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

Milton Vasconcelos da Gama Neto

Análise comparativa das técnicas de Explainable AI e um novo método para geração de explicações textuais

Recife

2022

Milton Vasconcelos da Gama Neto

Análise comparativa das técnicas de Explainable AI e um novo método para geração de explicações textuais

Trabalho apresentado 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 Mestre Profissional em Ciência da Computação.

Área de Concentração: Inteligência Computacional

Orientador: Germano Crispim Vasconcelos

Coorientador: Cleber Zanchettin

Recife

2022

Catalogação na fonte
Bibliotecária Monick Raquel Silvestre da S. Portes, CRB4-1217

G184a Gama Neto, Milton Vasconcelos da
Análise comparativa das técnicas de Explainable AI e um novo método para geração de explicações textuais / Milton Vasconcelos da Gama Neto. – 2022.
100 f.: il., fig., tab.

Orientador: Germano Crispim Vasconcelos.
Dissertação (Mestrado) – Universidade Federal de Pernambuco. CIn, Ciência da Computação, Recife, 2022.
Inclui referências e apêndices.

1. Inteligência computacional. 2. Aprendizagem de máquina. 3. Mineração de dados. I. Vasconcelos, Germano Crispim (orientador). II. Título.

006.31 CDD (23. ed.) UFPE - CCEN 2022 – 90

Milton Vasconcelos da Gama Neto

“Análise comparativa das técnicas de Explainable AI e um novo método para geração de explicações textuais”

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

Aprovado em: 10/03/2022.

BANCA EXAMINADORA

_____

Profa. Dra. Patricia Cabral de Azevedo Restelli Tedesco
Centro de Informática/UFPE

_____

Prof. Dr Jairson Barbosa Rodrigues
Colegiado de Engenharia da Computação/UNIVASF

_____

Prof. Dr. Germano Crispim Vasconcelos
Centro de Informática/UFPE
(**Orientador**)

Dedico este trabalho aos meus pais.

AGRADECIMENTOS

Gostaria de agradecer primeiro aos meus pais, Ywanoska e Pergentino, que tornaram essa jornada possível. Agradeço por todo incentivo à educação que sempre demonstraram. Por todo apoio para que isso fosse possível, pelo cuidado, carinho, amor e por acreditarem em mim. Sem dúvida, são minha maior inspiração.

Agradeço a toda minha família que me apoia e torce por mim. Em especial, ao meu tio, que me fez gostar de computação desde pequeno e se não fosse por ele, acredito que eu não estaria nessa área que me sinto tão realizado e feliz com o que faço. Obrigado por ser inspiração e por todas as dicas.

Um agradecimento muito especial também para minha namorada, Milena, que está sempre ao meu lado nos momentos bons e nos difíceis, que acompanhou essa jornada e sempre me deu forças para que eu conseguisse progredir. Obrigado por todo apoio e por ser tão companheira.

Não poderia deixar de agradecer aos meus amigos que torceram genuinamente por mim. Essas pessoas são fundamentais na minha vida, tornam ela mais leve, divertida, alegre e melhor. Dos amigos de infância a pessoas que começaram como colegas de trabalho, e com o passar do tempo se tornaram amigos e tão especiais para mim, muito obrigado.

Por fim, gostaria de agradecer ao meu orientador, Prof. Germano Vasconcelos, e ao meu coorientador, Prof. Cleber Zanchettin. Muito obrigado pela oportunidade, pelas orientações, ensinamentos e todo conhecimento transmitido durante esse tempo.

"Um passo à frente e você não está mais no mesmo lugar." (SCIENCE, 1996)

RESUMO

As soluções de Inteligência Artificial (IA), mais especificamente de Aprendizagem de Máquina (AM), têm alcançado crescentes níveis de desempenho devido à capacidade computacional, disponibilidade de dados e surgimento de novos métodos cada vez mais complexos. Essa complexidade tem aumentado a dificuldade de interpretar o funcionamento interno que conduz os modelos de AM na recomendação das decisões. Com objetivo de aumentar a interpretabilidade e manter a acurácia desses modelos complexos, surgiu a área de *Explainable AI* (XAI), com papel importante para impulsionar a confiança e o controle das soluções de IA. Este trabalho apresenta uma análise do estado da arte da área, propondo um mapa conceitual para organizar as taxonomias e abordagens. E realiza uma comparação entre as principais técnicas da literatura através de experimentos em uma base de dados reais para interpretar um modelo treinado para classificar o desempenho escolar, domínio no qual a interpretação dos resultados dos modelos é fundamental. Os resultados apontam as vantagens e desvantagens das abordagens, discussões sobre as saídas fornecidas, bem como uma forma de combinar estratégias. Diante das lacunas encontradas, um novo método é proposto nesta pesquisa, o Textual SHAP. O método busca endereçar necessidades da área como, por exemplo, considerar a perspectiva do usuário leigo como foco da explicabilidade. O método extrai as principais informações do gráfico da interpretação global do SHAP, técnica do estado da arte de XAI, e converte para um formato mais simples por meio de textos e apresenta em uma ferramenta com interface gráfica interativa. Foi realizada uma avaliação através de questionários com pessoas com conhecimento no domínio da educação e sem familiaridade com IA. Os resultados demonstraram que a abordagem proposta é menos complexa de interpretar e fornece maior nível de compreensão do que é exposto para os usuários. Com a abordagem de explicação textual, o método proposto apresenta potencial para alcançar explicações compreensivas e eficazes, contribuindo para os avanços das abordagens centradas nos humanos.

Palavras-chaves: explainable AI; interpretabilidade; aprendizagem de máquina; mineração de dados educacionais.

ABSTRACT

The Artificial Intelligence (AI) applications, more specifically Machine Learning (ML), have reached increasing levels of performance due to computational capacity, data availability and emergence of new and increasingly complex methods. This complexity has been increasing the difficulty of interpreting the internal mechanism which leads the ML model in decision recommendation. Aiming to improve the interpretability and maintain the accuracy of these complex models, Explainable AI has emerged, with an important role in boosting trust and control of AI solutions. This work presents an analysis of the state of the art in the area, proposing a conceptual map to organize taxonomies and approaches. And performs a comparison between the main techniques in literature through experiments on a real data set to interpret a trained model to classify school performance, a domain in which the interpretation of model results is fundamental. The results point out the advantages and disadvantages of the approaches, discussions about the outputs provided, as well as a way to combine strategies. Given the gaps found, a new method is proposed in this research, the Textual SHAP. The method seeks to address the needs of the area, for example, considering the lay user's perspective as the focus of explainability. The method extracts the main information from the chart from the global interpretation of SHAP, a state-of-the-art XAI technique, and converts it to a simpler format through texts and presents a tool with an interactive graphical interface. An evaluation was carried out with people with knowledge in the field of education and unfamiliar with AI. The results showed that the proposed approach is less complex to interpret and provides a higher level of understanding of what is exposed to users. With the textual explanation approach, the proposed method has the potential to achieve comprehensive and effective explanations, contributing to the advances of human-centered approaches.

Keywords: explainable AI; interpretability; machine learning; educational data mining.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

ALE	*Accumulated Local Effect*
AM	Aprendizagem de Máquina
AUC ROC	*Area Under the Receiver Operating Characteristic Curve*
IA	Inteligência Artificial
ICE	*Individual Conditional Expectation*
IHC	Interação Humano–Computador
k-NN	*k-Nearest Neighbors*
LIME	*Local Interpretable Model-agnostic Explanations*
Linear SVM	*Linear Support Vector Machine*
MDE	Mineração de Dados Educacionais
MLP	*Multilayer Perceptron*
PDP	*Partial Dependence Plot*
SARESP	Secretaria de Educação do Estado de São Paulo
XAI	*Explainable Artificial Intelligence*

SUMÁRIO

1 INTRODUÇÃO

1.1 CONTEXTO E MOTIVAÇÃO

Soluções de Inteligência Artificial (IA) estão amplamente presentes no cotidiano da sociedade, implantadas em diversas áreas (JORDAN; MITCHELL, 2015). Nas propagandas exibidas para o público alvo, em recomendações de produtos e filmes, na análise de crédito e nos diagnósticos médicos, são alguns exemplos em que soluções de IA são facilmente encontradas, ainda que as vezes nem notadas pelos usuários ou clientes. Diferentemente da visão fictícia do passado em que a IA surgia na forma de robôs, o que domina no mercado são modelos de Aprendizagem de Máquina (AM) realizando predições e influenciando decisões dentro de softwares.

Com aumento do poder computacional e os avanços nos algoritmos de AM, os modelos criados tornaram-se mais complexos, permitindo alcançar melhores performances que elevam as taxas de acertos, entretanto reduzem a interpretabilidade dos resultados (DOŠILOVIĆ; BRČIĆ; HLUPIĆ, 2018; GILPIN et al., 2018). Em alguns domínios, as predições afetam decisões sensíveis, o que reforça a necessidade de confiança e entendimento do resultado, bem como os casos em que o erro nas predições podem causar um grande impacto, por isso precisam de maior controle (DOSHI-VELEZ; KIM, 2017). Essas circunstancias são comuns em saúde (KNAPIČ et al., 2021), educação (QIN; LI; YAN, 2020), finanças (DEMAJO; VELLA; DINGLI, 2020) e direito (ANGWIN et al., 2016).

Esses modelos muito complexos não costumam fornecer interpretação do processo decisório, pois o objetivo principal de sua utilização são as predições geradas, o qual buscam encontrar a melhor performance. Em AM, a aplicação da modelagem com interesse apenas na saída fornecida a partir de um conjunto de entrada, é conhecida como caixa-preta (ou *black-box*). Os modelos com alta complexidade costumam ser chamados de *black-box* por não possibilitarem uma compreensão intuitiva, dentro da lógica humana, do conhecimento aprendido e respostas geradas pelo algoritmo. Um exemplo disso são as redes neurais, incluindo as redes neurais profundas, nas quais as estruturas internas formadas por camadas, pesos e conexões, não apresentam interpretação simples com termos habituais. Entretanto, devido ao crescimento de técnicas deste tipo e o aumento da utilização, surgiu a área *Explainable Artificial Intelligence*, também conhecida como *Explainable AI* ou XAI, ou pelo nome em português, porém menos utilizado, IA Explicável. Esta área busca endereçar o problema da falta de interpretabilidade

desses modelos e tem um crescimento notável na quantidade de buscas e publicações relacionadas à área nos últimos anos (LINARDATOS; PAPASTEFANOPOULOS; KOTSIANTIS, 2021; ADADI; BERRADA, 2018).

A educação é um dos principais domínios em que os modelos de AM precisam ser interpretados. Por exemplo, prever o desempenho das escolas em determinada avaliação não é suficiente para auxiliar à tomada de decisão dos gestores educacionais, pois apenas identificar previamente se o resultado será bom ou ruim não fornece insumos para entender e realizar ações. As técnicas de XAI fornecem mecanismos para interpretar os modelos de AM e justificar os resultados gerados. Nesse contexto, poderiam ser fornecidas as características que são mais relevantes para que as escolas tenham determinado desempenho, ou quais as condições que precisam ser satisfeitas para alcançar um resultado positivo. A análise também pode ser individual, isto é, para cada escola separadamente. Dessa forma, é possível entender o que ocorreu em um caso específico, explorar como as características dela impactaram o resultado, quais outras escolas são parecidas, o que deve ser alterado para modificar o desempenho, entre outras possibilidades. Portanto, XAI é fundamental para gerar explicações e possibilitar a utilização de modelos complexos, visto que a acurácia da solução também é desejada.

Esta área voltada para explicar os resultados da inteligência artificial, endereça a necessidade atual dos usuários entenderem e confiarem nas soluções, tal como fornece oportunidades para as aplicações. A motivação de explicar os sistemas de IA podem ser decorrentes de 4 razões, segundo Adadi e Berrada (2018), 1) justificar; 2) controlar; 3) melhorar; e 4) descobrir. Além de alguns domínios precisarem entender o que conduz as decisões, questões éticas e regulatórias tornaram-se assunto cada vez mais presente no contexto da utilização de IA, com intuito de proteger a sociedade do seu uso indevido e prejudicial (TJOA; GUAN, 2020). Desta forma, XAI é considerada como conceito central para conduzir a soluções de IA Responsável (em inglês, conhecido como *Responsible AI*), guiando para que atendam princípios éticos, de justiça e responsabilidade (ARRIETA et al., 2020; GUIDOTTI et al., 2018a).

Embora tenham surgido muitos métodos para aumentar a interpretabilidade dos modelos, e trabalhos para definir a taxonomia e agrupar as metodologias de acordo com suas peculiaridades, XAI é uma área relativamente nova e ainda em desenvolvimento. Isto reflete em problemas operacionais, desafios e questionamentos em aberto que devem ser considerados à medida que o campo de pesquisa progride (BELLE; PAPANTONIS, 2020). Guidotti et al. (2018a) fazem alguns questionamentos sobre o tema, como: *Quando uma explicação é compreensível? Qual a melhor forma de prover uma explicação? Como medir a qualidade da explicação?*

Quanto as interpretações são fidedignas ao processo decisório do modelo? Quanto é disposto a perder de acurácia e ganhar de interpretabilidade?

