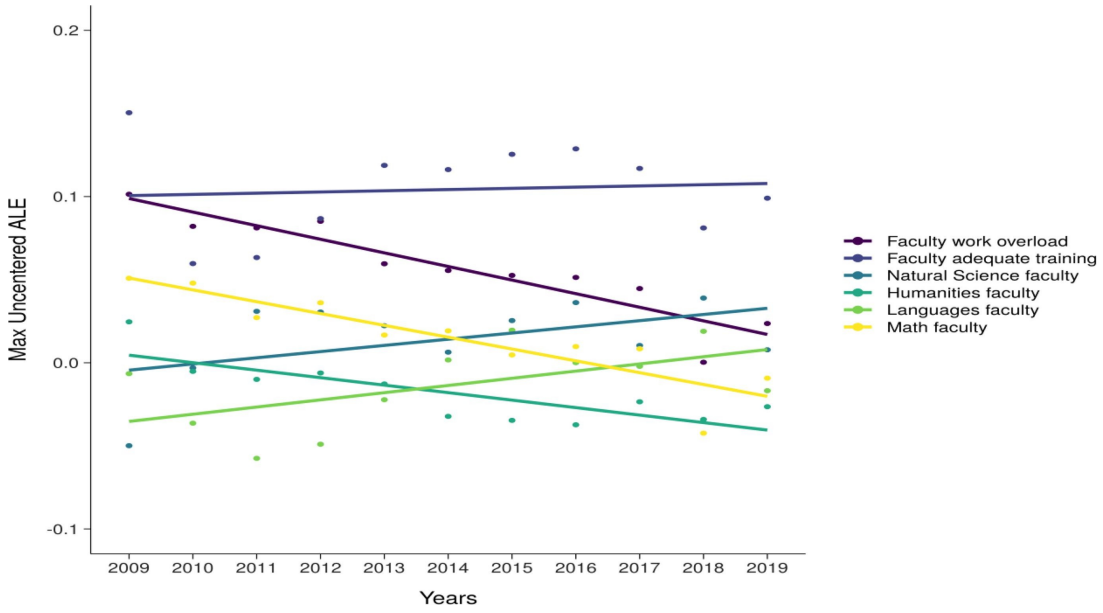


Source: self-provided

5.4.2 Closer examination of the faculty features

One relevant way to better understand the specific scenarios is to separate some of the features to observe them more deeply. For example, in Figure 22, the RF model highlighted some variance among the different faculty areas. These effects seem to derive from a nonlinear relationship with data since the LR model could not find them. Therefore, an additional line plot (Figure 23) may help to figure out how these features relate. In addition, *Faculty appropriate training* and *Faculty work overload* might enhance the analysis even more. As expected, behavior among features related to faculty domains is uneven since they are frequency encoded. Thus, a linear regression was embedded into the plot to obtain a better perception of their trends. The importance of *Natural Science faculty* and *Languages faculty* has increased over time, while the *Math faculty* has taken the opposite direction together with the *Humanities Faculty*. The indexes *Faculty appropriate training* and *Faculty work overload* demonstrate positive importance over the whole period, however with different behaviors. The former is stable with greater importance and seems to be strongly correlated with the *Languages faculty* and *Natural science faculty*. The importance of the latter has been decreasing, as in the case of the *Math faculty* and *Humanities faculty*.

Figure 23 – Specific feature effects size measured by MUA related to the faculty.

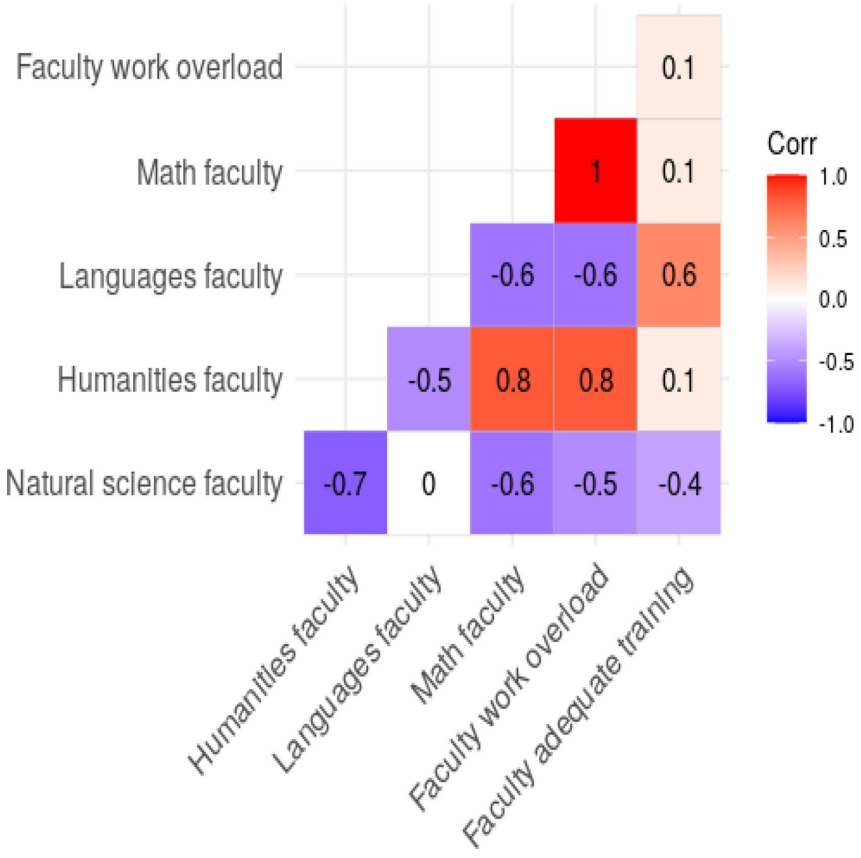


Source: self-provided

Figure 24 reveals the behavior detailing the Pearson's coefficients among the importance of these features. These findings suggest that the importance of the *Math faculty* and *Humanities*

Faculty (History, Geography, Philosophy, and Sociology) could be decreasing in some schools due to a heavy workload. Although this is difficult to verify in the training data since it is derived from an unknown nonlinear relationship, the data illustrates that 30% of humanities teachers teach more than one discipline. Moreover, 90% of the schools in the final dataset do not report a physics teacher during the period. Thus, it is probable that a math teacher might have to cover this subject, as already reported in (SANTOS; CURI, 2012). On the other hand, Foreign Languages, Arts, and Physical Education (subjects from the *language domain*) have the lowest indexes for the Faculty adequate training (0.5, 0.6, and 0.7 respectively, against an average of 0.8 for the others), and any endeavor to boost it could be contributing to the slight increase in the importance of *Faculty adequate training* over time. However, these results require caution and more studies with domain expert validation.

Figure 24 – Matrix correlation of MUA to the specific set of features related to the faculty.



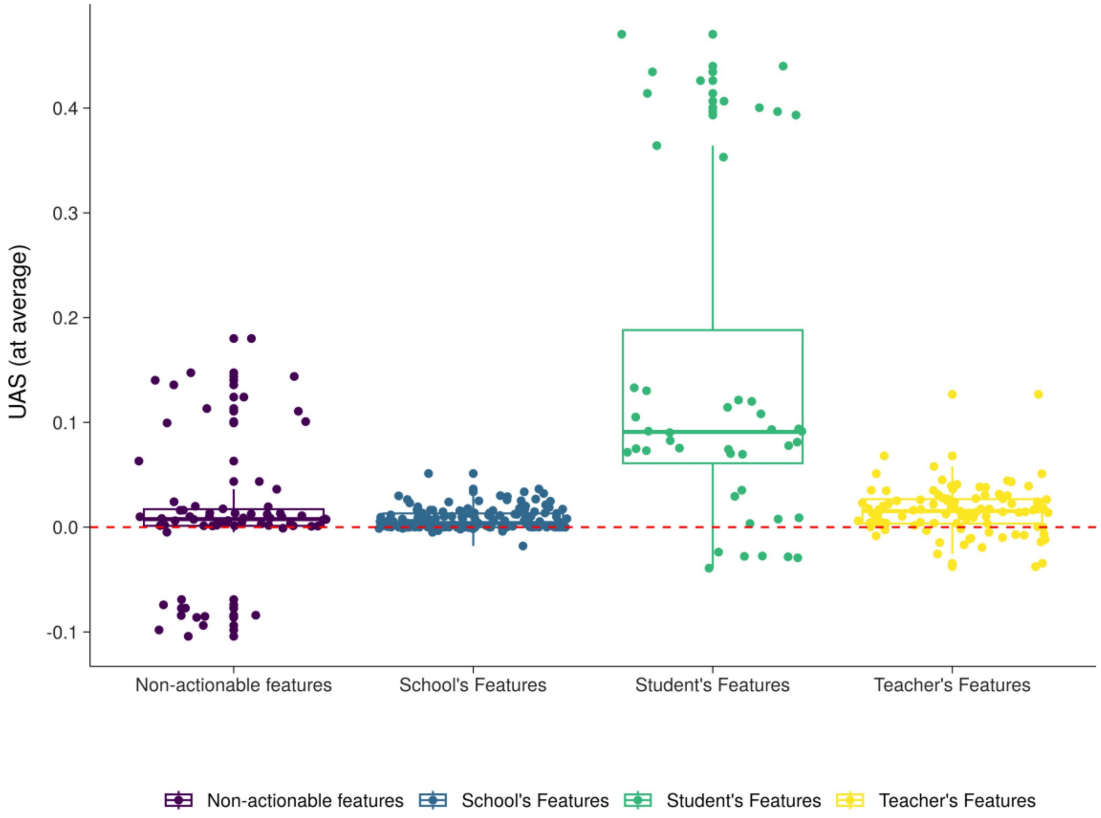
Source: self-provided

From another perspective, Figure 25 explains how different combinations of features have influenced school performance during the entire period. It demonstrates the influence of non-actionable features (*race* and *gender*), school features (related to infrastructure), faculty features, and student features (*parents' education* and *income*) in the period by using a box plot.

The metric used is the UAS (on average) from RF models.

The group of student features has more potential to improve educational performance, followed by the non-actionable features, which have a higher potential for damage. In general, the school features have a low influence, while the teacher features have a limited, although relevant, importance in improving the quality of schools.

Figure 25 – Feature effects size measured by UAS (on average) by group regardless of the time.



Source: self-provided

5.5 SUMMARY

This chapter illustrated the significance of the proposed metrics in a real-world application: Identify and track the relevance of features in secondary educational outcomes in the Brazilian public educational system. The application was defined as a repeated cross-sectional analysis, which required a definition of a new process since, to our knowledge, there are no other examples of this trend analysis using supervised learning.

The findings of this study may also provide valuable insights for researchers interested in Brazilian secondary education. While it is well-established in the literature that factors such as

income, *age*, *race*, and *parent's education* have a strong impact on educational achievement, it is still an open question as to how these factors evolve and influence achievement over the period of study. Moreover, student computers at school, which presents mixed findings regarding its effects in the literature, have been highlighted in this chapter as one of the most effective policies regarding variables related to schools. Additionally, the knowledge extraction process leverages hypotheses that have either only been discussed qualitatively or not at all. For example, the study has suggested that improving *Faculty education*, *Faculty appropriate training* (especially for language teachers), and addressing Faculty workload could be important for improving secondary school achievement. Nevertheless, this investigation is not able to provide causal conclusions, and further research by domain experts is needed to confirm the findings.

6 CONCLUSION

This chapter presents the concluding remarks of this thesis, emphasizing its contributions to the fields of XAI and supervised learning, clearly illustrated within the context of EDM. Moreover, this chapter delineated the inherent limitations encountered during the research and discussed a few themes that future works should focus on.

6.1 CONCLUDING REMARKS

The application of ML in data analysis represents a significant research opportunity. The recent growth in data collection enables the use of ML techniques that are versatile enough to capture intricate data relationships. However, using ML for this purpose presents challenges, as these techniques may lack transparency in how they adjust to the data to make predictions. Developing model explanations can help overcome this issue, enhancing knowledge extraction from complex data and supporting more informed, strategic data-driven decisions.

The objective of this thesis is to evaluate and propose methods to derive global explanations in the context of supervised ML. Specifically, it focuses on providing a more unbiased and isolated understanding of feature roles in predictive models.

With advancements in ML and XAI, many researchers increasingly use these tools to extract insights from data. In the educational domain, this kind of research is widely employed to identify significant predictor variables aiming to foster the educational ecosystem and advance the field of EDM. However, a comprehensive literature review reveals that conventional XAI methods used in EDM have interpretability limitations when applied to dependent data, frequent situation in educational contexts, which is a key motivation for using ML models. These methods often rely on assumptions that are frequently unmet, limiting the insights gained from the data.

The problem arises from the tendency of these methods to "extrapolate" existing data relationships when computing the contribution of variables within a predictive function. Additionally, existing techniques often do not align with the primary objectives of educational practitioners when analyzing effect sizes in statistical analysis. For instance, the score-based explanations often focus on a variable's impact on model performance rather than on its directly contribution to the model's predictions.

The ALE technique have been proposed as a prominent XAI technique that can minimize the extrapolation problem when computing feature effects. However, there is a limited assessment of these ALE capabilities when compared to other widely used explanation techniques. Additionally, ALE has only been defined for visualizing feature effects with limited discussion of use of ALE to derive score-based explanations.

Witin this context, this thesis goes to this problem formulating two main research questions:

RQ1 - How do widely used feature effects techniques compare with ALE in accurately identifying true feature effects considering different inter-data dependencies?

RQ2 - How effectively can score-based explanations derived from the ALE framework report individual and isolated attribution of the features in terms of their magnitude and direction compared to existing methods?

In response to RQ1, Chapter 3 benchmarked the most used explainable techniques on EDM and ALE, a recent contribution of XAI that employes some constraints to avoid data extrapolation. The benchmarking indentify ALE as the most suitable technique to report features effects when features are correlated. Building on this finding, Chapter 4 answer RQ2 by proposing a set of ALE-based metrics to enhance the clarity and utility of the supervised model explanations focused on overall variables effect size. Chapter 5 then demonstrated the practical application of these metrics in a real-world educational context.

6.1.1 Summary of contributions

6.1.1.1 A benchmarking of feature effects techniques

Answering RQ1 and aiming to provide empirical evidence regarding the robustness of ALE as compared to baseline methods in dependent data, a benchmarking of the feature effect technique was established. To the best of our knowledge, this is the first quantitative comparison of the accuracy of PD plots and ALE against a ground truth. Also, other techniques widely used in literature ME and SHAP were included, enhancing the benchmarking for a broader comparison. A new comparison metric, the ABX, was introduced to measure the area between the true and explained features.

The ALE outperformed in accurately recovering the feature effects in all scenarios under dependent data. Also, the experiments highlight the potential risk of explaining highly flexible algorithms, such as neural networks, using techniques that extrapolate the manifold even on

independent datasets. This phenomenon is already known and has also been simulated in this thesis (Chapter 4, Figure 16), but, to our knowledge, it has not yet been identified in empirical experimentation. These results may aid the XAI field, which can use the benchmarking framework as well for applied researchers, especially on EDM, that could identify the pitfalls of the most currently used XAI techniques to derive insights about the data.

6.1.1.2 New scores of feature effects size

Motivated by the robustness of ALE and the limitations of existing scores of feature importance, this contribution introduces four new model-agnostic measures of variable effect size based on ALE. These measures are designed for enhanced interpretability of feature roles in predictions, especially in scenarios involving dependent data. Three of these metrics offer distinct single-explanation perspectives, elucidating the extent and direction of feature effects in relation to the target variables. Each metric presents a unique interpretation, adding depth to the understanding of feature influence. The fourth metric offers a normalized ranking of feature impact, facilitating their comparison across different datasets and models.

In evaluations, these features exhibit similar or superior performance compared to existing metrics in the XAI literature, proving effective in identifying key variables in both synthetic and real datasets. The metrics can be employed either in cross-validation settings for more robust estimates or bootstrap, allowing yield confident intervals to account for variability and uncertainty inherent to the data and the model.

Calculating scores using ALE introduces specific limitations inherent to model-agnostic methods. The generalizability of the results largely depends on how representative the sample is of the population. Furthermore, an important aspect of ALE is its computation by segments rather than analyzing the entire dataset at once. As a result, it is necessary to consider the actual data sample used along with how each provided metric is computed for an appropriate interpretation and generalization of the explanation outputs. Despite this limitation, the local nature of ALE has partial benefits for the purpose of the metrics. The ALE ensures that explanations remain faithful to the relationships in the data. Additionally, ALE enables the computation of the isolated variable effects and their interactions within this interval.

However, relying solely on one score to represent the entire distribution produced by the ALE function may be problematic and conceal important aspects of the shape of variable effects, especially in cases where they are noisy or have been calculated based on a limited

number of data points. Following the ALE limitation, the metrics cannot also be computed for categorical variables without order relation. Finally, although ALE permits the computation of interaction effects, which have also been defined in the context of scores in this thesis, the empirical experimentation focused solely on assessing them by computing the main effects of feature size.

6.1.1.3 A empirical trend analysis of Brazilian secondary schools determinants

To demonstrate the usefulness and the meaningful of the proposed scores in the exploration of educational data, an empirical case study was presented. The real scenario seeks to identify and track the determinants of Brazilian public education from 2009 to 2019. To the best of our knowledge, we are the first to explore the impact of contextual features on educational outcomes through supervised learning over time. Previous studies have handled this problem only at a single point in time. While (FRANCO et al., 2020) used multiple years of ENEM data in Brazil to conduct similar research, their work did not aim to make results comparable, which does not allow for tracking the feature effects size over time.

Moreover, the defined process is also a contribution of this thesis to researchers interested in conducting repeated cross-sectional analysis using supervised learning. The process is flexible enough to be applied to any domain.

The findings of this case study also provided valuable new insights for researchers interested in Brazilian secondary education. Lastly, it should be noted that the preprocessed and standardized data used in this analysis is an additional contribution of this thesis and is available (FILHO, 2022) for other researchers interested in the quantitative analysis of Brazilian secondary education.

6.2 FUTURE WORKS

As the use of ML increases, so does the demand for interpretability, making XAI a rapidly growing field with numerous new interpretation methods being introduced. In this thesis, rather than developing a new method, the focus is on applying an established method, ALE, to the context of reporting global feature contributions in the educational domain. This approach aims to deepen and extend our understanding of its potential to enhance the interpretability of educational models.

The contribution of this thesis can be improved and further explored in future works. For instance, the benchmarking process of Chapter 3 could be expanded to include more algorithms as well as a wider range of data scenarios, including the presence of outliers and missing values. Similarly, these variations could be applied to the evaluation of the new metrics of Chapter 4. Specifically, while the potential of these metrics as a feature selection method was presented, detailed scrutiny was beyond the scope of this thesis. Consequently, there remains a need for empirical evidence to establish their effectiveness fully in this direction.

Moreover, from a broader perspective, by concentrating on using XAI in EDM to obtain global explanations, I believe that investigating the following related topics could substantially advance the field.

6.2.1 True to the model, true to the data, and true to the context

A central challenge in explainable methods lies in balancing fidelity to the model and the data. This tension forms a core part of this thesis's motivation. Overemphasis on the model can lead to unreliable explanations due to neglect of data relationships. Conversely, focusing solely on data may preclude leveraging complex functions that fit the data ALE have emerged as a promising solution to this dilemma, especially in the context of supervised learning for knowledge extraction from data. However, we argue there is another perspective that new XAI techniques must be aware of in education: the context perspective.

Contextual understanding involves comprehending how features semantically affect observations. While ALE and other global techniques effectively identify varying feature effects across their value range, there is limited exploration in contextualizing which observations correspond to each feature value. A potential breakthrough could be a technique that uncovers heterogeneous feature effects, pinpointing relevant groups based on their distinct responses to a feature within the model. This approach aligns with the existing literature on model fairness in EDM, which aims to identify the poor performance of ML models in sensitive groups. Viewing it through the lens of XAI at the feature level could significantly enhance the utility of EDM in providing valuable insights from educational data.

6.2.2 Data dependence is the real world

Historically, traditional linear models have been the most widely used method for extracting knowledge from data in the education domain. When specifying these models, the interpretation of coefficients is seen as the effect of a variable on the dependent variable. This perspective has guided researchers in supervised learning, with the primary aim of harnessing ML's powerful pattern recognition capabilities while maintaining a level of interpretability that tries to mimic traditional statistical models. Within this context, this thesis endeavors to introduce alternatives that better address the complexity of data dependence on educational datasets, dealing with the trade-off between being true to the data and true to the model.

However, a different approach to dealing with data dependence, treating it as an inherent part of the ML paradigm and also from the real world, can extend the meaningfulness of XAI in EDM. ML is inherently associative, and this property can be leveraged to gain deeper insights into the data-generating process. Instead of focusing solely on isolating the effects of individual features or their interactions, ML allows researchers to explore the network of relationships within the data. This approach recognizes the complexity and interconnectedness of educational environments as part of the problem. It enables a more holistic view, shifting the focus from measuring isolated feature effects to understanding the network of relationships within the data. For example, this could involve exploring how and when different aspects of the school environment interact with student backgrounds or how policy changes ripple through various layers of the education system.

6.2.3 Beyond a one-size-fits-all

This thesis has established the ALE as a robust XAI technique for reporting global feature contributions, particularly in the context of dependent data. While the potential of ALE is established, and the ALE framework has been further explored in this study in order to enhance data interpretability, it is important to recognize that XAI in and ML is inherently exploratory. No single method uniformly suits all scenarios. The efficacy of combining techniques, both for complementary insights and ensemble approaches, presents a significant area for exploration. This concept has been demonstrated in (FISHER; RUDIN; DOMINICI, 2018), where multiple models were utilized to generate more reliable scores. Furthermore, the adoption of frameworks that integrate various techniques aiming for more model-specific explanations has shown promise,

as discussed in (LI et al., 2019). We believe that the synergistic application of multiple XAI techniques and paradigms, particularly aimed at enhancing the quality of insights in educational contexts, represents a promising direction for future research.