Dentre esses questionamentos, nota-se que existem muitas reflexões relacionadas ao público que consome as explicações. Inspirado em (ROSSI, 2019), Arrieta et al. (2020) apresenta um diagrama com público formado por diferentes perfis que devem ser considerados para consumir os resultados das interpretações dos modelos de AM. Segundo Rossi (2019), cada perfil tem um nível de conhecimento específico no domínio e uma capacidade analítica de analisar explicações da modelagem. Esse foco nos usuários finais e humanos, é dada como uma das principais lacunas de XAI atualmente (ISLAM et al., 2021).

O sucesso das explicações depende da percepção de quem recebe. Sendo assim, surge uma relação de XAI com a área Interação Humano–Computador (IHC), evidenciando a importância da experiência humana e do design, através de uma perspectiva sociotécnica. Este campo de pesquisa mais recente é denominado de *Human-Centered XAI*, adicionando mais interdisciplinaridade, e como o próprio nome diz, adicionando os humanos no centro das decisões (EHSAN; RIEDL, 2020; SCHOONDERWOERD et al., 2021; WANG et al., 2019).

Ignorar quem consome a explicação fornecida pode reduzir sua eficácia quando o interlocutor não entende, seja buscando obter *insights* ou fornecer alguma justificativa como resultado. Pensar nos humanos ou usuários que são a audiência das interpretações geradas, é uma forma de adaptar a linguagem e mecanismos utilizados para conseguir transmitir uma mensagem com clareza, considerando as diferentes necessidades (LIAO; VARSHNEY, 2021).

1.2 OBJETIVOS

Este trabalho tem como objetivos principais 1) investigar e comparar as técnicas existentes de *Explainable AI* direcionadas para dados tabulares[1], do tipo post-hoc[2] para explicações de modelos de Aprendizagem de Máquina de forma agnóstica[3], utilizando como referência uma aplicação na área de educação, que naturalmente apresenta-se como um setor importante para técnicas de XAI; 2) propor um novo método que enderece deficiências nas abordagens investigadas. Como decorrência, foi proposto um novo método, o Textual SHAP, que considera a explicabilidade dos modelos na perspectiva dos usuários, considerando inclusive usuários não

[1] Tipo de dados estruturados com informações numéricas e categóricas. A estrutura é uma tabela, em que as colunas costumam representar as variáveis e as linhas são as observações

[2] Forma de extrair explicações após o treinamento do modelo

[3] Funcionamento que independe do tipo do modelo

especialistas. Para alcançar estes objetivos, foram determinadas as seguintes atividades:

- Revisar a literatura para identificar os principais métodos de *Explainable AI*;
- Identificar vantagens e desvantagens nos formatos das saídas geradas;
- Propor uma abordagem, o Textual SHAP, baseada em explicações textuais para modelos de AM voltada para o público não especialista;
- Avaliar o método Textual SHAP de forma qualitativa através de questionário enviado a um grupo de usuários;
- Alcançar uma explicabilidade mais compreensível e mais simples com o método proposto, Textual SHAP, em comparação ao SHAP original na avaliação com o público alvo.

1.3 ESTRUTURA DA DISSERTAÇÃO

O Capítulo 2 apresenta a fundamentação teórica, partindo das definições dos principais conceitos e fundamentos relacionados à área. É proposta uma organização das taxonomias e abordagens, comparando suas diferenças em relação a outras organizações existentes, e são detalhadas as técnicas de *Explainable AI* na literatura que embasam o trabalho. Outras definições e aprofundamentos da literatura são apresentados no decorrer dos outros capítulos a medida que são necessários detalhes mais específicos.

No Capítulo 3 é realizada uma análise comparativa entre as principais técnicas da área. O estudo é conduzido através de experimentos para interpretar um modelo LightGBM treinado previamente para classificar o desempenho educacional. As técnicas de explicabilidade utilizadas foram selecionadas de modo a contemplar os tipos de abordagens apresentados na organização do estudo, tanto para explicar as predições individuais como o comportamento geral do modelo.

No Capítulo 4 é apresentado o método proposto da pesquisa, o Textual SHAP para geração de explicações textuais do SHAP. Além dos detalhes técnicos para a implementação do método, são apresentados resultados de sua aplicação no mesmo problema de classificação de desempenho escolar apresentado na análise comparativa. A metodologia e os resultados da avaliação com usuários reais finalizam o capítulo.

Por fim, o Capítulo 6 traz as considerações finais do trabalho. Enumerando as contribuições do trabalho, as limitações do método e sugerindo oportunidades para trabalhos futuros como extensão da pesquisa e avanços na área de XAI.

2 FUNDAMENTAÇÃO TEÓRICA

2.1 APRENDIZAGEM DE MÁQUINA

Aprendizagem de Máquina (do inglês, *Machine Learning*) é um subcampo da Inteligência Artificial que possibilita que computadores aprendam comportamentos, detectem padrões e tomem decisões com uma intervenção mínima de humanos. Segundo Mitchell et al. (1997), Aprendizagem de Máquina é definida formalmente como um programa aprende a partir da experiência E, em relação a uma classe de tarefas T, com medida de desempenho P, se seu desempenho em T, medido por P, melhora com E. Em outras palavras, o objetivo é aprender determinada tarefa a partir dos dados, de forma que é possível medir o desempenho e os algoritmos de AM buscam reduzir o erro para melhorar a experiência (MITCHELL et al., 1997).

Os problemas de Aprendizagem de Máquina são divididos em três tipos:

- Aprendizado supervisionado: o modelo recebe os dados rotulados, isto é, as saídas desejadas para cada dado de entrada. Com essa informação, tenta aprender uma regra para mapear as entradas as saídas. O objetivo é que o modelo consiga aprender padrões que generalizem e consiga determinar corretamente a saída para exemplos ainda não vistos. Os problemas supervisionados costumam ser divididos em classificação e regressão, para os dados discretos e os contínuos, respectivamente.

- Aprendizado não-supervisionado: nesse paradigma não é fornecido o rótulo. O modelo analisa os dados de entrada fornecidos e busca encontrar padrões, construir agrupamentos e associações. Os padrões são capturados através da exploração dos dados sem intervenção humana. O problema mais comum deste paradigma é o agrupamento, que, em suma, busca construir grupos de acordo com a similaridade dos dados.

- Aprendizado por reforço: tipo de modelagem baseada em ações e respostas. Também conhecida pelo padrão tentativa e erro, o qual o modelo realiza uma ação a partir da percepção do ambiente que está inserido e aprende de acordo com a resposta, ou seja, com a experiência que vai obtendo ao passar das iterações.

2.2 TERMINOLOGIAS DE EXPLAINABLE AI

Com o crescimento acelerado das pesquisas de interpretação e explicação para sistemas de inteligência artificial, diversas técnicas e terminologias surgiram sem uma clara definição ou medida de avaliação. Apontando este problema e propondo taxonomias, Doshi-Velez e Kim (2017) definem interpretabilidade no contexto de AM como a habilidade de explicar ou apresentar o funcionamento do modelo em termos compreensíveis para um humano.

Enquanto a explicabilidade, termo frequentemente utilizada na área de XAI, ainda que muitas vezes tratados de forma igual a interpretabilidade, apresenta algumas diferenças. A explicabilidade está associada a clarificar as lógicas e mecanismos internos dos modelos de AM, com objetivo de criar explicações relacionadas aos processos decisórios que devem ser tanto acuradas como compreensíveis (GILPIN et al., 2018). A partir dessas definições, a explicabilidade pode ser considerada como um conhecimento mais detalhado do processo decisório do modelo, enquanto a interpretabilidade não necessita de tantos detalhes, mas permite que seja compreendido o comportamento e os resultados gerados. Devido a sutileza entre as diferenças e o uso de forma intercambiável na literatura, ao longo do texto as palavras também serão utilizadas dessas forma.

Outras nomenclaturas comuns na área são os termos *understandable*, *comprehensible*, *intelligible* (ou em português, entendível, compreensível e inteligível). Adadi e Berrada (2018) consideram que os termos não são muito específicos para permitir uma formalização. Entretanto, por meio de um trabalho mais recente, (ARRIETA et al., 2020) apresenta as similaridades e distinções. A *understandability*, equivalente a *intelligibility*, denota a característica de um modelo fazer o ser humano entender seu funcionamento sem explicar o processo interno do modelo de aprendizagem de máquina. E a *comprehensibility*, está relacionada à capacidade de um modelo de aprendizagem de representar seu conhecimento aprendido de uma forma que seja compreensível para o usuário. Já o termo *Transparency*, bastante comum também, é empregado para referenciar os modelos que por si só são entendíveis, ou seja, apresentam interpretação em sua forma.

A partir dessas definições, ainda que algumas vezes existam bastante similaridades, o conceito de *understandability* (ou *intelligibility*) é um dos aspectos mais importantes em *Explainable AI*. Com objetivo unir a compreensibilidade dos modelos e a humana. Gunning (2017) define que XAI criará um conjunto de técnicas de ML que permitirá os usuários humanos entenderem, confiarem apropriadamente e gerenciarem os recursos de IA. A definição de (BAROCAS et al.,

2018) acrescenta que essa área tem por objetivo garantir as explicações das decisões algorítmicas, bem como quaisquer dados que conduzam essas decisões, possibilitando a explicação em termos não técnicos para usuários finais ou outras partes interessadas.

O aumento das aplicações com modelos inteligentes trouxeram debates relacionados à ética das soluções atuais e futuras, sobre a utilização de inteligência artificial de forma responsável, área denominada *Responsable AI* (BENJAMINS; BARBADO; SIERRA, 2019). Este campo de pesquisa está fortemente relacionado com XAI, que engloba outras motivações além das que foram citadas anteriormente, e busca estabelecer uma série de princípios necessários para implantação de aplicações reais de IA. Arrieta (2020) apresenta uma visão de XAI como conceito central para garantir os princípios de *Responsable AI*. A Comissão Europeia publicou as diretrizes éticas para construção de soluções confiáveis de IA (HLEG, 2019), os princípios são: agência e supervisão humana; robustez técnica e segurança; privacidade e governança de dados; transparência, diversidade, não discriminação e justiça; bem-estar social e ambiental; e responsabilidade.

2.3 ESTRATÉGIAS DE EXPLICABILIDADE

Existem diferentes pontos de vista que devem ser levados em conta para classificar uma técnica de XAI. A classificação baseada na estratégia ou escopo da técnica é fundamental para guiar na escolha do método ideal de acordo com um determinado contexto (LINARDATOS; PAPASTEFANOPOULOS; KOTSIANTIS, 2021). As principais taxonomias dos métodos são: escopo, tipo do modelo, momento da geração das explicações e o tipo dos dados.

O escopo é uma dimensão referente ao nível da interpretação, podendo ser Global ou Local. (DOSHI-VELEZ; KIM, 2017) apresenta essa característica, em que a abordagem global é referente ao modelo como um todo, com objetivo de explicar o comportamento geral e a lógica que o modelo segue, por exemplo, os principais fatores de influência. Enquanto a interpretação local explica as razões de uma decisão específica, isto é, para predição de uma instância individualmente.

O tipo do modelo classifica a abrangência da técnica à respeito dos modelos suportados através de uma dicotomia, sendo *Specific-Model* (específica ao modelo) ou *Model-Agnostic* (agnóstica ao modelo). Se a técnica for restrita a um modelo ou família específica, a técnica é considerada *Specific-Model*, por exemplo, uma técnica que funciona apenas para redes neurais artificiais. Por outro lado, quando o método é aplicável para qualquer modelo, é chamado de

Model-Agnostic. As abordagens da natureza do primeiro caso buscam interpretar baseado em alguma estrutura interna do modelo para conseguir extrair informações, por isso são específicas e não podem ser aplicadas em outros algoritmos. Enquanto no segundo, costumam analisar as predições do modelo, além de usar suas funções de predição/decisão e dados de entrada, abstraindo os mecanismos internos.