REFERENCES

ABAD, F. M.; LÓPEZ, A. A. C. C. Data-mining techniques in detecting factors linked to academic achievement. *School Effectiveness and School Improvement*, Routledge, v. 28, n. 1, p. 39–55, 1 2017. ISSN 0924-3453. Available at: <http://dx.doi.org/10.1080/09243453.2016.1235591https://www.tandfonline.com/doi/full/10.1080/09243453.2016.1235591>.

ADEODATO, P. J. L. Data Mining Solution for Assessing Brazilian Secondary School Quality Based on ENEM and Census Data. In: *International Conference on Information Systems & Technology Management - CONTECSI*. [s.n.], 2016. p. 1112–1124. Available at: <http://www.contecsi.tecsi.org/index.php/contecsi/13CONTECSI/paper/view/3818>.

AFZAAL, M.; NOURI, J.; ZIA, A.; PAPAPETROU, P.; FORS, U.; WU, Y.; LI, X.; WEEGAR, R. Explainable AI for Data-Driven Feedback and Intelligent Action Recommendations to Support Students Self-Regulation. *Frontiers in Artificial Intelligence*, Frontiers Media S.A., v. 4, 11 2021. ISSN 2624-8212. Available at: <https://www.frontiersin.org/articles/10.3389/frai.2021.723447/full>.

AGUIAR, E.; LAKKARAJU, H.; BHANPURI, N.; MILLER, D.; YUHAS, B.; ADDISON, K. L. Who, when, and why. In: *Proceedings of the Fifth International Conference on Learning Analytics And Knowledge*. New York, NY, USA: ACM, 2015. v. 16-20-March-2015, p. 93–102. ISBN 9781450334174. Available at: <https://dl.acm.org/doi/10.1145/2723576.2723619>.

ALDOWAH, H.; AL-SAMARRAIE, H.; FAUZY, W. M. Educational data mining and learning analytics for 21st century higher education: A review and synthesis. *Telematics and Informatics*, Elsevier Ltd, v. 37, p. 13–49, 4 2019. ISSN 07365853. Available at: <https://linkinghub.elsevier.com/retrieve/pii/S0736585318304234>.

ALGHAMDI, A.; BARSHEED, A.; ALMSHJARY, H.; ALGHAMDI, H. A Machine Learning Approach for Graduate Admission Prediction. In: *Proceedings of the 2020 2nd International Conference on Image, Video and Signal Processing*. New York, NY, USA: Association for Computing Machinery, 2020. (IVSP '20), p. 155–158. ISBN 9781450376952. Available at: <https://doi.org/10.1145/3388818.3393716>.

ANDERSON, C. The end of theory: The data deluge makes the scientific method obsolete. *Wired magazine*, v. 16, n. 7, p. 16–17, 2008. Available at: <https://www.wired.com/2008/06/pb-theory/>.

ANDRADE, R. J. d.; SOARES, J. F. O efeito da escola básica brasileira. *Estudos em Avaliação Educacional*, v. 19, n. 41, p. 379–406, 12 2008. ISSN 1984-932X. Available at: <https://publicacoes.fcc.org.br/eae/article/view/2067>.

ANDREWS, M. *The Immortal Science of ML: Machine Learning & the Theory-Free Ideal*. [S.l.], 2023. Available at: <https://www.researchgate.net/publication/371982028_The_Immortal_Science_of_ML_Machine_Learning_the_Theory-Free_Ideal>.

APLEY, D. W.; ZHU, J. Visualizing the Effects of Predictor Variables in Black Box Supervised Learning Models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, v. 82, n. 4, p. 1059–1086, 9 2020. ISSN 1369-7412. Available at: <https://academic.oup.com/jrsssb/article/82/4/1059/7056085>.

ARAQUE, F.; ROLDÁN, C.; SALGUERO, A. Factors influencing university drop out rates. *Computers & Education*, v. 53, n. 3, p. 563–574, 11 2009. ISSN 03601315. Available at: <https://linkinghub.elsevier.com/retrieve/pii/S0360131509000815>.

ARYA, V.; BELLAMY, R. K. E.; CHEN, P.-Y.; DHURANDHAR, A.; HIND, M.; HOFFMAN, S. C.; HOUDE, S.; LIAO, Q. V.; LUSS, R.; MOJSILOVIĆ, A.; MOURAD, S.; PEDEMONTE, P.; RAGHAVENDRA, R.; RICHARDS, J.; SATTIGERI, P.; SHANMUGAM, K.; SINGH, M.; VARSHNEY, K. R.; WEI, D.; ZHANG, Y. One Explanation Does Not Fit All: A Toolkit and Taxonomy of AI Explainability Techniques. 9 2019. Available at: <http://arxiv.org/abs/1909.03012>.

ASHRAF, A.; ANWER, S.; KHAN, M. G. A Comparative Study of Predicting Student's Performance by use of Data Mining Techniques. *American Scientific Research Journal for Engineering*, ASRJETS. ISSN 2313-4402. Available at: <http://asrjetsjournal.org/>.

ATHEY, S.; IMBENS, G. W. Machine Learning Methods That Economists Should Know About. *Annual Review of Economics*, v. 11, n. 1, p. 685–725, 8 2019. ISSN 1941-1383. Available at: <https://www.annualreviews.org/doi/10.1146/annurev-economics-080217-053433>.

BABIC, B.; GERKE, S.; EVGENIOU, T.; COHEN, I. G. Beware explanations from AI in health care. *Science*, v. 373, n. 6552, p. 284–286, 2021. Available at: <https://www.science.org/doi/abs/10.1126/science.abg1834>.

BAKHSHI, A. K.; AHMED, M. M. Utilizing black-box visualization tools to interpret non-parametric real-time risk assessment models. *Transportmetrica A: Transport Science*, Taylor & Francis, v. 17, n. 4, p. 739–765, 2021. Available at: <https://doi.org/10.1080/23249935.2020.1810169>.

BARBOSA, A. Implicações dos baixos salários para o trabalho dos professores brasileiros / Low salaries implications on brazilian teachers' work. *Revista Educação e Políticas em Debate*, v. 1, n. 2, 2013. Available at: <http://www.seer.ufu.br/index.php/revistaeducaopoliticas/article/view/21902>.

BAREISS, E. R.; PORTER, B. W.; WIER, C. C. Protos: an exemplar-based learning apprentice. *International Journal of Man-Machine Studies*, v. 29, n. 5, p. 549–561, 1 1988. ISSN 00207373. Available at: <https://linkinghub.elsevier.com/retrieve/pii/S0020737388800129>.

BELLANTUONO, L.; PALMISANO, F.; AMOROSO, N.; MONACO, A.; PERAGINE, V.; BELLOTTI, R. Detecting the socio-economic drivers of confidence in government with eXplainable Artificial Intelligence. *Scientific Reports*, Nature Research, v. 13, n. 1, p. 839, 1 2023. ISSN 2045-2322. Available at: <https://www.nature.com/articles/s41598-023-28020-5>.

BERENS, J.; SCHNEIDER, K.; GÖRTZ, S.; OSTER, S.; BURGHOFF, J. *Early Detection of Students at Risk-Predicting Student Dropouts Using Administrative Student Data from German Universities and Machine Learning Methods*. [S.l.], 2019. v. 11, n. 3. Available at: <https://jedm.educationaldatamining.org/index.php/JEDM/article/view/389>.

BERGER, T. Explainable artificial intelligence and economic panel data: A study on volatility spillover along the supply chains. *Finance Research Letters*, Elsevier Ltd, v. 54, p. 103757, 6 2023. ISSN 15446123. Available at: <https://linkinghub.elsevier.com/retrieve/pii/S1544612323001307>.

BHATT, U.; XIANG, A.; SHARMA, S.; WELLER, A.; TALY, A.; JIA, Y.; GHOSH, J.; PURI, R.; MOURA, J. M. F.; ECKERSLEY, P. Explainable machine learning in deployment. In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. New York, NY, USA: ACM, 2020. p. 648–657. ISBN 9781450369367. Available at: <https://dl.acm.org/doi/10.1145/3351095.3375624>.

BOWLES, S. Towards an educational production function. In: *Education, income, and human capital*. NBER, 1970. p. 11–70. Available at: <https://www.nber.org/system/files/chapters/c3276/c3276.pdf>.

BREIMAN, L. Random Forests. *Machine Learning*, v. 45, n. 1, p. 5–32, 2001. ISSN 1573-0565. Available at: <https://doi.org/10.1023/A:1010933404324>.

BREIMAN, L. Statistical Modeling: The Two Cultures (with comments and a rejoinder by the author). *Statistical Science*, v. 16, n. 3, p. 199–215, 8 2001. ISSN 0883-4237. Available at: <https://projecteuclid.org/journals/statistical-science/volume-16/issue-3/Statistical-Modeling--The-Two-Cultures-with-comments-and-a/10.1214/ss/1009213726.full>.

BUCK, N.; ERMISCH, J.; JENKINS, S. Choosing a longitudinal survey design: the issues. *Occasional Paper*, n. September, p. 96–1, 1995. Available at: <https://api.semanticscholar.org/CorpusID:158010184>.

BUSSMANN, N.; GIUDICI, P.; MARINELLI, D.; PAPENBROCK, J. Explainable Machine Learning in Credit Risk Management. *Computational Economics*, Springer, v. 57, n. 1, p. 203–216, 1 2021. ISSN 0927-7099. Available at: <https://link.springer.com/10.1007/s10614-020-10042-0>.

CALDAS, S. J.; BANKSTON, C. Effect of School Population Socioeconomic Status on Individual Academic Achievement. *The Journal of Educational Research*, Routledge, v. 90, n. 5, p. 269–277, 5 1997. ISSN 0022-0671. Available at: <http://www.tandfonline.com/doi/abs/10.1080/00220671.1997.10544583>.

CANDÈS, E.; FAN, Y.; JANSON, L.; LV, J. Panning for Gold: 'Model-X' Knockoffs for High Dimensional Controlled Variable Selection. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, v. 80, n. 3, p. 551–577, 6 2018. ISSN 1369-7412. Available at: <https://academic.oup.com/jrsssb/article/80/3/551/7048447>.

CAO, L. Introduction to domain driven data mining. *Data Mining for Business Applications*, p. 3–10, 2009. Available at: <https://www-staff.it.uts.edu.au/~lbcao/publication/dmba-dddm.pdf>.

CAO, L.; LIN, L.; CHENGQI, Z. Domain Driven in Depth Pattern Discovery: APractical Methodology. *Proceedings 4th Australasion Data Mining Conference AusDM05*, The University of Technology, Sydney, v. 6, p. 101 – 114, 2005. Available at: <http://hdl.handle.net/10453/1903>.

CARNOY, M.; MAROTTA, L.; LOUZANO, P.; KHAVENSON, T.; GUIMARÃES, F. R. F.; CARNAUBA, F. Intranational Comparative Education: What State Differences in Student Achievement Can Teach Us about Improving Education—the Case of Brazil. *Comparative Education Review*, v. 61, n. 4, p. 726–759, 11 2017. ISSN 0010-4086. Available at: <https://www.journals.uchicago.edu/doi/10.1086/693981>.

CARNOY, M.; ROSA, L.; SIMõES, A. Trends in the academic achievement gap between high and low social class children: The case of Brazil. *International Journal of Educational Development*, v. 94, n. July, p. 102650, 10 2022. ISSN 07380593. Available at: <https://linkinghub.elsevier.com/retrieve/pii/S0738059322001006>.

CHANG, D. F.; CHEN, C. C. Cluster analysis for student performance in PISA2015 among OECD economies. *ICIC Express Letters, Part B: Applications*, v. 9, n. 11, p. 1139–1146, 2018. ISSN 21852766. Available at: <http://www.icicelb.org/ellb/contents/2018/11/elb-09-11-07.pdf>.

CHAPMAN, P.; CLINTON, J.; KERBER, R.; REINARTZ, T. K. T.; SHEARER, C.; WIRTH, R. *Cross industry standard process for data mining (crisp-dm)–step by step data mining guide*. Jelentés, 1999. Available at: <https://api.semanticscholar.org/CorpusID:59777418>.

CHATURAPRUEK, S.; DEE, T. S.; JOHARI, R.; KIZILCEC, R. F.; STEVENS, M. L. How a data-driven course planning tool affects college students' GPA. *L@S*, p. 1–63, 2018.

CHEN, C.; LIN, K.; RUDIN, C.; SHAPOSHNIK, Y.; WANG, S.; WANG, T. Globally-Consistent Rule-Based Summary-Explanations for Machine Learning Models: Application to Credit-Risk Evaluation. *Journal of Machine Learning Research*, v. 24, p. 1–44, 2023. Available at: <https://jmlr.org/papers/volume24/21-0488/21-0488.pdf>.

CHEN, H.; JANIZEK, J. D.; LUNDBERG, S.; LEE, S.-I. True to the Model or True to the Data? 6 2020. Available at: <http://arxiv.org/abs/2006.16234>.

CHEN, H.-C.; PRASETYO, E.; TSENG, S.-S.; PUTRA, K. T.; Prayitno; KUSUMAWARDANI, S. S.; WENG, C.-E. Week-Wise Student Performance Early Prediction in Virtual Learning Environment Using a Deep Explainable Artificial Intelligence. *Applied Sciences*, MDPI, v. 12, n. 4, p. 1885, 2 2022. ISSN 2076-3417. Available at: <https://www.mdpi.com/2076-3417/12/4/1885>.

CHEN, J.; ZHANG, Y.; HU, J. Synergistic effects of instruction and affect factors on high- and low-ability disparities in elementary students' reading literacy. *Reading and Writing*, Springer Netherlands, v. 34, n. 1, p. 199–230, 1 2021. ISSN 0922-4777. Available at: <https://link.springer.com/10.1007/s11145-020-10070-0>.

CHEN, V.; JOHNSON, N.; TOPIN, N.; PLUMB, G.; TALWALKAR, A. *Use-Case-Grounded Simulations for Explanation Evaluation*. Advances in Neural Information Processing Systems, 2022. Available at: <https://openreview.net/pdf?id=48Js-sP8wnv>.

CHITTI, M.; CHITTI, P.; JAYABALAN, M. Need for Interpretable Student Performance Prediction. In: *2020 13th International Conference on Developments in eSystems Engineering (DeSE)*. IEEE, 2020. v. 2020-December, p. 269–272. ISBN 978-1-6654-2238-3. Available at: <https://ieeexplore.ieee.org/document/9450735/>.

CHIU, M.-S. Gender Differences in Predicting STEM Choice by Affective States and Behaviors in Online Mathematical Problem Solving: Positive-Affect-to-Success Hypothesis. *Journal of Educational Data Mining*, v. 12, n. 2, p. 48–77, 2020. Available at: <https://jedm.educationaldatamining.org/index.php/JEDM/article/view/409>.

CICHY, R. M.; KAISER, D. Deep Neural Networks as Scientific Models. *Trends in Cognitive Sciences*, Elsevier Ltd, v. 23, n. 4, p. 305–317, 4 2019. ISSN 13646613. Available at: <https://linkinghub.elsevier.com/retrieve/pii/S1364661319300348>.

CLAASSEN, T.; MOOIJ, J. M.; HESKES, T. Learning sparse causal models is not NP-hard. *Uncertainty in Artificial Intelligence - Proceedings of the 29th Conference, UAI 2013*, p. 172–181, 2013. Available at: <https://dl.acm.org/doi/10.5555/3023638.3023656>.

COLEMAN, J. S. Equality of educational opportunity. *Integrated Education*, Taylor \& Francis, v. 6, n. 5, p. 19–28, 1968. Available at: <https://eric.ed.gov/?id=ED012275>.

COLEMAN, J. S. *Equality And Achievement In Education*. Routledge, 2019. ISBN 9780429690693. Available at: <https://www.taylorfrancis.com/books/9780429690693>.

CORTEZ, P.; SILVA, A. Using data mining to predict secondary school student performance. *15th European Concurrent Engineering Conference 2008, ECEC 2008 - 5th Future Business Technology Conference, FUBUTEC 2008*, n. May, p. 5–12, 2008.

DARLING-HAMMOND, L. How Teacher Education Matters. *Journal of Teacher Education*, v. 51, n. 3, p. 166–173, 5 2000. ISSN 0022-4871. Available at: <http://journals.sagepub.com/doi/10.1177/002248710051003002>.

DEBEER, D.; STROBL, C. Conditional permutation importance revisited. *BMC Bioinformatics*, BioMed Central, v. 21, n. 1, p. 307, 12 2020. ISSN 1471-2105. Available at: <https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-020-03622-2>.

DEPREN, S. K.; AŞKıN, E.; ÖZ, E. Identifying the Classification Performances of Educational Data Mining Methods: A Case Study for TIMSS. *Educational Sciences: Theory & Practice*, v. 17, n. 5, p. 1605–1623, 2017. ISSN 21487561. Available at: <https://jestp.com/index.php/estp/article/view/426>.

DOEWES, A.; PECHENIZKIY, M. Structural Explanation of Automated Essay Scoring. In: *Educational Data Mining*. [s.n.], 2020. p. 1–4. Available at: <https://educationaldatamining.org/files/conferences/EDM2020/papers/paper_259.pdf>.

DONG, X.; HU, J. An Exploration of Impact Factors Influencing Students' Reading Literacy in Singapore with Machine Learning Approaches. *International Journal of English Linguistics*, v. 9, n. 5, p. 52, 8 2019. ISSN 1923-8703. Available at: <http://www.ccsenet.org/journal/index.php/ijel/article/view/0/40486>.

DOSHI-VELEZ, F.; KIM, B. Towards A Rigorous Science of Interpretable Machine Learning. 2 2017. Available at: <http://arxiv.org/abs/1702.08608>.

DUA, D.; GRAFF, C. *{UCI} Machine Learning Repository*. 2017. Available at: <http://archive.ics.uci.edu/ml>.

EHSAN, U.; WINTERSBERGER, P.; LIAO, Q. V.; MARA, M.; STREIT, M.; WACHTER, S.; RIENER, A.; RIEDL, M. O. Operationalizing Human-Centered Perspectives in Explainable AI. In: *Conference on Human Factors in Computing Systems - Proceedings*. [S.l.]: Association for Computing Machinery, 2021. ISBN 9781450380959.

FAGANT, L. M.; SHORTLIFFE, E. H.; BUCHANAN, B. G. Computer-based Medical Decision Making: From MYCIN to VM. *Automedica*, v. 3, p. 97–106, 1980. Available at: <https://stacks.stanford.edu/file/druid:xt779dh5744/xt779dh5744.pdf>.

FAN, C.; XU, J.; NATARAJAN, B. Y.; MOSTAFAVI, A. Interpretable machine learning learns complex interactions of urban features to understand socio-economic inequality. *Computer-Aided Civil and Infrastructure Engineering*, John Wiley and Sons Inc, v. 38, n. 14, p. 2013–2029, 9 2023. ISSN 1093-9687. Available at: <https://onlinelibrary.wiley.com/doi/10.1111/mice.12972>.

FANCSALI, S. E.; YUDELSON, M. V.; BERMAN, S. R.; RITTER, S. Intelligent instructional hand offs. *Proceedings of the 11th International Conference on Educational Data Mining, EDM 2018*, p. 198–207, 2018. Available at: <https://files.eric.ed.gov/fulltext/ED593100.pdf>.

FAWCETT, T. An introduction to ROC analysis. *Pattern Recognition Letters*, v. 27, n. 8, p. 861–874, 6 2006. ISSN 01678655. Available at: <https://linkinghub.elsevier.com/retrieve/pii/S016786550500303X>.

FILHO, R. L. C. S.; ADEODATO, P. J. L.; BRITO, K. dos S. Interpreting Classification Models Using Feature Importance Based on Marginal Local Effects. In: . BRACIS, São Paulo, BR: Springer International Publishing, 2021. v. 2, p. 484–497. Available at: <https://link.springer.com/10.1007/978-3-030-91702-9_32>.

FILHO, R. L. C. S.; BRITO, K.; ADEODATO, P. J. L. A data mining framework for reporting trends in the predictive contribution of factors related to educational achievement. *Expert Systems with Applications*, Elsevier Ltd, v. 221, p. 119729, 7 2023. ISSN 09574174. Available at: <https://linkinghub.elsevier.com/retrieve/pii/S0957417423002300>.

FILHO, R. L. C. S.; BRITO, K.; ADEODATO, P. J. L. Leveraging Causal Reasoning in Educational Data Mining: An Analysis of Brazilian Secondary Education. *Applied Sciences*, MDPI, v. 13, n. 8, p. 5198, 4 2023. ISSN 2076-3417. Available at: <https://www.mdpi.com/2076-3417/13/8/5198>.

FILHO, R. L. C. S.; GARG, A.; BRITO, K.; ADEODATO, P. J. L.; CARNOY, M. Beyond scores: A machine learning approach to comparing educational system effectiveness. *PLOS ONE*, Public Library of Science, v. 18, n. 10, p. e0289260, 10 2023. ISSN 1932-6203. Available at: <https://dx.plos.org/10.1371/journal.pone.0289260>.

FILHO, R. L. S. *EnemCensus2009-2019*. Harvard Dataverse, 2022. Available at: <https://doi.org/10.7910/DVN/WEWDHL>.

FILHO, R. L. S.; ADEODATO, P. J. Data Mining Solution for Assessing the Secondary School Students of Brazilian Federal Institutes. In: *2019 8th Brazilian Conference on Intelligent Systems (BRACIS)*. IEEE, 2019. p. 574–579. ISBN 978-1-7281-4253-1. Available at: <https://ieeexplore.ieee.org/document/8923965/>.

FISCHER, C.; PARDOS, Z. A.; BAKER, R. S.; WILLIAMS, J. J.; SMYTH, P.; YU, R.; SLATER, S.; BAKER, R.; WARSCHAUER, M. Mining Big Data in Education: Affordances and Challenges. *Review of Research in Education*, v. 44, n. 1, p. 130–160, 3 2020. ISSN 0091-732X. Available at: <http://journals.sagepub.com/doi/10.3102/0091732X20903304>.