O momento da geração das explicações é formado por duas possibilidades, durante a construção do modelo ou através da exploração após o modelo ser treinado, denominadas de *Intrinsic Interpretable Model* (modelo interpretável intrínseco) e *Post-hoc explainability techniques* (técnicas de explicabilidade post-hoc), respectivamente. A abordagem intrínseca também é conhecida como *Transparent* ou *White-Box Model*, neste caso a interpretação dos modelos está contida em sua estrutura interna, eles são auto-explicáveis. Exemplos comuns de interpretação intrínseca são as árvores de decisão por meio das ramificações (QUINLAN, 1986) e os modelos lineares com seus coeficientes (WEISBERG, 2005). Já as explicações post-hoc estão associadas às técnicas que são aplicadas no modelo após seu treinamento para criar explicações. Os trabalhos (ADADI; BERRADA, 2018; GUIDOTTI et al., 2018b) apresentam esses conceitos associando a classificação de acordo com a complexidade dos modelos, em que as técnicas post-hoc são voltadas para modelos mais complexos, também chamados de *black-box*. Mesmo os modelos *black-box* necessitando de técnicas post-hoc para explicá-los, as técnicas desta natureza não se limitam a complexidade do modelo e podem ser aplicadas em modelos mais simples, em geral, quando é desejado obter uma explicação diferente da forma habitual do modelo interpretável. Ademais, a técnica ser post-hoc não garante que é agnóstica ao modelo, visto que pode ser uma forma de obter uma interpretação voltada para um tipo específico, por exemplo, um método para entender as ativações de uma rede neural.

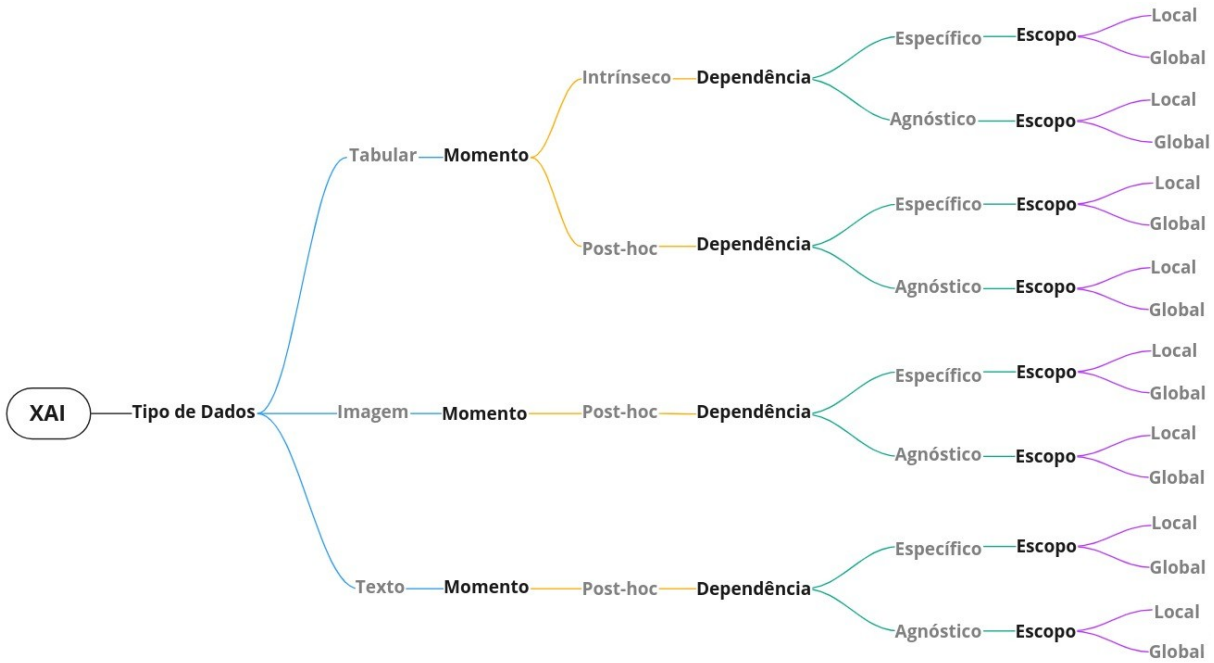
O tipo dos dados é outro fator que deve ser considerado para classificar ou selecionar o método, indicando em quais tipos eles podem ser aplicados. Os principais tipos de dados encontrados na literatura são: tabular, imagem e texto (GUIDOTTI et al., 2018b). A representação das informações em um formato tabular são uma maneira comum de utilizar os dados, os modelos de AM costumam receber esse tipo de entrada, mas é preciso um metadados para auxiliar os usuários na compreensão dos significados das colunas e valores presentes. Enquanto as informações apresentadas em imagens ou textos são mais fáceis de entender e não devem necessitar de arquivo auxiliar para explicação. Entretanto esses formatos precisam ser transformados em uma representação vetorial de modo que os modelos de AM consigam processar os dados e capturar os padrões. Após essa transformação, os dados utilizados não são mais

intuitivos. Devido a essas especificidades, o tipo de dados é um aspecto importante, pois tanto o modelo de AM pode ser específico quanto a forma de interpretar de acordo com a natureza dos dados, sendo assim, é fundamental na escolha e análise da técnica de XAI.

2.4 MAPA CONCEITUAL DE EXPLAINABLE AI

Baseado nas principais taxonomias e técnicas existentes, este trabalho propõe um mapa conceitual de XAI para guiar na escolha da abordagem ideal a partir do problema e as necessidades associadas. A figura 1 apresenta o mapa conceitual proposto das taxonomias de XAI.

Figura 1 – Mapa Conceitual das Taxonomias de XAI



Fonte: Elaborado pelo autor

A sequência adotada segue uma linha de raciocínio lógico para adoção de alguma técnica. Primeiro, deve ser considerado o tipo dos dados do problema em questão, ainda que algumas técnicas funcionem para tipos distintos de dados, essa escolha inicial ajudará a filtrar as possibilidades adequadas para o problema. Segundo, o momento em que a explicação é desejada, seja através de um modelo autoexplicável (intrínseco/transparente) ou numa análise posterior (post-hoc). A proposta considera um panorama geral das abordagens, em que os modelos construídos para imagens e textos costumam não apresentar interpretação em suas estruturas, enquanto modelos clássicos de aprendizagem de máquina para dados tabulares como Árvore de

Decisão e Regressão Linear, são exemplos de explicações intrínsecas. Terceiro, a dependência do tipo do modelo na explicação, é importante decidir se as explicações desejadas são para um modelo específico ou deve funcionar para qualquer tipo. Por último, o escopo da explicação para definir se será global ou local. É importante ressaltar que para um problema, pode ser explorado um conjunto de técnicas, podendo passar por ramificações diferentes, principalmente na última etapa, que define o escopo.

O trabalho (ADADI; BERRADA, 2018) apresenta uma pseudo ontologia da taxonomia dos métodos de XAI com informações equivalentes ao Escopo, Momento e a Dependência, ressaltando a importância desses aspectos para classificar as técnicas, entretanto não aborda o tipo dos dados para decisão. A taxonomia apresentada em (ARRIETA et al., 2020) não contempla o escopo da explicação como uma ramificação fundamental na escolha da técnica, mas sim como uma forma ou meio de explicar. Dessa forma, as Explicações Locais estão no mesmo nível que Explicação Visual, Explicação por Simplificação e Explicação de Relevância de Atributo, sendo estas as mesmas características de (BELLE; PAPANTONIS, 2020). Ainda que essa área tenha bastante sobreposição entre as abordagens e seja difícil isolar cada uma, esse tipo de classificação dificulta ainda mais a separação das técnicas, visto que todas as outras formas de explicar que foram apresentadas pelos autores também podem ser locais. Essa classificação não separa o que é global e local, informação de extrema importância para adotar a técnica desejada. Ainda sobre (ARRIETA et al., 2020), em cada técnica apresentada os autores marcam quais tipos de dados elas englobam, enquanto (BELLE; PAPANTONIS, 2020) não fazem menção a isto. Em (GUIDOTTI et al., 2018b) é apresentado uma taxonomia simples para "*open the black box problems*", limitada a modelos *black box versus transparent box* e ao escopo da explicação, porém esta última é dividida em *Model Explanation*, *Outcome Explanation* e Model *Inspection*.

Com o foco deste trabalho em modelos agnósticos para dados tabulares, o mapa conceitual proposto demonstra os tipos de explicações fornecidas nesta ramificação com as principais técnicas. A figura 2 apresenta o mapa conceitual das técnicas de XAI a partir do escopo, contemplando as explicações locais e globais, para tipo de dados tabular, de forma post-hoc e agnóstica ao modelo (o caminho pode ser visualizado na figura 1).

Nessa organização proposta, após definir o escopo da explicação desejada, o próximo passo é escolher qual abordagem da explicação, o que está fortemente associado a forma de explicar. São apresentadas abordagens comuns utilizadas pelos humanos para explicar decisões, considerando os trabalhos existentes na área de XAI (ADADI; BERRADA, 2018; ARRIETA et al.,

Figura 2 – Mapa Conceitual das Técnicas de XAI

Fonte: Elaborado pelo autor

2020; BELLE; PAPANTONIS, 2020). As formas apresentadas podem ocorrer tanto no escopo local quanto global, ainda que algumas sejam mais comuns para determinado escopo. As quatro abordagens selecionadas são:

- Visualização: permite que seja realizada uma interpretação através de visualizações de imagens e gráficos, por exemplo. O intuito é que seja fornecido insumos que facilitem a descoberta de conhecimento e explicações do comportamento. A visualização de dados é um recurso frequentemente utilizado por diversas técnicas, o que torna difícil de isolar os

métodos, entretanto são classificados nesta abordagem quando o mecanismo principal da explicação é a visualização.

- Relevância dos atributos: forma de explicar o comportamento do modelo por meio do ranqueamento e mensuração do impacto causado pelos atributos na saída do modelo. Após mensurar esse impacto, é possível inferir o foi mais relevante para decisão.

- Explicação baseada em instâncias: essa abordagem tem fundamentação na forma humana de explicar algo a partir de exemplos. Desta forma, o comportamento do modelo é explicado a partir de uma ou mais instâncias.

- Extração de regras: abordagem que busca extrair regras compreensíveis pelos humanos para representar o comportamento do modelo. Este tipo de explicação é desafiador, pois costuma ser esperado regras simples e de fácil entendimento, entretanto modelos *black-box*, em geral, são complexos.

Para cada ramificação da abordagem da explicação foram selecionadas pelo menos duas técnicas que a representassem, ressaltando que a mesma pode ocorrer em ambos os escopos (local e global). Para realizar essa seleção, foram considerados os principais os principais métodos de acordo com a revisão da literatura, identificando os mais populares e evoluídos, também foi baseado na disponibilidade dos códigos para realização de experimentos posteriormente. Alguns autores apresentam uma separação após a abordagem de explicação para determinar qual tipo da técnica baseado no método, segmentando ainda mais (ARRIETA et al., 2020; BELLE; PAPANTONIS, 2020). No diagrama proposto neste trabalho não é realizada essa separação, visto que são apresentados apenas os principais métodos do estado da arte, levando em conta questões práticas também, e não abordando todas as técnicas de maneira exaustiva.

2.5 TÉCNICAS DE EXPLAINABLE AI

Esta seção apresenta as principais técnicas da literatura analisadas no trabalho, direcionadas para o contexto post-hoc de XAI que são agnósticas ao modelo e focadas em dados tabulares. As técnicas são agrupadas conforme a organização ilustrada na figura 2, com abordagens do tipo: visualização, explicações baseadas em instâncias, relevância dos atributos e extração de regras.

2.5.1 Explicações Baseadas em Instâncias

Esse tipo de abordagem busca explicar através de exemplos o comportamento do modelo ou a distribuição dos dados subjacentes. Similar a uma das formas humanas de justificar alguma decisão com exemplos representativos.

A Seleção de Protótipos é utilizada para explicar o comportamento através de instâncias similares, encontrando grupos de instâncias que representem (KIM; KHANNA; KOYEJO, 2016) propôs o método *MMD-criticism* que além de capturar os protótipos dos dados bem representados, também foca nas críticas (*criticism*, ou também conhecidas como *outliers*). O método foi estendido por (GURUMOORTHY et al., 2019) na construção do algoritmo *ProtoDash*, que associa pesos não negativos correspondente à contribuição dos protótipos selecionados. As explicações desse método são comumente utilizadas no escopo local, indicando as instâncias similares como forma de embasar a decisão do modelo para uma predição. Entretanto, por possibilitar representações da base, podem ser utilizadas para o entendimento dos dados ou dos modelos através da separação dos dados após predição, sendo assim, considerada também como explicação global, como classificada em (CARVALHO; PEREIRA; CARDOSO, 2019).

Wachter et al. (2017) propuseram um método de interpretabilidade baseado em explicações contrafactuais, chamado de *Counterfactual*. No qual a explicação contrafactual descreve uma situação ou evento que não aconteceu, mas poderia ter acontecido. Com objetivo de identificar e apresentar os fatores externos necessários para mudar o resultado da predição, em que o modelo produz uma saída desejada, uma forma diferente de outros métodos que buscam clarificar as decisões e mecanismos internos. Considerando que as "causas" são os atributos e o "evento" é o alvo (ou saída), o algoritmo cria instâncias contrafactuais com um conjunto mínimo de modificações nos atributos que são suficientes para o modelo predizer a saída desejada. Para isto, é realizada a otimização da função de perda que computa as distâncias entre a predição da instância contrafactual e a saída predefinida, e da distância entre as características da instância original a ser explicada e da instância contrafactual gerada.