FISHER, A.; RUDIN, C.; DOMINICI, F. All Models are Wrong, but Many are Useful: Learning a Variable's Importance by Studying an Entire Class of Prediction Models Simultaneously. *Journal of Machine Learning Research*, v. 20, p. 1–81, 1 2018. Available at: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8323609/>.

FLEISHER, W. Understanding, Idealization, and Explainable AI. *Episteme*, Cambridge University Press, v. 19, n. 4, p. 534–560, 12 2022. ISSN 1742-3600. Available at: <https://www.cambridge.org/core/product/identifier/S1742360022000399/type/journal_article>.

FLORES, A. W.; BECHTEL, K.; LOWENKAMP, C. T. False Positives, False Negatives, and False Analyses: A Rejoinder to "Machine Bias: There's }Software Used Across the Country to Predict Future Criminals. And it's Biased Against Blacks. *Federal Probation Journal*, v. 80, n. 2, 2016. Available at: <https://www.crj.org/assets/2017/07/9_Machine_bias_rejoinder.pdf>.

FRANCO, J. J.; MIRANDA, F. L. d. A.; STIEGLER, D.; DANTAS, F. R.; BRANCHER, J. D.; NOGUEIRA, T. D. C. Usando Mineração de Dados para Identificar Fatores mais Importantes do Enem dos Últimos 22 Anos. In: *Anais do XXXI Simpósio Brasileiro de Informática na Educação (SBIE 2020)*. Sociedade Brasileira de Computação, 2020. p. 1112–1121. Available at: <https://sol.sbc.org.br/index.php/sbie/article/view/12867>.

FREIESLEBEN, T.; KÖNIG, G.; MOLNAR, C.; TEJERO-CANTERO, A. Scientific Inference With Interpretable Machine Learning: Analyzing Models to Learn About Real-World Phenomena. *CoRR*, 6 2022.

FRIEDMAN, J. H. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, v. 29, n. 5, p. 1189–1232, 10 2001. ISSN 0090-5364. Available at: <https://projecteuclid.org/journals/annals-of-statistics/volume-29/issue-5/Greedy-function-approximation-A-gradient-boosting-machine/10.1214/aos/1013203451.full>.

FRYE, C.; MIJOLLA, D. de; BEGLEY, T.; COWTON, L.; STANLEY, M.; FEIGE, I. Shapley Explainability on the Data Manifold. In: *International Conference on Learning Representations*. [s.n.], 2021. p. 1–14. Available at: <https://openreview.net/pdf?id=OPyWRrcjVQw>.

GABRIEL, F.; SIGNOLET, J.; WESTWELL, M. A machine learning approach to investigating the effects of mathematics dispositions on mathematical literacy. *International Journal of Research & Method in Education*, v. 41, n. 3, p. 306–327, 5 2018. ISSN 1743-727X. Available at: <https://www.tandfonline.com/doi/full/10.1080/1743727X.2017.1301916>.

GAMAZO, A.; MARTÍNEZ-ABAD, F. An Exploration of Factors Linked to Academic Performance in PISA 2018 Through Data Mining Techniques. *Frontiers in Psychology*, v. 11, n. November, p. 1–17, 11 2020. ISSN 1664-1078. Available at: <https://www.frontiersin.org/articles/10.3389/fpsyg.2020.575167/full>.

GKOLEMIS, V.; DALAMAGAS, T.; DIOU, C.; KHAN, E.; GÖNEN, M. DALE: Differential Accumulated Local Effects for efficient and accurate global explanations. In: *Proceedings of Machine Learning Research*. [s.n.], 2022. v. 189, p. 2022–2022. Available at: <https://proceedings.mlr.press/v189/gkolemis23a/gkolemis23a.pdf>.

GKOLEMIS, V.; DALAMAGAS, T.; NTOUTSI, E.; DIOU, C. RHALE: Robust and Heterogeneity-aware Accumulated Local Effects. 9 2023. Available at: <http://arxiv.org/abs/2309.11193>.

GOLDSTEIN, A.; KAPELNER, A.; BLEICH, J.; PITKIN, E. Peeking Inside the Black Box: Visualizing Statistical Learning With Plots of Individual Conditional Expectation. *Journal of Computational and Graphical Statistics*, Taylor & Francis, v. 24, n. 1, p. 44–65, 1 2015. ISSN

1061-8600. Available at: <http://www.tandfonline.com/doi/full/10.1080/10618600.2014.907095>.

GOMES, C. M. A.; JELIHOVSCHI, E. Presenting the Regression Tree Method and its application in a large-scale educational dataset. *International Journal of Research & Method in Education*, Routledge, v. 43, n. 2, p. 201–221, 3 2020. ISSN 1743-727X. Available at: <https://www.tandfonline.com/doi/full/10.1080/1743727X.2019.1654992>.

GOMES, M.; HIRATA, G.; OLIVEIRA, J. B. A. e. Student composition in the PISA assessments: Evidence from Brazil. *International Journal of Educational Development*, v. 79, n. November, p. 102299, 11 2020. ISSN 07380593. Available at: <https://linkinghub.elsevier.com/retrieve/pii/S0738059320304582>.

GOROSTIAGA, A.; ROJO-ÁLVAREZ, J. L. On the use of conventional and statistical-learning techniques for the analysis of PISA results in Spain. *Neurocomputing*, Elsevier, v. 171, p. 625–637, 1 2016. ISSN 09252312. Available at: <https://linkinghub.elsevier.com/retrieve/pii/S092523121500942X>.

GREENWELL, B. M.; BOEHMKE, B. C.; MCCARTHY, A. J. A Simple and Effective Model-Based Variable Importance Measure. p. 1–27, 5 2018. Available at: <http://arxiv.org/abs/1805.04755>.

GREGORUTTI, B.; MICHEL, B.; SAINT-PIERRE, P. Correlation and variable importance in random forests. *Statistics and Computing*, Springer New York LLC, v. 27, n. 3, p. 659–678, 5 2017. ISSN 0960-3174. Available at: <http://link.springer.com/10.1007/s11222-016-9646-1>.

GRINSZTAJN, L.; OYALLON, E.; VAROQUAUX, G. Why do tree-based models still outperform deep learning on typical tabular data? In: KOYEJO, S.; MOHAMED, S.; AGARWAL, A.; BELGRAVE, D.; CHO, K.; OH, A. (Ed.). *36th Conference on Neural Information Processing Systems*. Curran Associates, Inc., 2022. v. 35, p. 507–520. Available at: <https://proceedings.neurips.cc/paper_files/paper/2022/file/0378c7692da36807bdec87ab043cdadc-Paper-Datasets_and_Benchmarks.pdf>.

GUIDOTTI, R.; MONREALE, A.; RUGGIERI, S.; TURINI, F.; GIANNOTTI, F.; PEDRESCHI, D. A survey of methods for explaining black box models. *ACM Computing Surveys*, v. 51, n. 5, 2018. ISSN 15577341.

HASIB, K. M.; RAHMAN, F.; HASNAT, R.; ALAM, M. G. R. 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 2022*. [S.l.]: Institute of Electrical and Electronics Engineers Inc., 2022. p. 399–405. ISBN 9781665483032.

HASTIE ROBERT TIBSHIRANI, J. F. T. Assessment and Selection. In: *The Business of Giving*. Palgrave Macmillan, 2014. Available at: <http://www.palgraveconnect.com/doifinder/10.1057/9780230355033.0018>.

HERNÁNDEZ-TORRANO, D.; COURTNEY, M. G. R. Modern international large-scale assessment in education: an integrative review and mapping of the literature. *Large-scale Assessments in Education*, v. 9, n. 1, p. 17, 12 2021. ISSN 2196-0739. Available at: <https://largescaleassessmentsineducation.springeropen.com/articles/10.1186/s40536-021-00109-1>.

HONG, J.; KIM, H.; HONG, H.-G. Random Forest Analysis of Factors Predicting Science Achievement Groups: Focusing on Science Activities and Learning in School. *Asia-Pacific Science Education*, Brill Rodopi, v. 8, n. 2, p. 424–451, 12 2022. ISSN 2364-1177. Available at: <https://brill.com/view/journals/apse/8/2/article-p424_6.xml>.

HOOKER, G.; MENTCH, L.; ZHOU, S. Unrestricted Permutation forces Extrapolation: Variable Importance Requires at least One More Model, or There Is No Free Variable Importance. p. 1–15, 5 2019. Available at: <http://arxiv.org/abs/1905.03151>.

HOQ, M.; BRUSILOVSKY, P.; AKRAM, B. Analysis of an Explainable Student Performance Prediction Model in an Introductory Programming Course. 2023. Available at: <https://doi.org/10.5281/zenodo.8115693>.

HU, J.; DONG, X.; PENG, Y. Discovery of the key contextual factors relevant to the reading performance of elementary school students from 61 countries/regions: insight from a machine learning-based approach. *Reading and Writing*, Springer Netherlands, v. 35, n. 1, p. 93–127, 1 2022. ISSN 0922-4777. Available at: <https://link.springer.com/10.1007/s11145-021-10176-z>.

HUANG, G.; REISER, M.; PARKER, A.; MUNIEC, J.; SALVUCCI, S.; RALPH, J. Institute of Education Science Findings from Interviews with Education Policymakers. 2003. Available at: <http://files.eric.ed.gov/fulltext/ED480144.pdf>.

JANZING, D.; MINORICS, L.; BLÖBAUM, P. Feature relevance quantification in explainable AI: A causal problem. In: Chiappa; Silvia; Calandra (Ed.). *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*. [s.n.], 2020. p. 2907–2916. Available at: <http://proceedings.mlr.press/v108/janzing20a/janzing20a.pdf>.

JAUHIAINEN, S.; KAUPPI, J.-P.; LEPPÄNEN, M.; PASANEN, K.; PARKKARI, J.; VASANKARI, T.; KANNUS, P.; ÄYRÄMÖ, S. New Machine Learning Approach for Detection of Injury Risk Factors in Young Team Sport Athletes. *International Journal of Sports Medicine*, v. 42, n. 02, p. 175–182, 2 2021. ISSN 0172-4622. Available at: <http://www.thieme-connect.de/DOI/DOI?10.1055/a-1231-5304>.

JAYAPRAKASH, S. M.; MOODY, E. W.; LAURÍA, E. J.; REGAN, J. R.; BARON, J. D. Early Alert of Academically At-Risk Students: An Open Source Analytics Initiative. *Journal of Learning Analytics*, v. 1, n. 1, p. 6–47, 5 2014. ISSN 1929-7750. Available at: <https://learning-analytics.info/index.php/JLA/article/view/3249>.

JEBEILE, J.; KENNEDY, A. G. Explaining with Models: The Role of Idealizations. *International Studies in the Philosophy of Science*, Routledge, v. 29, n. 4, p. 383–392, 10 2015. ISSN 0269-8595. Available at: <https://www.tandfonline.com/doi/full/10.1080/02698595.2015.1195143>.

JIANG, W.; PARDOS, Z. A.; WEI, Q. Goal-based Course Recommendation. In: *Proceedings of the 9th International Conference on Learning Analytics & Knowledge*. New York, NY, USA: ACM, 2019. p. 36–45. ISBN 9781450362566. Available at: <https://dl.acm.org/doi/10.1145/3303772.3303814>.

JOHANSSON, S. International large-scale assessments: what uses, what consequences? *Educational Research*, Routledge, v. 58, n. 2, p. 139–148, 4 2016. ISSN 0013-1881. Available at: <http://www.tandfonline.com/doi/full/10.1080/00131881.2016.1165559>.

KAPLAN, D.; HUANG, M. Bayesian probabilistic forecasting with large-scale educational trend data: a case study using NAEP. *Large-scale Assessments in Education*, Springer US, v. 9, n. 1, p. 15, 7 2021. ISSN 2196-0739. Available at: <https://largescaleassessmentsineducation.springeropen.com/articles/10.1186/s40536-021-00108-2>.

KHAN, A.; GHOSH, S. K. Student performance analysis and prediction in classroom learning: A review of educational data mining studies. *Education and Information Technologies*, Springer, v. 26, n. 1, p. 205–240, 1 2021. ISSN 1360-2357. Available at: <https://link.springer.com/10.1007/s10639-020-10230-3>.

KITCHIN, R. Big Data, new epistemologies and paradigm shifts. *Big Data & Society*, SAGE Publications Ltd, v. 1, n. 1, p. 205395171452848, 4 2014. ISSN 2053-9517. Available at: <http://journals.sagepub.com/doi/10.1177/2053951714528481>.

KOLMOGOROV, A. c. *Inst. Ital. Attuari, Giorn.*, v. 4, p. 83–91, 1933.

KOVALEV, S.; KOLODENKOVA, A.; MUNTYAN, E. Educational Data Mining: Current Problems and Solutions. In: *2020 V International Conference on Information Technologies in Engineering Education (Inforino)*. IEEE, 2020. p. 1–5. ISBN 978-1-7281-4810-6. Available at: <https://ieeexplore.ieee.org/document/9111699/>.

LAKKARAJU, H.; BACH, S. H.; LESKOVEC, J. Interpretable Decision Sets. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York, NY, USA: ACM, 2016. v. 13-17-August-2016, p. 1675–1684. ISBN 9781450342322. Available at: <https://dl.acm.org/doi/10.1145/2939672.2939874>.

LEE, Y.-G.; OH, J.-Y.; KIM, D.; KIM, G. SHAP Value-Based Feature Importance Analysis for Short-Term Load Forecasting. *Journal of Electrical Engineering & Technology*, Springer Nature Singapore, v. 18, n. 1, p. 579–588, 1 2023. ISSN 1975-0102. Available at: <https://link.springer.com/10.1007/s42835-022-01161-9>.

LEEPER, T. J. Interpreting Regression Results using Average Marginal Effects with R's margins. *https://cran.r-project.org/web/packages/margins/vignettes/TechnicalDetails.pdf*, p. 32, 2021. Available at: <https://rdrr.io/cran/margins/f/inst/doc/TechnicalDetails.pdf%0Ahttps://cran.r-project.org/web/packages/margins/vignettes/TechnicalDetails.pdf>.

LEI, J.; G'SELL, M.; RINALDO, A.; TIBSHIRANI, R. J.; WASSERMAN, L. Distribution-Free Predictive Inference for Regression. *Journal of the American Statistical Association*, Taylor & Francis, v. 113, n. 523, p. 1094–1111, 7 2018. ISSN 0162-1459. Available at: <https://www.tandfonline.com/doi/full/10.1080/01621459.2017.1307116>.

LEZHNINA, O.; KISMIHÓK, G. Combining statistical and machine learning methods to explore German students' attitudes towards ICT in PISA. *International Journal of Research & Method in Education*, v. 45, n. 2, p. 180–199, 3 2022. ISSN 1743-727X. Available at: <https://www.tandfonline.com/doi/full/10.1080/1743727X.2021.1963226>.

LI, X.; WANG, Y.; BASU, S.; KUMBIER, K.; YU, B. A Debiased MDI Feature Importance Measure for Random Forests. In: H. Wallach; H. Larochelle; A. BeygelzimerF.; R. Garnett (Ed.). *NeuroIPS*. [s.n.], 2019. Available at: <https://proceedings.neurips.cc/paper_files/paper/2019/file/702cafa3bb4c9c86e4a3b6834b45aedd-Paper.pdf>.

LI, Z.; WEI, Z.; ZHANG, Y.; KONG, X.; MA, C. Applying an interpretable machine learning framework to study mobility inequity in the recovery phase of COVID-19 pandemic. *Travel Behaviour and Society*, Elsevier Ltd, v. 33, p. 100621, 10 2023. ISSN 2214367X. Available at: <https://linkinghub.elsevier.com/retrieve/pii/S2214367X23000728>.

LIEBERSON, S.; HORWICH, J. Implication Analysis: A Pragmatic Proposal for Linking Theory and Data in the Social Sciences. *Sociological Methodology*, v. 38, n. 1, p. 1–50, 8 2008. ISSN 0081-1750. Available at: <http://journals.sagepub.com/doi/10.1111/j.1467-9531.2008.00199.x>.

LINARDATOS, P.; PAPASTEFANOPOULOS, V.; KOTSIANTIS, S. Explainable AI: A Review of Machine Learning Interpretability Methods. *Entropy*, MDPI AG, v. 23, n. 1, p. 18, 12 2020. ISSN 1099-4300. Available at: <https://www.mdpi.com/1099-4300/23/1/18>.

LIPTON, Z. C. The Mythos of Model Interpretability. *Queue*, Association for Computing Machinery, New York, NY, USA, v. 16, n. 3, p. 31–57, 6 2018. ISSN 1542-7730. Available at: <https://dl.acm.org/doi/10.1145/3236386.3241340>.

LIU, C. *The Three-Body Problem*. Paperback. New York, NY: Tor Books, 2016. (The Three-Body Problem, v. 1). ISBN 978-0765382030.

LIU, X.; RUIZ, M. E. Using data mining to predict K–12 students' performance on large-scale assessment items related to energy. *Journal of Research in Science Teaching*, v. 45, n. 5, p. 554–573, 5 2008. ISSN 0022-4308. Available at: <https://onlinelibrary.wiley.com/doi/10.1002/tea.20232>.

LIVIERIS, I. E.; KARACAPILIDIS, N.; DOMALIS, G.; TSAKALIDIS, D. An Advanced Explainable and Interpretable ML-Based Framework for Educational Data Mining. In: *Lecture Notes in Networks and Systems*. Springer Science and Business Media Deutschland GmbH, 2023. v. 769 LNNS, p. 87–96. Available at: <https://link.springer.com/10.1007/978-3-031-42134-1_9>.

LONG, J. S.; LONG, J. S. *Regression models for categorical and limited dependent variables*. [S.l.]: Sage, 1997. 1–328 p.

LONG, J. S.; MUSTILLO, S. A. Using Predictions and Marginal Effects to Compare Groups in Regression Models for Binary Outcomes. *Sociological Methods & Research*, v. 50, n. 3, p. 1284–1320, 8 2021. ISSN 0049-1241. Available at: <http://journals.sagepub.com/doi/10.1177/0049124118799374>.

LOYOLA-GONZALEZ, O. Black-Box vs. White-Box: Understanding Their Advantages and Weaknesses From a Practical Point of View. *IEEE Access*, Institute of Electrical and Electronics Engineers Inc., v. 7, p. 154096–154113, 2019. ISSN 2169-3536. Available at: <https://ieeexplore.ieee.org/document/8882211/>.

LUNDBERG, S. M.; ERION, G.; CHEN, H.; DEGRAVE, A.; PRUTKIN, J. M.; NAIR, B.; KATZ, R.; HIMMELFARB, J.; BANSAL, N.; LEE, S.-I. From Local Explanations to Global Understanding with Explainable AI for Trees. *Nature machine intelligence*, Nature Research, v. 2, n. 1, p. 56–67, 1 2020. ISSN 2522-5839. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/32607472http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC7326367>.

LUNDBERG, S. M.; LEE, S.-I. A Unified Approach to Interpreting Model Predictions. In: *Proceedings of the 31st International Conference on Neural Information Processing Systems*. long Beach: Curran Associates Inc., 2017. (NIPS'17), p. 4768–4777. ISBN 9781510860964.

MAIA, J. d. S. Z.; BUENO, A. P. A.; SATO, J. R. Assessing the educational performance of different Brazilian school cycles using data science methods. *PLOS ONE*, v. 16, n. 3, p. e0248525, 3 2021. ISSN 1932-6203. Available at: <https://dx.plos.org/10.1371/journal.pone.0248525>.

MALLINSON, V.; NOAH, H. J.; ECKSTEIN, M. A. Towards a Science of Comparative Education. *British Journal of Educational Studies*, v. 17, n. 3, p. 334, 10 1969. ISSN 00071005. Available at: <https://www.jstor.org/stable/10.2307/3119655?origin=crossref>.

MANGALATHU, S.; KARTHIKEYAN, K.; FENG, D.-C.; JEON, J.-S. Machine-learning interpretability techniques for seismic performance assessment of infrastructure systems. *Engineering Structures*, Elsevier Ltd, v. 250, p. 112883, 1 2022. ISSN 01410296. Available at: <https://linkinghub.elsevier.com/retrieve/pii/S0141029621010336>.

MARTÍNEZ-ABAD, F. Identification of Factors Associated With School Effectiveness With Data Mining Techniques: Testing a New Approach. *Frontiers in Psychology*, v. 10, n. November, p. 1–13, 11 2019. ISSN 1664-1078. Available at: <https://www.frontiersin.org/article/10.3389/fpsyg.2019.02583/full>.

MARTÍNEZ-ABAD, F.; GAMAZO, A.; RODRÍGUEZ-CONDE, M. J. Big Data in Education. In: *Proceedings of the Sixth International Conference on Technological Ecosystems for Enhancing Multiculturality*. New York, NY, USA: ACM, 2018. p. 145–150. ISBN 9781450365185. Available at: <https://dl.acm.org/doi/10.1145/3284179.3284206>.

MARTÍNEZ-ABAD, F.; GAMAZO, A.; RODRÍGUEZ-CONDE, M.-J. Educational Data Mining: Identification of factors associated with school effectiveness in PISA assessment. *Studies in Educational Evaluation*, v. 66, n. March, p. 100875, 9 2020. ISSN 0191491X. Available at: <https://linkinghub.elsevier.com/retrieve/pii/S0191491X20301231>.

MASCI, C.; JOHNES, G.; AGASISTI, T. Student and school performance across countries: A machine learning approach. *European Journal of Operational Research*, Elsevier B.V., v. 269, n. 3, p. 1072–1085, 9 2018. ISSN 03772217. Available at: <https://linkinghub.elsevier.com/retrieve/pii/S0377221718301462>.

MATETIC, M. Mining Learning Management System Data Using Interpretable Neural Networks. In: *2019 42nd International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*. IEEE, 2019. p. 1282–1287. ISBN 978-953-233-098-4. Available at: <https://ieeexplore.ieee.org/document/8757113/>.