Para abordar algumas limitações apresentadas no algoritmo Counterfactual (WACHTER; MITTELSTADT; RUSSELL, 2017), foi proposta em (LOOVEREN; KLAISE, 2019) uma abordagem para explicações contrafactuais das predições de classificadores incorporando protótipos das classes. Através desses protótipos, foi possível aumentar a velocidade para encontrar as instâncias contrafactuais, sendo uma versão mais escalável, visto que o algoritmo original se tornava mais lento à medida que a dimensionalidade do espaço das características aumentava.

Os protótipos, construídos por meio de *autoencoders* ou *k-d trees*, são utilizados para guiar e acelerar a busca contrafactual para o protótipo mais próximo diferente da classe original, fazendo as perturbações se moverem para a distribuição típica de uma instância. Por fim, esta versão também apresenta uma forma eficiente de produzir perturbações nas características categóricas. Enquanto o algoritmo original não realizava nenhum tratamento específico para características desse tipo, lidando apenas com variáveis numéricas, o que poderia gerar uma instância contrafactual com valores impossíveis no mundo real.

A técnica *Contrastive Explanations Method* (CEM) (DHURANDHAR et al., 2018) gera uma explicação local através de exemplos para apresentar termos minimamente suficientes para produzir uma predição específica, e também quais características deveriam estar ausentes. Esses exemplos encontrados, representados por vetores de características, são denominados de Positivo Pertinente (PP) e Negativo Pertinente (NP), para manter e mudar a classificação, respectivamente.

Uma forma de interpretar através de exemplos, são as técnicas conhecidas como *Influential Instances*. Como o nome sugere, o intuito da técnica é analisar as instâncias mais influentes no modelo treinado, atuando na depuração do modelo, inspeção, identificação de vieses ou possíveis erros. *Deletion Diagnostics* é a técnica agnóstica ao modelo que fornece esse tipo de informação, ideia proposta inicialmente em (COOK, 1977) para computar a influência das instâncias na performance e nos parâmetros da Regressão Linear. A identificação é realizada através de um processo iterativo, em que cada instância do conjunto de treinamento é removida e o modelo é retreinado com esse novo conjunto, o desempenho e os parâmetros são medidos e as grandes oscilações indicam a influência. Cook (1977) propôs a técnica para regressão linear, mas a ideia foi estendida para qualquer modelo. Por meio de *Influential Instances* é possível obter explicações globais, sobre como o modelo se comporta após a remoção e o que está induzindo aqueles resultados, como também é possível inspecionar localmente os resultados, analisando o comportamento da predição individual.

A principal desvantagem de utilizar *Deletion Diagnostics* é associada ao custo computacional, dado que é necessário retreinar o modelo para cada instância da base. Endereçando este problema, foi proposta a técnica *Influence Functions*(KOH; LIANG, 2017), uma forma de encontrar as instâncias mais influentes sem retreinar o modelo, com a necessidade apenas de acesso a versão oráculo do modelo com acesso a gradientes e produtos de vetor de Hessiano. Alguns autores classificam como uma abordagem agnóstica ao modelo (BELLE; PAPANTONIS, 2020; LINARDATOS; PAPASTEFANOPOULOS; KOTSIANTIS, 2021), entretanto nem todos os modelos de

aprendizado de máquina utilizam parâmetros de pesos, o que torna restrito a algoritmos como regressão logística, redes neurais e support vector machine, e não aplicável a ensembles, por exemplo.

2.5.2 Relevância dos Atributos

As explicações através das relevâncias dos atributos são uma forma de explicar o funcionamento do modelo por meio da importância e influência dos atributos nas predições. As técnicas dessa natureza quantificam a contribuição de cada variável de entrada e determinam os fatores mais importantes nas decisões.

Uma das técnicas mais populares de interpretabilidade, o LIME (do inglês, *Local Interpretable Model-agnostic Explanations*) (RIBEIRO; SINGH; GUESTRIN, 2016), gera explicações de predições individuais de modelos *black-box* indicando a contribuição de cada variável com a magnitude e o sentido. Seguindo uma abordagem de *surrogate model*, o qual consiste em treinar um modelo interpretável para aproximar o funcionamento de outro modelo mais complexo. Este treinamento adjacente é realizado com um conjunto de dados gerado artificialmente através de perturbações na instância original que será explicada, criando então vizinhos neste espaço de entrada ponderadas pela proximidade. Na proposta original, os autores utilizam um modelo linear para aproximar e explicar as predições, capturando as interações entre os atributos localmente.

Para medir a importância dos atributos de forma global, o método *Permutation Feature Importance* foi introduzido por (BREIMAN, 2001) para o modelo *Random Forest*. O método consiste em realizar permutações nos valores dos atributos e depois medir o erro do modelo pré-treinado com estas variações. Os atributos importantes são identificados por meio do aumento do erro em função das permutações, já os atributos que são alterados e não afetam o resultado do modelo, indicam que o modelo não depende dele, logo são pouco relevantes. Baseado nesta ideia, foi proposta uma versão agnóstica ao modelo para calcular importância dos atributos chamada *Model Class Reliance* (MCR) (FISHER; RUDIN; DOMINICI, 2018).

Uma técnica que fornece interpretação global e local, com forte embasamento matemático e seguindo propriedades como *local accuracy, missingness e consistency*, é o método SHapley Additive exPlanations (SHAP) (LUNDBERG; LEE, 2017), que tem obtido bastante destaque na área de *Explainable AI* (LINARDATOS; PAPASTEFANOPOULOS; KOTSIANTIS, 2021). Com objetivo de explicar as predições do modelo através das contribuições de cada atributo, SHAP é

baseada na teoria dos jogos (MORGENSTERN; NEUMANN, 1953), utilizando a técnica *Shapley values* (SHAPLEY, 1953) por meio da média das contribuições marginais de todas as permutações. Através da agregação das contribuições dos atributos individualmente, SHAP fornece a interpretação global da importância dos atributos. Além disso, combina a importância com a distribuição dos *Shapley values* para melhor compreensão do efeito dos atributos no modelo. A principal desvantagem do algoritmo é o custo computacional, visto que ele cobre todas as possibilidades, entretanto há otimizações que tornam o processo mais rápido, principalmente para modelos baseados em árvores.

2.5.3 Visualização

Explicações através de visualizações é uma forma de apresentar o comportamento e as relações capturadas pelo modelo através de visualizações de dados, como gráficos, e promover uma forma simples de usuários não familiarizados com AM entenderem.

Para explicar as relações entre os atributos e o alvo por meio de visualizações, (FRIEDMAN, 2001) propôs o método Partial Dependence Plot (PDP). A técnica post-hoc e agnóstica ao modelo cria uma representação gráfica para visualizar o efeito marginal de uma ou mais variáveis nas predições médias do modelo de AM. O PDP tem por objetivo mostrar como os atributos estão impactando de acordo com seu comportamento. Podendo capturar relações complexas, é possível realizar a análise simultaneamente com até três atributos, gerando uma representação 3D, porém esta é a quantidade máxima devido a limitação de construir gráficos de fácil entendimento com mais dimensões.

Com objetivo de criar o mesmo tipo de visualização do PDP, porém endereçando algumas limitações, (APLEY; ZHU, 2020) propôs o método Accumulated Local Effects (ALE). O PDP assume independência entre as características, e não lida bem com casos em que existe uma alta correlação, resultando na inclusão de áreas preditas de forma irrealistas. Enquanto o ALE substitui a distribuição marginal por uma distribuição condicional para mitigar esse problema de correlação, tornando-se preferível para estes casos.

Enquanto o PDP apresenta explicações globais do efeito médio dos atributos, o método Individual Conditional Expectation (ICE) (GOLDSTEIN et al., 2015) foi proposto para gerar explicações de predições individuais. ICE plot apresenta uma visualização com linha para cada instância com a dependência entre os valores do atributo sendo analisado e o valor predito.

Uma extensão do PDP e ICE, é o método *Ceteris Paribus Profile* (BIECEK, 2018). Neste

trabalho, os autores apresentam uma metodologia para explicações locais em cenários do tipo *"What If"*, em que a visualização dos dados informa como os atributos impactam na predição à medida que os valores são modificados para determinada instância. O termo *Ceteris Paribus* vem do Latim e significa "todo o resto inalterado", ou seja, as variações são realizadas apenas nos atributos de interesse enquanto os outros permanecem o mesmo. A técnica *Ceterius Paributs Profile* é fortemente embasada no PDP e ICE, fornecendo algumas visualizações iguais, entretanto um dos principais diferenciais é que o método fornece a visualização de uma instância separadamente, enquanto o ICE apresenta todas juntas. Os autores investem na visualização de dados, fornecendo gráficos bem elaborados e interativos.

Diferente das técnicas apresentadas para gerar determinada explicação, o grupo de pesquisa PAIR (*People + AI Research*) desenvolveu o *What-If Tool* (WIT) (WEXLER et al., 2019), uma ferramenta voltada para interpretação dos modelos e exploração de cenários. A ferramenta integra algumas técnicas de explicabilidade já consolidadas na literatura, mas além disso, introduz a aplicação do Facets Dive para fornecer uma ampla visualização dos dados. Enquanto a técnica *Ceteris Paribus Profile* menciona os cenários hipotéticos com a modificação de valores, o WIT permite que o usuário faça as modificações em sua interface e indica o impacto causado, além disso, é possível alterar diversos valores simultaneamente.

As explicações visuais agnósticas ao modelo, costumam ser mais escassas, principalmente para explicações locais, devido complexidade de generalizar. As visualizações das explicações também estão presentes em outras abordagens, como por exemplo, a relevância dos atributos, o que dificulta a separação de abordagens que usam os recursos visuais como motivação principal da resposta (ARRIETA et al., 2020). Existem muitas abordagens no escopo local voltada para imagens ou específicas para modelos de *Deep Learning*, enquanto para o contexto tabular e agnóstico ao modelo, são menos comuns.

2.5.4 Extração de Regras

O algoritmo Anchors (RIBEIRO; SINGH; GUESTRIN, 2018), propõe a geração de explicações locais no formato de regras *if-then*. Essas regras são denominadas de âncoras, pois são suficientes para garantir determinada predição mesmo com modificação em valores de outros atributos, apresentando alta precisão e clara cobertura. A abordagem de âncora combina os benefícios das explicações locais agnósticas ao modelo com a interpretabilidade das regras, resultando em uma forma mais simples e compreensível para humanos.

Voltado para fornecer explicações locais por meio de regras, o algoritmo *LOcal Rule-based Explanations* (LORE) (GUIDOTTI et al., 2018a) treina um modelo interpretável com dados sintéticos gerados por meio de algoritmos genéticos para representar a vizinhança da instância a ser explicada. O método proposto considera árvores de decisões no modelo interpretável, e extrai as regras posteriormente para explicar as decisões de um modelo. Além das regras para representar a decisão do modelo para determinada instância, a técnica fornece também regras contrafactuais para indicar modificações mínimas que mudem o resultado da predição.

Para extrair regras de modelos complexos, uma abordagem clássica é a utilização de um *surrogate model*, treinando algoritmos de construção de regras ou árvores para tentar aproximar o comportamento original, porém fornecendo padrões interpretáveis. Seguindo essa abordagem, destaca-se o algoritmo TREPAN (CRAVEN; SHAVLIK, 1995), proposto inicialmente para extrair representações simbólicas de redes neurais. Entretanto, por tratar a rede neural como um oráculo, o método pode ser estendido para outros classificadores. A árvore de decisão aprendida por TREPAN é similar a algoritmos clássicos como CART e C4.5, porém com algumas modificações fundamentais no processo de aprendizagem. As principais são na expansão dos nós, que ao invés da forma tradicional de expandir em profundidade por meio de *depht-first*, o algoritmo usa a abordagem *best-first*, considerando como melhor o nó que apresenta maior potencial de aumentar a fidelidade entre a árvore de decisão e o oráculo. Outro ponto importante é na amostragem realizada durante o treinamento, que através do particionamento da árvore, os dados tendem a diminuir, porém no método proposto novos exemplos são construídos e o oráculo é utilizado para gerar o rótulo.

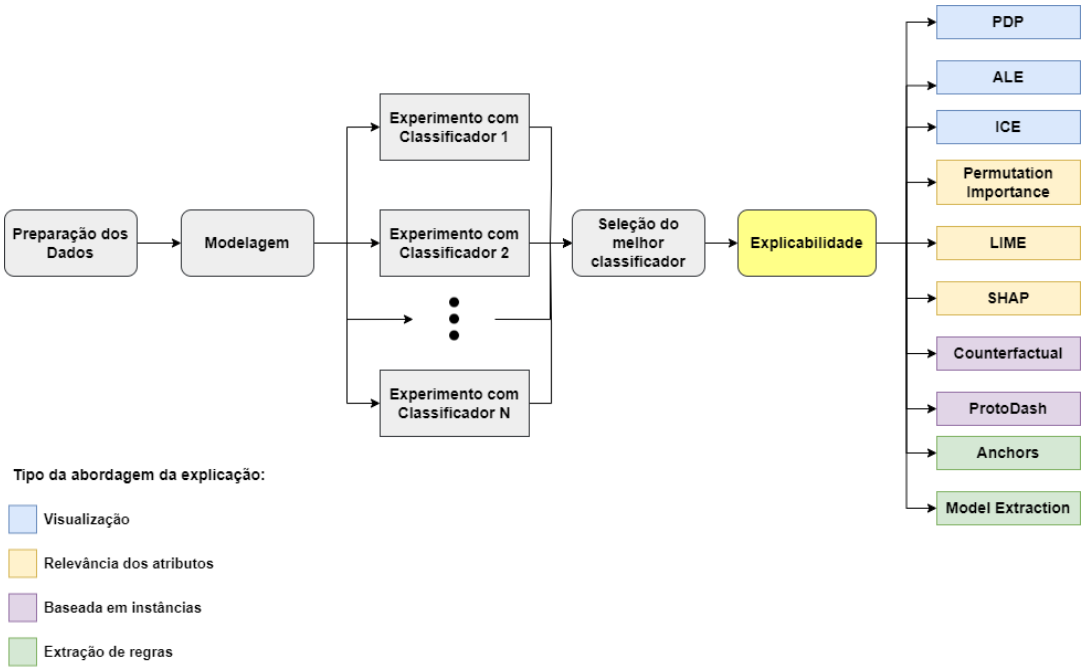
Uma outra abordagem que tem similaridade com o ramo anterior, é o método *Model Extraction* (BASTANI; KIM; BASTANI, 2017). Esta abordagem é voltada para extrair uma simplificação de um modelo black-box. Este *surrogate model* é utilizado para construir uma árvore de decisão de forma gulosa (*greedy*). Na primeira etapa o algoritmo gera um Modelo de Mistura Gaussiana para agrupar os dados do conjunto de treinamento utilizando Maximização da Expectativa, que posteriormente servem como entrada para construir uma árvore de decisão com algoritmo CART, adaptando a função de ganho.

3 AVALIAÇÃO EXPERIMENTAL E RESULTADOS

A figura 3 apresenta a metodologia experimental adotada para comparação das técnicas. O fluxo ilustrado consiste em duas etapas principais. A primeira, com caixas cinzas, voltadas para a modelagem do desempenho das escolas. E a segunda, com caixas coloridas, os experimentos com diferentes técnicas de XAI, contemplando as ramificações das abordagens de acordo com a organização da literatura proposta na pesquisa (seção 2.4).

No detalhamento dos experimentos realizados são informadas as bibliotecas públicas com a implementação dos métodos utilizados.

Figura 3 – Metodologia experimental



Fonte: Elaborado pelo autor

3.1 CLASSIFICAÇÃO DO DESEMPENHO ESCOLAR

Os experimentos realizados neste trabalho com as diferentes técnicas de *Explainable AI* foram direcionados para interpretar o mesmo modelo previamente treinado, um classificador *LightGBM* (KE et al., 2017) para determinar o desempenho das escolas. Esta seção apresenta uma contextualização do problema, dos dados utilizados e da modelagem, o estudo completo foi publicado em (NETO; VASCONCELOS; ZANCHETTIN, 2021).

A Mineração de Dados Educacional (MDE) (BAKER; ISOTANI; CARVALHO, 2011) é um campo de pesquisa com objetivo de explorar os conjuntos de dados coletados no contexto da educação, o qual busca descobrir padrões dos alunos, professores, processos de aprendizagem e gestão para educação. Um tipo de aplicação comum é a predição do desempenho escolar (LACRUZ; AMÉRICO; CARNIEL, 2019; SILVA et al., 2018; PINTO et al., 2019), provendo a indicação do desempenho esperado a partir de algumas características. Nesse tipo de abordagem, saber previamente o desempenho de um aluno ou escola não é suficiente, é necessário entender o que influencia para alcançar o respectivo desempenho para conseguir levantar ações acionáveis.

Por meio dos dados da Secretaria de Educação do Estado de São Paulo, o estudo apresentado classifica o desempenho das escolas em bom ou ruim a partir das notas do SARESP (Sistema de Avaliação de Rendimento Escolar do Estado de São Paulo) de 2018. A avaliação é realizada através das provas de Língua Portuguesa e Matemática com estudantes no Ensino Fundamental I e II, e Ensino Médio. Foi empregada a metodologia CRISP-DM para construir uma solução de MDE.

Além do SARESP, outros dados públicos da Secretaria de Educação[1] foram utilizados para coletar características das escolas e criar atributos preditivos. Dentre as informações estavam dados das classes e turmas, endereços, históricos de mudanças na gestão nos últimos 5 anos, servidores ativos na rede de ensino, carga horária e formação dos profissionais. Além desses, alguns dados externos foram integrados na base, como um compilado de informações sobre os municípios brasileiros[2] e resultados da Avaliação Nacional da Alfabetização (ANA) de 2016. Os dados coletados foram processados por meio de engenharia de características para construir uma base de dados tabular para ser utilizada na modelagem. A tabela 1 resume os atributos.

[1] Disponíveis em: https://dados.educacao.sp.gov.br/
[2] https://www.kaggle.com/crisparada/brazilian-cities

Tabela 1 – Resumo dos atributos utilizados pelo modelo de AM

Nome	Descrição
MUNICIPIO CAPITAL	Indica se é capital ou não (variável binária)
MUNICIPIO AREA	Área do município
MUNICIPIO POPULACAO	População do município
MUNICIPIO AREA RURAL	Indica se é área rural ou não (variável binária)
MUNICIPIO VALOR ACRESCENTADO BRUTO	Produto Acrescido Público
MUNICIPIO PIB PER CAPITA	Produto Interno Bruto (PIB) per Capita
DEPENDENCIAS	Quantidade de dependências. Uma coluna para cada informação: salas de aula, sala dos professoreres, laboratório de ciência e laboratório de informática.
FORMACAO	Distribuição da formação dos professores. Um coluna para cada formação com seu percentual (sem informação, ensino médio, bacharelado, licenciatura, especialização, mestrado, doutorado)
QTD SERVIDORES	Quantidade de servidores
QTD PROFESSORES	Quantidade de professores
MEDIA FORMACOES	Valor médio da formação dos professores (conversão das categorias para números, quanto maior a formação maior o valor)
QTD FORMACAO CONTINUADA	Quantidade de professores com pós-graduação
QTD CARGOS DISTINTOS	Quantidade de cargos distintos
QTD TOTAL ALUNOS	Quantidade total de alunos
QTD CLASSES	Quantidade de classes
MEDIA ALUNOS SALA	Média de alunos nas salas
QTD CLASSES TIPO ENSINO	Quantidade de classes de acordo com o tipo de ensino. Uma coluna para cada modalidade (fundamental e médio).
QTD ALUNOS TIPO ENSINO	Quantidade de alunos de acordo com o tipo de ensino. Uma coluna para cada modalidade (fundamental e médio).
DIRETORES QTD 2018	Quantidade de diretores em 2018
COORDENADORES QTD 2018	Quantidade de coordenadores em 2018
DIRETORES QTD 5 ANOS	Quantidade de diretores nos últimos 5 anos
COORDENADORES QTD 5 ANOS	Quantidade de coordenadores nos últimos 5 anos
DIRETOR IDADE	Idade do(a) diretor(a)
DIRETOR CARGO CLAS EXER IGUAIS	Indica se o diretor tem o cargo de contrato igual ao de exercício (informação binária)
DIRETOR ANOS TRAB CARGO C	Anos de trabalho do diretor no cargo de contrato
DIRETOR ANOS TRAB CARGO E	Anos de trabalho do diretor no cargo de exercício
JORNADA QTD DISCIPLINAS MEDIA	Quantidade média de disciplinas dos professores
JORNADA QTD DISCIPLINAS MAX	Quantidade máxima de disciplinas dos professores
JORNADA QTD TOTAL AULAS MEDIA	Quantidade média de aulas dos professores
JORNADA QTD TOTAL AULAS MAX	Quantidade máxima de aulas dos professores
SERVIDORES IDADE MEDIA	Idade média dos servidores
SERVIDORES TEMPO CARGO C MEDIA	Tempo médio de contrato dos servidores
SERVIDORES CAT FUNCIONAL	Distribuição dos servidores de acordo com a categoria funcional. Uma coluna com percentual para cada tipo (A, F e O)
ANA NOTAS	Notas no exame ANA em Matemática, Escrita e Leitura (uma coluna para cada)
RELACAO ALUNO POR SERVIDOR	Relação entre o número de alunos para o de servidores
RELACAO ALUNO POR PROFESSOR	Relação entre o número de alunos para o de professores

Fonte: Baseada em (NETO; VASCONCELOS; ZANCHETTIN, 2021)

A base de dados final gerada após a etapa de preparação foi composta por 4.523 escolas, que representam as observações desse conjunto tabular de dados, e teve 58 atributos. O desempenho escolar, alvo do problema, foi definido com uma classificação binária para determinar se o desempenho foi ruim ou bom. A discretização da avaliação do SARESP considera os níveis de proficiência esperados para cada classe e respectiva prova realizada. Através da agregação das notas obtidas em cada nível, é formado o indicador para a escola, granularidade tratada neste trabalho.

Os dados foram divididos em conjunto de treinamento e teste, com a proporção de 80% (3.618 escolas) e 20% (904 escolas), respectivamente. O conjunto de treinamento foi utilizado para definir o melhor modelo para o problema através da experimentação. Para garantir a robustez da análise, foi empregada a técnica de validação cruzada, com o método *k-fold* (com $k = 10$). Os resultados obtidos são apresentados na tabela 2. Foram utilizados os classificadores *Decision Tree*, *Random Forest*, *Gradient Boosting*, *LightGBM*, *k-NN*, *MLP* e *Linear SVM*, todos utilizando os hiperparâmetros padrões, e as métricas de avaliação Acurácia, AUC ROC (do inglês, *Area Under the Receiver Operating Characteristic Curve*), e o valor macro das métricas de Precisão, *Recall* e *F1-score*, que computa a média do resultado obtido para cada classe.

Tabela 2 – Resultados dos experimentos com diferentes classificadores

Classificador	AUC ROC	Acurácia	Precisão	Recall	F1-score	Tempo
Decision Tree	$0,80$	$89,0\%$	$0,79$	$0,80$	$0,79$	**1s**
Random Forest	**0,96**	$92,3\%$	**0,85**	$0,87$	**0,86**	$5s$
Gradient Boosting	$0,95$	$91,9\%$	$0,84$	$0,86$	$0,85$	$17s$
k-NN	$0,91$	$90,5\%$	$0,83$	$0,81$	$0,82$	**1s**
LightGBM	**0,96**	$91,5\%$	$0,84$	$0,85$	$0,84$	**1s**
MLP	**0,96**	$91,5\%$	$0,84$	$0,85$	$0,84$	$21s$
Linear SVM	**0,96**	**92,4%**	**0,85**	**0,88**	**0,86**	$7s$

Fonte: Baseada em (NETO; VASCONCELOS; ZANCHETTIN, 2021)

Na análise dos resultados observou-se o altíssimo desempenho obtido, alcançando $0,96$ na AUC ROC (quando comparado ao máximo de 1 para a métrica), indicando uma alta capacidade do modelo de discriminar as classes Bom e Ruim. Entretanto, os modelos mais simples, *Decision Tree* e *k-NN*, apresentaram desempenhos inferiores em comparação com os outros mais complexos.

Para selecionar o melhor modelo, foi realizado o teste estatístico não paramétrico Kruskal-

Wallis (KRUSKAL; WALLIS, 1952) com múltiplas amostras para comparar se os resultados obtidos pelos classificadores são diferentes. O teste foi realizado em duas etapas, ambas com nível de significância de $0,05$. 1) com todos os classificadores, resultando um *p-value* de aproximadamente 0, permitindo rejeitar a hipótese nula que as amostras são iguais, e fornecendo forte evidência que as distribuições são diferentes. 2) removendo os classificadores mais simples que obtiveram menor desempenho (*Decision Tree* e *k-NN*), o *p-value* do teste estatístico 0.09, que falha em rejeitar a hipótese nula, então é possível assumir que são da mesma distribuição.

Como os modelos complexos apresentaram resultados estatisticamente iguais, o tempo necessário para treinar e predizer foi o critério para seleção final. O *LightGBM* foi o mais rápido dentre os que obtiveram melhores resultados, por isso foi selecionado. O conjunto de teste foi utilizado para obter a avaliação final do desempenho e para a fase de interpretabilidade. Neste conjunto, o *LightGBM* obteve $93,3\%$ de acurácia e AUC ROC de $0,97$.