MAULANA, A.; NOVIANDY, T. R.; SASMITA, N. R.; PARISTIOWATI, M.; SUHENDRA, R.; YANDRI, E.; SATRIO, J.; IDROES, R. Optimizing University Admissions: A Machine Learning Perspective. *Journal of Educational Management and Learning*, PT. Heca Sentra Analitika, v. 1, n. 1, p. 1–7, 6 2023. ISSN 3025-1117. Available at: <https://heca-analitika.com/jeml/article/view/46>.

MILLER, G. A. The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological review*, American Psychological Association, v. 63, n. 2, p. 81, 1956. Available at: <https://labs.la.utexas.edu/gilden/files/2016/04/MagicNumberSeven-Miller1956.pdf>.

MILLER, T. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, Elsevier B.V., v. 267, p. 1–38, 2019. ISSN 00043702. Available at: <https://doi.org/10.1016/j.artint.2018.07.007>.

MITTELSTADT, B.; RUSSELL, C.; WACHTER, S. Explaining Explanations in AI. In: *Proceedings of the Conference on Fairness, Accountability, and Transparency*. New York, NY, USA: ACM, 2019. (FAT* '19), p. 279–288. ISBN 9781450361255. Available at: <https://dl.acm.org/doi/10.1145/3287560.3287574>.

MIZE, T. D.; DOAN, L.; LONG, J. S. A General Framework for Comparing Predictions and Marginal Effects across Models. *Sociological Methodology*, v. 49, n. 1, p. 152–189, 8 2019. ISSN 0081-1750. Available at: <http://journals.sagepub.com/doi/10.1177/0081175019852763>.

MOLINA, M.; GARIP, F. Annual Review of Sociology Machine Learning for Sociology. 2019. Available at: <https://doi.org/10.1146/annurev-soc-073117->.

MOLNAR, C. *Model-agnostic interpretable machine learning*. 1–260 p. Phd Thesis (PhD Thesis) — lmu, 2022. Available at: <https://d-nb.info/1265378495/34>.

MOLNAR, C. *Interpretable Machine Learning*. 2. ed. [s.n.], 2023. Available at: <christophm.github.io/interpretable-ml-book/>.

MOLNAR, C.; CASALICCHIO, G.; BISCHL, B. Quantifying Model Complexity via Functional Decomposition for Better Post-Hoc Interpretability. 4 2019. Available at: <http://arxiv.org/abs/1904.03867http://dx.doi.org/10.1007/978-3-030-43823-4_17>.

MOLNAR, C.; KÖNIG, G.; BISCHL, B.; CASALICCHIO, G. Model-agnostic feature importance and effects with dependent features: a conditional subgroup approach. *Data Mining and Knowledge Discovery*, Springer, 1 2023. ISSN 1384-5810. Available at: <https://link.springer.com/10.1007/s10618-022-00901-9>.

MOLNAR, C.; KÖNIG, G.; HERBINGER, J.; FREIESLEBEN, T.; DANDL, S.; SCHOLBECK, C. A.; CASALICCHIO, G.; GROSSE-WENTRUP, M.; BISCHL, B. General Pitfalls of Model-Agnostic Interpretation Methods for Machine Learning Models. In: GOEBEL, R. H. A.; RUTH, F.; TAESUP, M.; KLAUS-ROBERT, M.; WOJCIECH, S. (Ed.). *xxAI - Beyond Explainable AI: International Workshop, Held in Conjunction with ICML 2020, July 18, 2020, Vienna, Austria, Revised and Extended Papers*. Cham: Springer International Publishing, 2022. p. 39–68. Available at: <https://link.springer.com/10.1007/978-3-031-04083-2_4>.

MOOD, C. Logistic regression : Uncovering unobserved heterogeneity. p. 1–25, 2017.

MOUSSAVI, R.; GOBERT, J.; PEDRO, M. S. The effect of scaffolding on the immediate transfer of students' data interpretation skills within science topics. In: *Proceedings of International Conference of the Learning Sciences, ICLS*. [s.n.], 2016. v. 2, p. 1002–1005. ISBN 9780990355083. ISSN 18149316. Available at: <https://repository.isls.org/bitstream/1/364/1/157.pdf>.

MU, T.; ANDREAJETTEN, A. J. W. C. H.; BRUNSKILL, w. E. Towards Suggesting Actionable Interventions for Wheel-Spinning Students. In: *International Conference on Educational Data Mining*. [s.n.], 2020. p. 183–193. Available at: <https://educationaldatamining.org/files/conferences/EDM2020/papers/paper_201.pdf>.

MULLAINATHAN, S.; SPIESS, J. Machine Learning: An Applied Econometric Approach. *Journal of Economic Perspectives*, American Economic Association, v. 31, n. 2, p. 87–106, 5 2017. ISSN 0895-3309. Available at: <https://pubs.aeaweb.org/doi/10.1257/jep.31.2.87>.

NAIMI, A. I.; WESTREICH, D. J. Big Data: A Revolution That Will Transform How We Live, Work, and Think. *American Journal of Epidemiology*, Houghton Mifflin Harcourt, v. 179, n. 9, p. 1143–1144, 5 2014. ISSN 0002-9262. Available at: <https://academic.oup.com/aje/article-lookup/doi/10.1093/aje/kwu085>.

NAMOUN, A.; ALSHANQITI, A. Predicting Student Performance Using Data Mining and Learning Analytics Techniques: A Systematic Literature Review. *Applied Sciences*, MDPI AG, v. 11, n. 1, p. 237, 12 2020. ISSN 2076-3417. Available at: <https://www.mdpi.com/2076-3417/11/1/237>.

NEMBRINI, S. Bias in the intervention in prediction measure in random forests: illustrations and recommendations. *Bioinformatics*, Oxford University Press, v. 35, n. 13, p. 2343–2345, 7 2019. ISSN 1367-4803. Available at: <https://academic.oup.com/bioinformatics/article/35/13/2343/5194684>.

NICODEMUS, K. K. Letter to the Editor: On the stability and ranking of predictors from random forest variable importance measures. *Briefings in Bioinformatics*, v. 12, n. 4, p. 369–373, 7 2011. ISSN 1467-5463. Available at: <https://academic.oup.com/bib/article-lookup/doi/10.1093/bib/bbr016>.

OCDE. *PISA 2012 Results: What makes schools sucessful? Resources, policies and practices*. [s.n.], 2013. ISBN 9789264201156. Available at: <http://dx.doi.org/10.1787/9789264201156-en>.

OLIVEIRA, H.; MELLO, R. F.; ROSA, B. A. B.; RAKOVIC, M.; MIRANDA, P.; CORDEIRO, T.; ISOTANI, S.; BITTENCOURT, I.; GASEVIC, D. Towards Explainable Prediction of Essay Cohesion in Portuguese and English. In: *LAK23: 13th International Learning Analytics and Knowledge Conference*. New York, NY, USA: Association for Computing Machinery, 2023. (LAK2023), p. 509–519. ISBN 9781450398657. Available at: <https://doi.org/10.1145/3576050.3576152>.

PEARL, J. [Bayesian Analysis in Expert Systems]: Comment: Graphical Models, Causality and Intervention. *Statistical Science*, v. 8, n. 3, p. 266–269, 8 1993. ISSN 0883-4237. Available at: <https://projecteuclid.org/journals/statistical-science/volume-8/issue-3/Bayesian-Analysis-in-Expert-Systems--Comment--Graphical-Models/10.1214/ss/1177010894.full>.

PEDREGOSA, F.; VAROQUAUX, G.; GRAMFORT, A.; MICHEL, V.; THIRION, B.; GRISEL, O.; BLONDEL, M.; PRETTENHOFER, P.; WEISS, R.; DUBOURG, V.; VANDERPLAS, J.; PASSOS, A.; COURNAPEAU, D.; BRUCHER, M.; PERROT, M.; DUCHESNAY, E. Scikit-learn: Machine Learning in {P}ython. *Journal of Machine Learning Research*, v. 12, p. 2825–2830, 2011. Available at: <http://jmlr.org/papers/v12/pedregosa11a.html>.

PELLAGATTI, M.; MASCI, C.; IEVA, F.; PAGANONI, A. M. Generalized mixed-effects random forest: A flexible approach to predict university student dropout. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, v. 14, n. 3, p. 241–257, 6 2021. ISSN 1932-1864. Available at: <https://onlinelibrary.wiley.com/doi/10.1002/sam.11505>.

PROVOST, F.; FAWCETT, T. Data Science and its Relationship to Big Data and Data-Driven Decision Making. *Big Data*, Mary Ann Liebert Inc., v. 1, n. 1, p. 51–59, 3 2013. ISSN 2167-6461. Available at: <http://www.liebertpub.com/doi/10.1089/big.2013.1508>.

RANGONE, G. N.; PIZARRO, C.; MONTEJANO, G. Automation of an Educational Data Mining Model Applying Interpretable Machine Learning and Auto Machine Learning. In: *Smart Innovation, Systems and Technologies*. Springer Science and Business Media Deutschland GmbH, 2022. v. 259 SIST, p. 22–30. Available at: <https://link.springer.com/10.1007/978-981-16-5792-4_3>.

RASTROLLO-GUERRERO, J. L.; GÓMEZ-PULIDO, J. A.; DURÁN-DOMÍNGUEZ, A. Analyzing and Predicting Students' Performance by Means of Machine Learning: A Review. *Applied Sciences*, v. 10, n. 3, p. 1042, 2 2020. ISSN 2076-3417. Available at: <https://www.mdpi.com/2076-3417/10/3/1042>.

RAZAVIAN, N.; BLECKER, S.; SCHMIDT, A. M.; SMITH-MCLALLEN, A.; NIGAM, S.; SONTAG, D. Population-Level Prediction of Type 2 Diabetes From Claims Data and Analysis of Risk Factors. *Big Data*, v. 3, n. 4, p. 277–287, 12 2015. ISSN 2167-6461. Available at: <http://www.liebertpub.com/doi/10.1089/big.2015.0020>.

REALINHO, V.; MACHADO, J.; BAPTISTA, L.; MARTINS, M. V. Predicting Student Dropout and Academic Success. *Data*, v. 7, n. 11, p. 146, 10 2022. ISSN 2306-5729. Available at: <https://www.mdpi.com/2306-5729/7/11/146>.

REBAI, S.; YAHIA, F. B.; ESSID, H. A graphically based machine learning approach to predict secondary schools performance in Tunisia. *Socio-Economic Planning Sciences*, Elsevier, v. 70, n. June, p. 100724, 6 2020. ISSN 00380121. Available at: <https://linkinghub.elsevier.com/retrieve/pii/S0038012118302908>.

RODRIGUES, D.; REGIO, M.; MUSSE, S.; MANSSOUR, I. Data Mining on the Prediction of Student's Performance at the High School National Examination. In: *Proceedings of the 13th International Conference on Computer Supported Education*. SCITEPRESS - Science and Technology Publications, 2021. p. 92–99. ISBN 978-989-758-502-9. Available at: <https://www.scitepress.org/DigitalLibrary/Link.aspx?doi=10.5220/0010408000920099>.