3.2 ANÁLISE DAS TÉCNICAS

Esta seção apresenta uma análise comparativa dos resultados das técnicas de *Explainable AI*. A análise é composta por experimentos com diferentes técnicas para interpretar um modelo de aprendizado de máquina voltado para predizer o desempenho escolar, detalhado na seção anterior.

As explicações são voltadas para dados tabulares, de maneira *post-hoc*, do tipo agnósticas ao modelo e tanto o escopo global como o local foram explorados. As técnicas selecionadas contemplam pelo menos uma possibilidade para cada tipo de abordagem de acordo com a organização proposta (representado na figura 2), que são a visualização, relevância dos atributos, explicação baseadas em instâncias e extração de regras. Após a aplicação de cada técnica, os resultados gerados são discutidos e analisados.

Um dos principais desafios da área, é a metodologia para avaliar as técnicas, dado que, cada uma pode apresentar sua saída em um formato diferente. Esse é um dos motivos para existir uma dificuldade em medir de modo automatizado, ao contrário das métricas de avaliação para algoritmos de aprendizagem de máquina supervisionados. Uma forma comum nas abordagens centradas nos humanos, ocorre através de entrevistas e validações com usuários, entretanto a análise comparativa das técnicas da literatura neste trabalho, não contempla essa avaliação, ao invés disso apresenta uma discussão dos diferentes aspectos.

Os experimentos são realizados com o mesmo modelo e base de dados com intuito de

discutir as vantagens e desvantagens, exibindo questões técnicas e sobre a forma de exibir as explicações. Enquanto outros estudos da literatura apresentam apenas a parte teórica dos métodos (CARVALHO; PEREIRA; CARDOSO, 2019; ADADI; BERRADA, 2018; ARRIETA et al., 2020) ou suas aplicações em diferentes contextos (LINARDATOS; PAPASTEFANOPOULOS; KOTSIANTIS, 2021; VILONE; LONGO, 2020; CONFALONIERI et al., 2021), o que torna mais difícil comparar os resultados. Existem algumas limitações por não considerar o mesmo contexto, em alguns casos os tipos dos dados são diferentes, por exemplo, explicação para classificação de imagens e outro para classificação binária com dados tabulares.

3.2.1 Visualização

As explicações baseadas em visualizações têm um apelo maior para exploração de recursos visuais que ajudem o usuário a interpretar o comportamento do modelo. Muitos métodos do estado da arte são voltados para problemas que o tipo de dados são imagens, e algumas técnicas são específicas para modelos como redes neurais profundas. Para estas, o objetivo costuma ser mostrar as estruturas aprendidas e principais ativações, outras para mostrar as principais regiões de uma imagem para o modelo interpretado. As técnicas selecionadas para os experimentos são voltadas para dados do tipo tabular e agnósticas ao modelo.

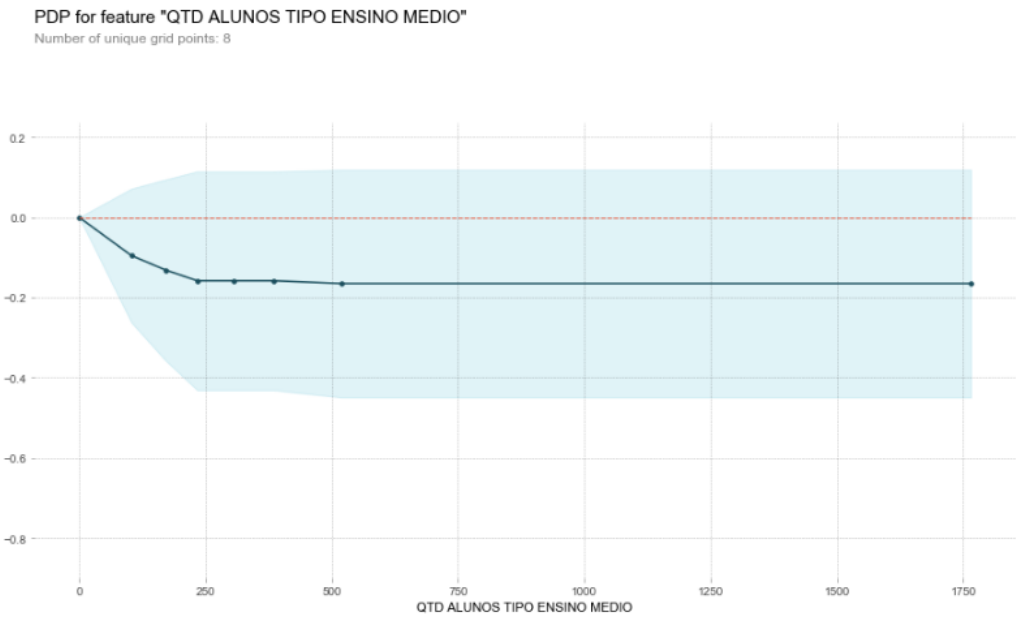
3.2.1.1 PDP

O *Partial Dependence Plot*[3], ou apenas PDP, é o método mais conhecido desta classificação. Os recursos visuais apresentados nesta técnica serviram como embasamento para diversas outras. O objetivo principal é analisar o comportamento de um atributo ao longo de suas variações. O método é de cunho global, pois as explicações são para o modelo como um todo, porém voltadas para um atributo individualmente ou no máximo três. A figura 4 apresenta o resultado da aplicação da técnica no modelo treinado para explicar o atributo *QTD ALUNOS TIPO ENSINO MEDIO*.

A forma que o método possibilita explorar o comportamento de um atributo fornece informações de como o impacto ocorre de acordo com o valor do atributo, o PDP é capaz de capturar relações não lineares. No eixo Y da figura 4 está presente o valor do impacto na saída do modelo, enquanto no eixo X estão os valores do atributo analisado. Neste exemplo,

[3] https://pdpbox.readthedocs.io/en/latest/

Figura 4 – PDP com um atributo



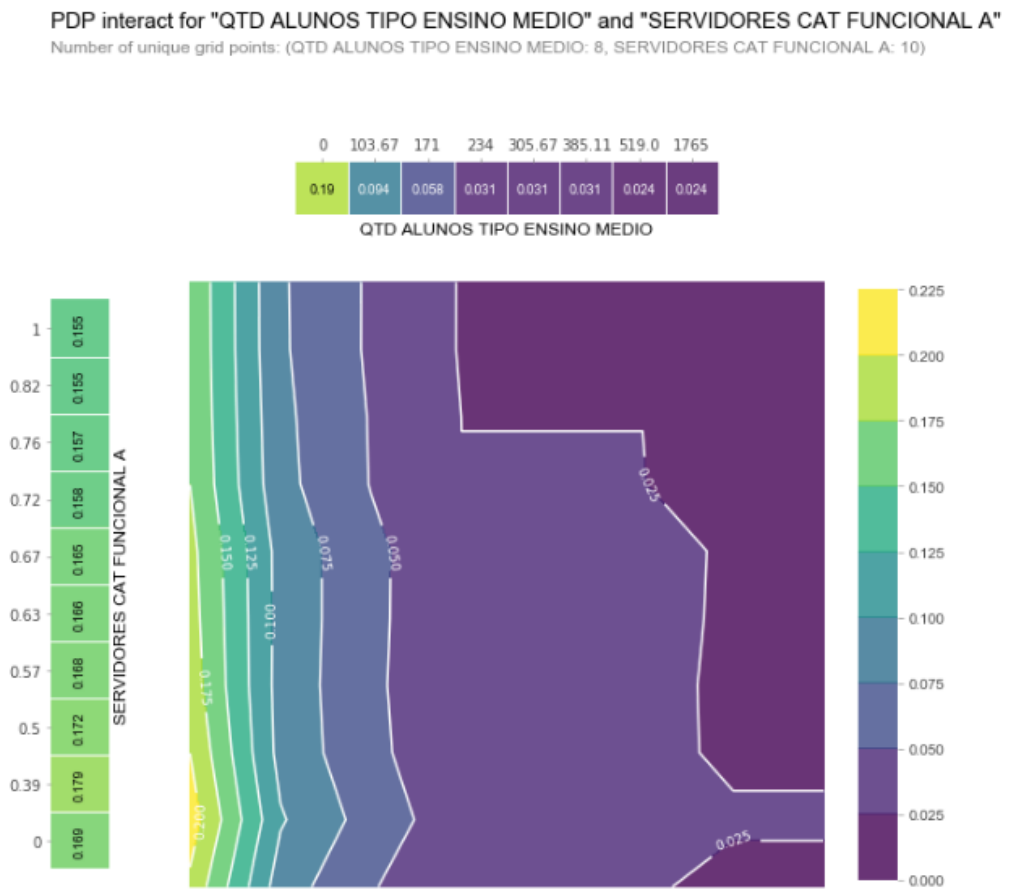
Fonte: Elaborado pelo autor

é possível perceber que após o atributo alcançar determinado limiar, quando a *QTD ALUNOS TIPO ENSINO MEDIO* é aproximadamente 250, as contribuições saturam e não variam significativamente.

Alguns métodos da literatura, como *Permutation Importance*, estão focados na magnitude da contribuição global, sem grandes preocupações com o sentido ou forma que eles ocorrem, enquanto o PDP tem foco justamente oposto. O método apresenta uma forma útil de explorar a relação entre o atributo e a contribuição. Numa regressão linear o peso atribuído a determinada característica será sempre o mesmo, produzindo uma reta no impacto à medida que os valores são alterados. Com PDP é possível capturar comportamentos não lineares na oscilação do impacto, identificando regiões mais críticas, pontos de inflexão, entre outras interpretações.

Uma das principais limitações do PDP é que sua interpretação deve ser realizada para atributos de forma individual, pois aumentar a quantidade de atributos implica no aumento das dimensões do gráfico, o que torna humanamente impossível uma análise com diversas dimensões. A figura 5 apresenta uma visualização de duas dimensões, com a variável *QTD ALUNOS TIPO ENSINO MEDIO* no eixo X e *SERVIDORES CAT FUNCIONAL A* no eixo Y. Este tipo de visualização é uma forma de identificar a relação entre dois atributos. É possível também gerar um gráfico com três variáveis, resultando numa visualização 3D, o que aumenta bastante a complexidade para interpretar.

Figura 5 – PDP com dois atributos



Fonte: Elaborado pelo autor

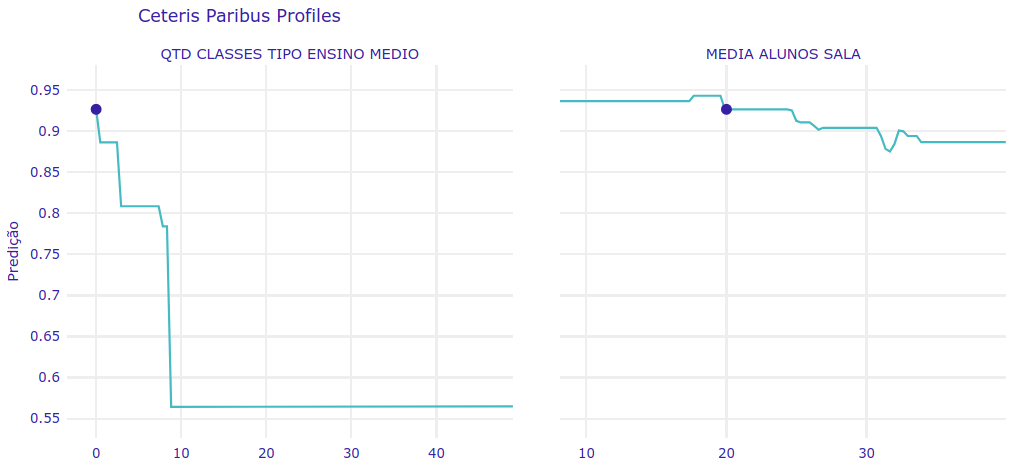
Na figura 5, a região inferior da lateral esquerda contém as colorações mais claras, que indicam quando ocorrem os maiores impactos na saída com a combinação entre as variáveis analisadas. Os detalhes do mapa de calor da contribuição na predição estão apresentados na barra à direita da imagem, enquanto as outras barras estão relacionadas com os valores e impactos de cada atributo separadamente.

A análise individual que o PDP fornece não considera as correlações com os outros atributos, sendo essa uma das principais limitações do método, o que não afeta a forma de visualizar, mas sim os valores exibidos.

3.2.1.2 ICE

O método ICE (*Individual Conditional Expectation*)[4] é uma versão equivalente ao PDP, porém voltado para explicações locais. Com objetivo de visualizar como variações no atributo estão relacionadas com as predições de instâncias individualmente, o método constrói uma visualização similar ao PDP, mas apresentando cada instância separadamente. Na figura 6 está o resultado da aplicação no modelo de predição de desempenho escolar.

Figura 6 – ICE



Fonte: Elaborado pelo autor

Na figura 6 são exibidos os comportamentos de dois atributos, *QTD CLASSES TIPO ENSINO MEDIO* e *MEDIA ALUNOS SALA*. A construção do gráfico de cada atributo é realizada de forma independente, em que o método considera o impacto que cada um afeta na predição à medida que seus valores são alterados e as outras informações são mantidas constantes. De acordo com os resultados, é possível perceber que o primeiro atributo sofre uma influência maior quando seus valores aumentam, dado que o valor original da instância, ponto no gráfico, está no valor mínimo. Enquanto as variações do segundo atributo causam um impacto menor.

A forma de interpretar o resultado localmente através do ICE permite identificar comportamento e mudanças para alterar a predição de forma bastante intuitiva. O método costuma apresentar uma visualização com várias instâncias no gráfico simultaneamente, entretanto a visualização de forma individual, como mostrado no exemplo, ou com poucos valores, torna mais simples e melhor para a proposta local. Assim como o PDP, o método tem a limitação

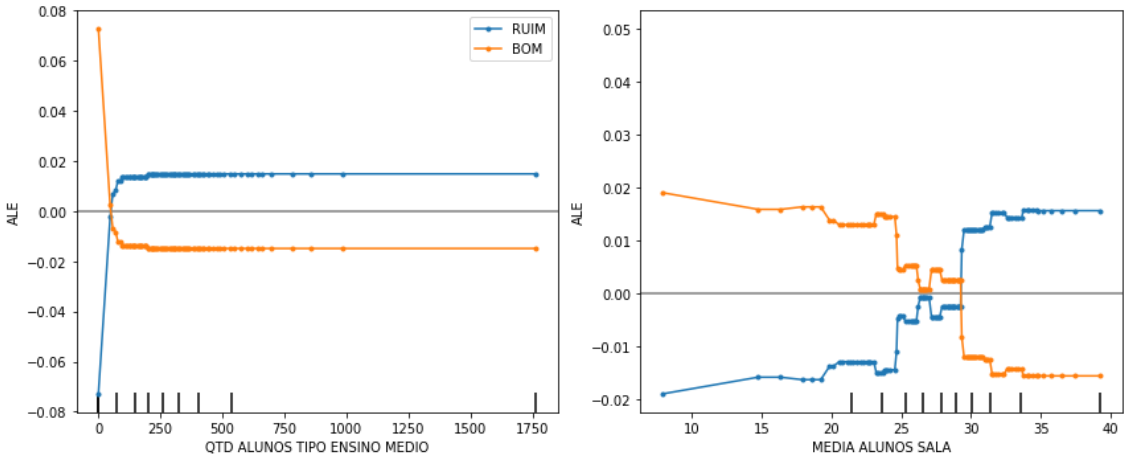
[4] https://dalex.drwhy.ai/python/api/

da quantidade de dimensões a serem consideradas. Ainda que seja possível identificar regiões que impactam na predição, a visualização pode induzir a um cenário que não seria possível dado as outras configurações da instância, isto é, uma simulação que não reflete na realidade, um cenário improvável ou até impossível, visto que as outras características permanecem inalteradas.

3.2.1.3 ALE

O resultado gerado pelo método ALE[5], apresentado na figura 7, é bem similar ao PDP. A principal diferença está no cálculo realizado, em que o ALE computa o efeito médio nas predições, enquanto o PDP lida com os efeitos marginais. O método ALE funciona razoavelmente bem com atributos correlacionados, diferente do PDP.

Figura 7 – ALE



Fonte: Elaborado pelo autor

Algumas diferenças podem ser percebidas devido a forma que os diferentes pacotes implementaram os gráficos e exploraram os recursos visuais. Na figura 7 são apresentadas as curvas para as duas classes, enquanto na figura 4 apenas uma classe é considerada, dado que o comportamento da outra é simétrico não há perda de informação. Por este motivo, do ponto de vista do usuário, os dois métodos entregam o mesmo tipo de informação.

Nos exemplos da figura 7, estão os atributos *QTD ALUNOS TIPO ENSINO MEDIO* e *MEDIA ALUNOS SALA*. No primeiro caso, o resultado gerado é bastante similar ao PDP, com a mesma saturação na contribuição em determinada faixa conforme os valores do atributo

[5] https://docs.seldon.io/projects/alibi/en/stable/

aumentam. Já no segundo exemplo, é possível notar um comportamento não linear com muitas variações. A quantidade baixa de alunos na sala tem um impacto positivo, quando os valores aumentam muito, com uma média por volta de 30 alunos, o modelo considera o impacto para o desempenho negativo, isto é, um desempenho escolar ruim.

3.2.2 Relevância dos Atributos

Entender a relevância dos atributos de um modelo, é uma forma muito útil para obter *insights*, avaliar vieses, descobrir quais informações irrelevantes ou quais são as prioritárias, entre outras utilidades, que servem como oportunidades para tanto para pessoas de negócio como também cientistas de dados melhorarem a solução. Enquanto na regressão linear os coeficientes são um indicativo do que é mais importante e na árvore de decisão por meio da posição dos atributos, há uma grande necessidade de entender a importância dos atributos de black-box.

3.2.2.1 Permutation Importance

Com intuito de obter a relevância global do modelo, isto é, quantificar o impacto que cada atributo gera de forma geral nas predições, o método *Permutation Importance*[6] é baseado na permutação, alterando os valores dos atributos e depois avaliando as variações causadas na saída do modelo previamente treinado. A figura 8 apresenta o resultado obtido através da aplicação do método no modelo analisado.

A saída deste método fornece uma tabela ordenada pelos atributos mais importantes, indicando os respectivos pesos encontrados e uma coloração que varia a intensidade à medida que os atributos são mais relevantes. O resultado é a relevância, representada por um valor absoluto, o qual pode impactar de forma positiva ou negativa à medida que seus valores aumentam, porém o método não identifica isto, sendo uma das primeiras limitações encontradas.

No resultado apresentado, o modelo considera o atributo *QTD ALUNOS TIPO ENSINO MEDIO* como mais relevante com diferença significativa para o segundo colocado.

[6] https://eli5.readthedocs.io/en/latest/

Figura 8 – Permutation Importance

Weight	Feature
0.0904 ± 0.0053	QTD ALUNOS TIPO ENSINO MEDIO
0.0195 ± 0.0037	QTD CLASSES TIPO ENSINO MEDIO
0.0076 ± 0.0023	FORMACAO BACHARELADO/TECNIOLOGO
0.0073 ± 0.0017	FORMACAO ENSINO MEDIO
0.0046 ± 0.0018	SERVIDORES TEMPO CARGO C MEDIA
0.0032 ± 0.0014	JORNADA QTD TOTAL AULAS MEDIA
0.0030 ± 0.0013	ANA MATEMATICA
0.0017 ± 0.0010	JORNADA QTD DISCIPLINAS MAX
0.0017 ± 0.0008	SERVIDORES CAT FUNCIONAL O
0.0015 ± 0.0011	DIRETOR ANOS TRAB CARGO E
0.0013 ± 0.0009	FORMACAO DOUTORADO
0.0010 ± 0.0006	FORMACAO ESPECIALIZACAO
0.0010 ± 0.0009	COORDENADORES QTD 5 ANOS
0.0008 ± 0.0006	JORNADA QTD DISCIPLINAS MEDIA
0.0007 ± 0.0003	DEPENDENCIAS TOT SALA ANA LEITURA
0.0007 ± 0.0006	SERVIDORES CAT FUNCIONAL A
0.0007 ± 0.0006	MUNICIPIO MUNICIPIO AREA
0.0007 ± 0.0007	MEDIA ALUNOS SALA
0.0006 ± 0.0006	QTD CLASSES TIPO ENSINO FUNDAMENTAL
0.0005 ± 0.0006	SERVIDORES CAT FUNCIONAL F
	... 38 more ...

Fonte: Elaborado pelo autor

3.2.2.2 LIME

Enquanto a técnica anterior, *Permutation Importance*, apresentou explicações globais, o algoritmo LIME constrói explicações locais, isto é, calcula a importância de cada atributo para uma dada instância. A figura 9 apresenta o resultado obtido na aplicação do LIME [7] para uma escola que o modelo indiciou um bom desempenho. A saída conta com as probabilidades das classes, o efeito causado por cada atributo com indicativo do sentido, ou seja, qual classe o impacto foi direcionado, e os valores dos atributos da instância analisada.

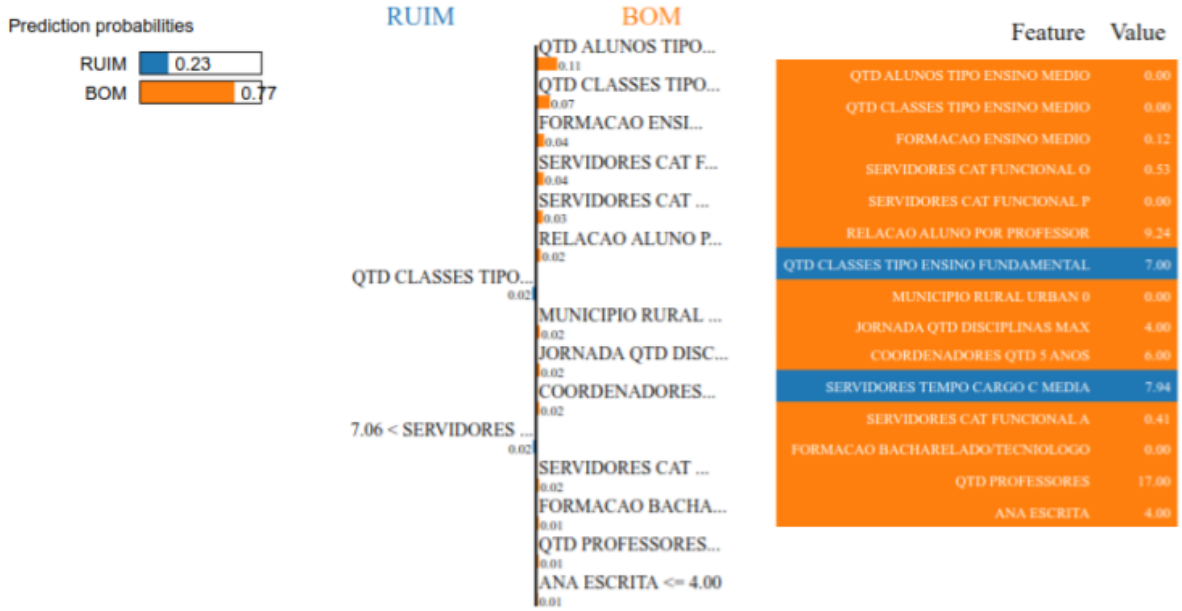
O método é formado por uma solução simples e eficiente, construindo um modelo transparente na vizinhança da instância interpretada. Dessa forma é possível alcançar uma generalização, funcionando para qualquer modelo e apresentando um tempo curto de resposta, visto que o *surrogate model* construído ao entorno da instância original possui uma estrutura simples.

O algoritmo não apresenta uma explicação completa, mas sim um resultado a partir de uma aproximação. A definição da vizinhança pode ser sensível, causando explicações diferentes para exemplos similares. Os parâmetros do *surrogate model* podem variar e são mais um fator que pode modificar a explicação gerada.

A análise de relevância local permite encontrar cenários que as características mudam de impacto de acordo com seu contexto, ao invés de encontrar um valor fixo para todos os casos.

[7] https://github.com/marcotcr/lime

Figura 9 – LIME



Fonte: Elaborado pelo autor

Entretanto, espera-se que para dados similares o comportamento seja parecido. O LIME é um dos métodos mais importantes da literatura de XAI, sua abordagem de explicação local inspirou outros trabalhos posteriormente.

3.2.2.3 SHAP

Diferente dos métodos de relevância dos atributos apresentados, que só contemplam um dos escopos de interpretabilidade, o SHAP [8] fornece ambos, tanto as explicações globais como locais. O método inspirado na teoria dos jogos Shapley e no LIME, implementa versões otimizadas para diferentes tipos de modelo com intuito de melhorar a performance e manter-se agnóstico ao modelo.