ROHANI, N.; GAL, K.; GALLAGHER, M.; MANATAKI, e. A. Early Prediction of Student Performance in a Health Data Science MOOC. In: *Inter- national Conference on Educational Data Mining*. [s.n.], 2023. p. 325–333. Available at: <https://educationaldatamining.org/EDM2023/proceedings/2023.EDM-short-papers.32/2023.EDM-short-papers.32.pdf>.

ROMERO, C.; VENTURA, S. Educational data mining and learning analytics: An updated survey. *WIREs Data Mining and Knowledge Discovery*, v. 10, n. 3, p. 1–21, 5 2020. ISSN 1942-4787. Available at: <https://wires.onlinelibrary.wiley.com/doi/10.1002/widm.1355>.

RUDIN, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, v. 1, n. 5, p. 206–215, 5 2019. ISSN 2522-5839. Available at: <https://www.nature.com/articles/s42256-019-0048-x>.

SAARELA, M.; YENER, B.; ZAKI, M. J.; KÄRKKÄINEN, T. Predicting Math Performance from Raw Large-Scale Educational Assessments Data: A Machine Learning Approach. In: *International Conference on Machine Learning*. New York: JMLR, 2016. v. 48, n. 1, p. 1–8. Available at: <https://www.cs.rpi.edu/~zaki/PaperDir/MLDEAS16.pdf>.

SANTOS, C. A. B. d.; CURI, E. A formação dos professores que ensinam física no ensino médio TT - Training of teachers in High School physics. *Ciência & Educação (Bauru)*, v. 18, n. 4, p. 837–849, 2012. ISSN 1516-7313. Available at: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S1516-73132012000400007&lang=pt%0Ahttp://www.scielo.br/pdf/ciedu/v18n4/v18n4a07.pdf>.

SCAVUZZO, C. M.; SCAVUZZO, J. M.; CAMPERO, M. N.; ANEGAGRIE, M.; ARAMENDIA, A. A.; BENITO, A.; PERIAGO, V. Feature importance: Opening a soil-transmitted helminth machine learning model via SHAP. *Infectious Disease Modelling*, The Authors, v. 7, n. 1, p. 262–276, 3 2022. ISSN 24680427. Available at: <https://linkinghub.elsevier.com/retrieve/pii/S2468042722000045>.

SCHEERENS, J. Process indicators of school functioning: A selection based on the research literature on school effectiveness. *Studies in Educational Evaluation*, v. 17, n. 2-3, p. 371–403, 1 1991. ISSN 0191491X. Available at: <https://linkinghub.elsevier.com/retrieve/pii/S0191491X05800914>.

SCHILTZ, F.; MASCI, C.; AGASISTI, T.; HORN, D. Using regression tree ensembles to model interaction effects: a graphical approach. *Applied Economics*, Routledge, v. 50, n. 58, p. 6341–6354, 12 2018. ISSN 0003-6846. Available at: <https://www.tandfonline.com/doi/full/10.1080/00036846.2018.1489520>.

SCHOLBECK, C. A.; MOLNAR, C.; HEUMANN, C.; BISCHL, B.; CASALICCHIO, G. Sampling, Intervention, Prediction, Aggregation: A Generalized Framework for Model-Agnostic Interpretations. In: *Communications in Computer and Information Science*. [s.n.], 2020. v. 1167 CCIS, p. 205–216. Available at: <https://link.springer.com/10.1007/978-3-030-43823-4_18>.

SHAPLEY, L. S.; others. A value for n-person games. Princeton University Press Princeton, 3 1953. Available at: <https://www.rand.org/content/dam/rand/pubs/papers/2021/P295.pdf>.

SHMUELI, G. To Explain or to Predict? *Statistical Science*, v. 25, n. 3, p. 289–310, 8 2010. ISSN 0883-4237. Available at: <https://projecteuclid.org/journals/statistical-science/volume-25/issue-3/To-Explain-or-to-Predict/10.1214/10-STS330.full>.

SLACK, D.; HILGARD, S.; JIA, E.; SINGH, S.; LAKKARAJU, H. Fooling LIME and SHAP. In: *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. New York, NY, USA: ACM, 2020. p. 180–186. ISBN 9781450371100. Available at: <https://dl.acm.org/doi/10.1145/3375627.3375830>.

SOUZA, D. G. d.; OLIVEIRA, A. M. d.; NASCIMENTO, T. O. S. O Exame Nacional do Ensino Médio: o que Revelam os Dados por Área de Conhecimento num Período Decenal? *COLLOQUIUM HUMANARUM*, Associacao Prudentina de Educacao e Cultura (APEC), v. 17, n. 1, p. 61–74, 5 2020. ISSN 1809-8207.

STIGLIC, G.; KOCBEK, P.; FIJACKO, N.; ZITNIK, M.; VERBERT, K.; CILAR, L. Interpretability of machine learning-based prediction models in healthcare. *WIREs Data Mining and Knowledge Discovery*, v. 10, n. 5, p. 1–13, 9 2020. ISSN 1942-4787. Available at: <https://wires.onlinelibrary.wiley.com/doi/10.1002/widm.1379>.

STROBL, C.; BOULESTEIX, A.-L.; KNEIB, T.; AUGUSTIN, T.; ZEILEIS, A. Conditional variable importance for random forests. *BMC Bioinformatics*, v. 9, n. 1, p. 307, 12 2008. ISSN 1471-2105. Available at: <https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-9-307>.

SULLIVAN, E. Understanding from machine learning models. *The British Journal for the Philosophy of Science*, The University of Chicago Press, v. 73, n. 1, 2022.

SUNDARARAJAN, M.; NAJMI, A. The Many Shapley Values for Model Explanation. In: *Proceedings of the 37th International Conference on Machine Learning*. [s.n.], 2020. p. 9269–9278. Available at: <http://proceedings.mlr.press/v119/sundararajan20b/sundararajan20b.pdf>.

USTUN, B.; RUDIN, C. Supersparse linear integer models for optimized medical scoring systems. *Machine Learning*, Springer New York LLC, v. 102, n. 3, p. 349–391, 3 2016. ISSN 0885-6125. Available at: <http://link.springer.com/10.1007/s10994-015-5528-6>.

VANSCHOREN, J.; RIJN, J. N. van; BISCHL, B.; TORGO, L. OpenML. *ACM SIGKDD Explorations Newsletter*, ACM, New York, NY, USA, v. 15, n. 2, p. 49–60, 6 2014. ISSN 1931-0145. Available at: <https://dl.acm.org/doi/10.1145/2641190.2641198>.

VAPNIK, V. An overview of statistical learning theory. *IEEE Transactions on Neural Networks*, v. 10, n. 5, p. 988–999, 1999. ISSN 10459227. Available at: <http://ieeexplore.ieee.org/document/788640/>.

Varkey Foundation. *Global Teacher Status Index | 2018 | Brasil - GTSI STATISTICS*. [S.l.], 2018. Available at: <https://www.varkeyfoundation.org/media/4833/gtsi-brazil-chart-findings.pdf>.

WANG, C.; HAN, B.; PATEL, B.; RUDIN, C. In Pursuit of Interpretable, Fair and Accurate Machine Learning for Criminal Recidivism Prediction. *Journal of Quantitative Criminology*, Springer, v. 39, n. 2, p. 519–581, 6 2023. ISSN 0748-4518. Available at: <https://link.springer.com/10.1007/s10940-022-09545-w>.

WANG, D.; YANG, Q.; ABDUL, A.; LIM, B. Y. Designing Theory-Driven User-Centric Explainable AI. In: *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. New York, NY, USA: ACM, 2019. p. 1–15. ISBN 9781450359702. Available at: <https://dl.acm.org/doi/10.1145/3290605.3300831>.

WATSON, D. S. Conceptual challenges for interpretable machine learning. *Synthese*, Springer Science and Business Media B.V., v. 200, n. 2, p. 65, 4 2022. ISSN 0039-7857. Available at: <https://link.springer.com/10.1007/s11229-022-03485-5>.

WATSON, D. S.; WRIGHT, M. N. Testing conditional independence in supervised learning algorithms. *Machine Learning*, Springer, v. 110, n. 8, p. 2107–2129, 8 2021. ISSN 0885-6125. Available at: <https://link.springer.com/10.1007/s10994-021-06030-6>.

WEI, P.; LU, Z.; SONG, J. Variable importance analysis: A comprehensive review. *Reliability Engineering & System Safety*, Elsevier Ltd, v. 142, p. 399–432, 10 2015. ISSN 09518320. Available at: <https://linkinghub.elsevier.com/retrieve/pii/S0951832015001672>.

YANG, H.; RUDIN, C.; SELTZER, M. Scalable Bayesian Rule Lists. In: *Proceedings of the 34th International Conference on Machine Learning*. [s.n.], 2017. v. 70, p. 3291–3930. Available at: <http://proceedings.mlr.press/v70/yang17h/yang17h.pdf>.

YANG, J.; WANG, H. Interpretability Analysis of Academic Achievement Prediction Based on Machine Learning. In: *Proceedings - 11th International Conference on Information Technology in Medicine and Education, ITME 2021*. [S.l.]: Institute of Electrical and Electronics Engineers Inc., 2021. p. 475–479. ISBN 9781665406796.

YıLMAZ, N.; SEKEROGLU, B. Student Performance Classification Using Artificial Intelligence Techniques. In: ALIEV, R. A.; KACPRZYK, J.; PEDRYCZ, W.; JAMSHIDI, M.; BABANLI, M. B.; SADIKOGLU, F. M. (Ed.). *10th International Conference on Theory and Application of Soft Computing, Computing with Words and Perceptions - ICSCCW-2019*. Cham: Springer International Publishing, 2020. p. 596–603. ISBN 978-3-030-35249-3. Available at: <https://link.springer.com/chapter/10.1007/978-3-030-35249-3_76>.

ZABRISKIE, C.; YANG, J.; DEVORE, S.; STEWART, J. Using machine learning to predict physics course outcomes. *Physical Review Physics Education Research*, American Physical Society, v. 15, n. 2, p. 020120, 8 2019. ISSN 2469-9896. Available at: <https://link.aps.org/doi/10.1103/PhysRevPhysEducRes.15.020120>.

ZEDNIK, C. Solving the Black Box Problem: A Normative Framework for Explainable Artificial Intelligence. *Philosophy & Technology*, Springer Science and Business Media B.V., v. 34, n. 2, p. 265–288, 6 2021. ISSN 2210-5433. Available at: <https://link.springer.com/10.1007/s13347-019-00382-7>.

ZHAO, Q.; HASTIE, T. Causal Interpretations of Black-Box Models. *Journal of Business & Economic Statistics*, v. 39, n. 1, p. 272–281, 1 2021. ISSN 0735-0015. Available at: <https://www.tandfonline.com/doi/full/10.1080/07350015.2019.1624293>.

ZHAO, Y.; XU, Q.; CHEN, M.; WEISS, G. M. Predicting Student Performance in a Master of Data Science Program using Admissions Data. In: *Proceedings of The 13th International Conference on Educational Data Mining*. [s.n.], 2020. p. 325–334. Available at: <https://files.eric.ed.gov/fulltext/ED607995.pdf>.