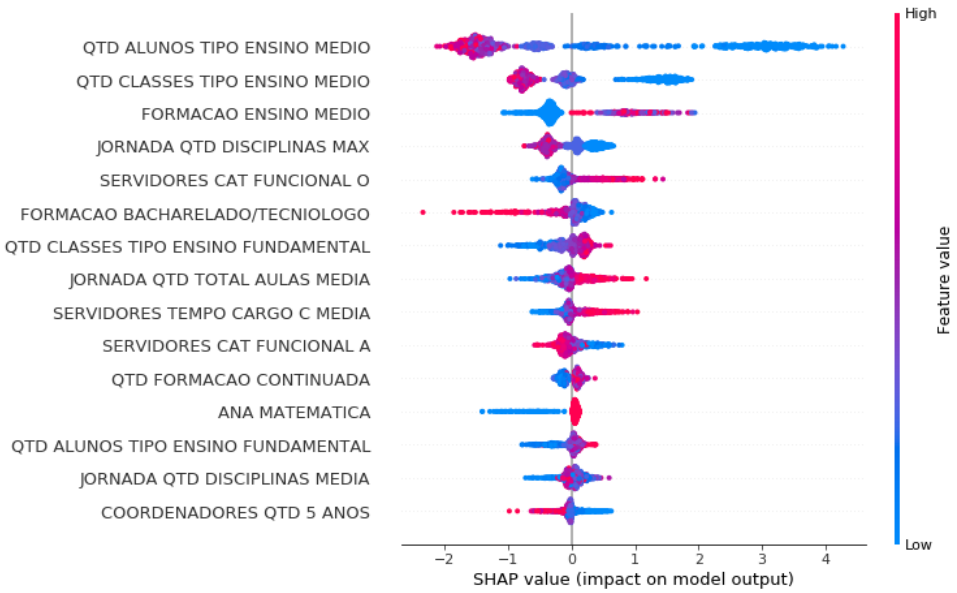
O resultado da interpretação global, apresentado na figura 10, indica os principais atributos, através da ordenação decrescente da lista no eixo Y, o impacto causado por cada predição em análise, representada por cada ponto, no eixto X está seu valor SHAP, por fim, as cores são utilizadas para indicar o valor presente no atributo. Esse gráfico global, conhecido como gráfico de resumo (em inglês, *summary plot*), consiste em exibir em um único gráfico os valores SHAP de cada predição, ou seja, cada explicação local. Através desse gráfico é possível perceber qual

[8] https://shap.readthedocs.io/en/latest/index.html

sentido os atributos impactam, à esquerda com valores negativos, ou seja, para o desempenho escolar ruim, à direita com valores positivos, direcionando para o desempenho escolar bom. As colorações usam uma escala para entre os menores valores até os maiores, dessa forma todos os atributos podem compartilhar da mesma representação, pois não há detalhes do valor real.

O gráfico global apresenta uma gama de informações. Há bastante riqueza e profundidade nos detalhes, visto que os pontos mostram a distribuição do impacto e não apenas um valor agregado e a coloração insere uma nova dimensão para ser analisada. Esse grau de detalhe também atribui uma alta complexidade para analisar e extrair explicações do comportamento geral do modelo.

Figura 10 – SHAP global



Fonte: Elaborado pelo autor

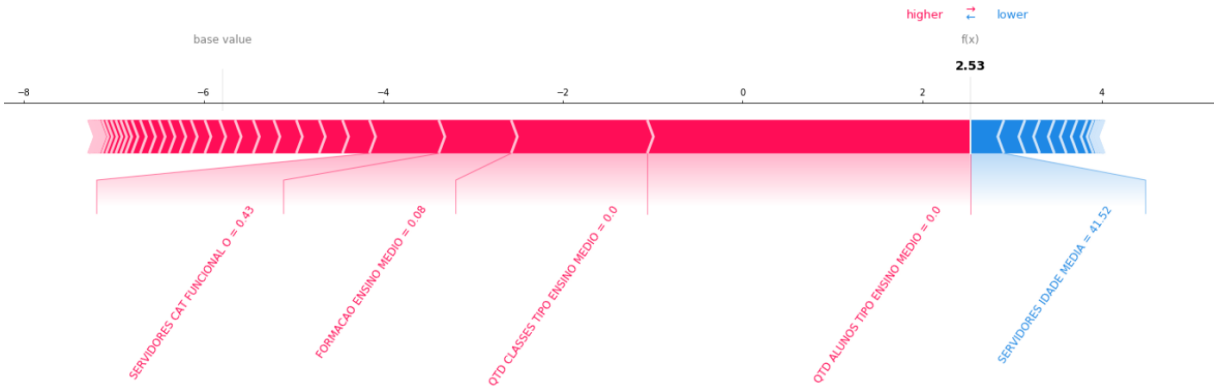
Neste exemplo, a variável mais importante foi a *QTD ALUNOS TIPO ENSINO MEDIO*, o mesmo foi encontrado através do *Permutation Importance*. A maior concentração do impacto ocorre em uma parte do valor SHAP negativo, com pontos roxos e vermelhos, o que seria equivalente aos valores medianos e os maiores, respectivamente. Essa similaridade no impacto da saída do modelo para esses valores diferentes, é um comportamento similar ao que foi indicado pelo PDP (figura 4), o qual a partir de determinado valor de entrada, a contribuição permanecia constante. Ainda em comparação com o PDP, as maiores influências exercidas ocorreram quando os valores deste atributo apresentaram valores muito baixos, no gráfico global do SHAP, é possível notar pelo eixo X esse mesmo comportamento, neste caso representado pelos pontos azuis na extremidade da direita.

O formato que o método constrói o gráfico com a dispersão dos valores, impacto causado e as concentrações, tudo isso para vários atributos simultaneamente, é uma visão bem detalhada, mas para extrair o melhor entendimento do resultado, cada linha do gráfico deve ser analisada individualmente. Mesmo que esta interpretação global seja formada pela junção de explicações locais, não há identificação no gráfico a respeito da instância e nem é possível identificar a contribuição de diferentes atributos para a mesma instância.

A ordem que os atributos aparecem é uma forma simples de identificar os mais relevantes, porém para obter essa informação, o método calcula a média dos valores SHAP absolutos de cada característica da base. Também vale ressaltar que o resultado do SHAP captura relações não-lineares, como pode ser notado na contribuição do atributo *JORNADA QTD DISCIPLINAS MEDIA*, o qual valores pequenos (azuis), contribuíram para ambas as classes, assinalando valores SHAP positivos e negativos.

Enquanto na abordagem local, conforme o resultado apresentado na figura 11, as barras vermelhas indicam atributos que contribuíram para o desempenho positivo da escola, enquanto as azuis são referentes a características negativas, ou seja, tendência para o desempenho ruim. Além das cores, há um formato de seta nas barras para apontar a direção da contribuição. O gráfico é conhecido como gráfico de forças, pois mostra como cada atributo força a predição para determinada classe. O tamanho da barra significa a magnitude de sua importância. O exemplo apresenta a interpretação de uma predição com desempenho bom, o que pode ser notado pelo resultado, $f(x)$, superior ao *base value* (valor base). O valor base costuma ser a média da saída do modelo no conjunto de treinamento, o intuito é criar uma referência para o caso de não conhecer as características de entrada e comparar com as contribuições posteriormente.

Figura 11 – SHAP local



Fonte: Elaborado pelo autor

A explicação local por meio do SHAP fornece um gráfico que auxilia na interpretação, em que todos os atributos são apresentados em uma única linha com tamanhos diferentes para sua relevância. Dessa maneira, é possível identificar rapidamente o que gerou maior impacto, também ter uma noção do que as pequenas contribuições representam quando somadas. A comparação entre positivos de um lado e negativos do outro, facilita a entender o valor final. Com o tipo de informação semelhante à do LIME, porém apresentada de maneira diferente, o SHAP explora melhor os recursos visuais, visto que um gráfico de barra com muitos atributos pode ser inviável de analisar na prática.

Atualmente SHAP é um dos métodos mais populares (LINARDATOS; PAPASTEFANOPOULOS; KOTSIANTIS, 2021; BELLE; PAPANTONIS, 2020; LIAO; VARSHNEY, 2021), principalmente entre os cientistas de dados. O método tem por objetivo fornecer a relevância dos atributos, porém os autores investiram bastante na visualização das informações, produzindo diferentes gráficos e formas de apresentar o resultado. O valor SHAP é uma estrutura atômica para as análises, por este motivo existe uma consistência na relação entre o escopo local e global, o que pode não ser encontrado quando são combinados métodos distintos, por exemplo o LIME e o PDP. Com informações bem complexas, o SHAP necessita de bom conhecimento em análise de dados, aprendizagem de máquina e no próprio método para poder ser compreendido.

3.2.3 Explicações baseadas em instâncias

A abordagem de explicação através de instâncias é baseada no paradigma utilizado pelos humanos para explicar algo através de exemplos. Para isso, costumam ser selecionados exemplos representativos, destacar a diferença entre alguns casos específicos, apresentar alguma justificava pela similaridade, entre outros. Os experimentos desta seção apresentam alguns métodos desta natureza, contemplando exemplos de explicações globais e locais.

3.2.3.1 Counterfactual

O algoritmo Counterfactual[9], segue o paradigma que há uma relação de causa e efeito para construir sua explicação. Em outas palavras, a abordagem é baseada em "Se o evento X ocorreu, o resultado será Y", ou, "se o evento X não ocorreu, o resultado não será Y". O método costuma ser utilizado para modificar uma instância, criando o que seria equivalente

[9] https://docs.seldon.io/projects/alibi/en/stable/

a uma situação hipotética, representando desta maneira o que seria necessário para mudar o resultado da predição do modelo. A figura 12 apresenta a saída do algoritmo Counterfactual através de uma tabela formada pelas seguintes colunas: valores originais da instância (X), valores da instância contrafactual gerada (CF) e a diferença entre elas (X - CF), nesta ordem.

Figura 12 – Counterfactual

FORMACAO DOUTORADO	0.000000	0.000000	0.000000
FORMACAO ENSINO MEDIO	0.253378	0.253378	0.000000
FORMACAO ESPECIALIZACAO	0.034749	0.041614	0.006860
FORMACAO LICENCIATURA	0.861004	0.861004	0.000000
FORMACAO MESTRADO	0.000000	0.000000	0.000000
FORMACAO S/INFO	0.000000	0.000000	0.000000
QTD SERVIDORES	0.215278	0.215278	0.000000
QTD PROFESSORES	0.198529	0.198529	0.000000
MEDIA FORMACOES	0.034749	0.041614	0.006860
QTD FORMACAO CONTINUADA	0.035714	0.035714	0.000000
QTD CARGOS DISTINTOS	0.400000	0.400000	0.000000
QTD TOTAL ALUNOS	0.263936	0.263936	0.000000
QTD CLASSES	0.238095	0.238095	0.000000
MEDIA ALUNOS SALA	0.742453	0.742453	0.000000

Fonte: Elaborado pelo autor

O método apresenta o resultado de uma instância contrafactual que satisfaz as condições estabelecidas para mudar a predição, entretanto há múltiplas instâncias que também podem atender os requisitos, estarem corretas e até refletir situações melhores para o negócio. Encontrar a instância ideal é mais um desafio desta metodologia.

No exemplo em questão, a instância gerada por meio do algoritmo indica o que precisa ser modificado para que a escola que obteve um desempenho ruim conseguisse atingir um desempenho positivo. A instância original deste caso, apresenta probabilidade de 76% de ser da classe negativa, enquanto a nova instância tem probabilidade de 64% de ser da classe positiva, ambas informações segundo o modelo treinado previamente.

Similar a outros métodos que utilizam perturbação ou alguma variação para modificar a instância, ocorre o risco de criar uma instância irreal, isto é, que os dados não reflitam a algo possível no mundo real. Este problema comum, tenta ser minimizado no Counterfactual através da limitação dos espaços disponíveis nos atributos com a indicação dos mínimos e máximos,

e por meio da utilização de *Autoencoder* para manter a fidelidade do dado. O *Autoencoder* é um tipo de rede neural não supervisionado utilizado para aprender a representação dos dados, dessa forma, eles são úteis para manter a instância construída dentro das características da base de dados.

Para encontrar a instância contrafactual, o algoritmo utiliza uma otimização multi-objetiva, que considera a quantidade de atributos modificados, o valor que será alterado, entre outras informações a fim de minimizar esses critérios. Indicar esse valor ideal para o usuário, apresenta a vantagem de detectar de forma automatizada quais são as modificações necessárias e mínimas, em contrapartida do usuário precisar simular diferentes situações.

O tipo de interpretação fornecida por este método é uma forma muito boa de explicar algumas circunstâncias, podendo ser aplicada em diversos segmentos. A instância resultante necessita de uma análise, contudo quando comparada com os dados originais e direcionando o foco para as diferenças encontradas, a explicação fica mais simples para entregar a um usuário final.

3.2.3.2 ProtoDash

Os protótipos gerados pelo algoritmo ProtoDash[10] são uma forma de detectar as instâncias importantes para representar a base de dados. O algoritmo apresenta uma forma versátil de abordar diferentes escopos, servindo numa etapa pré-modelagem para exploração dos dados, mas também para explicações globais e locais. Na exploração dos dados, os protótipos são encontrados de acordo com a classe alvo, enquanto na interpretação do modelo são utilizadas as predições. A figura 13 apresenta uma abordagem global baseado em instâncias para interpretar as predições de bom desempenho escolar, o resultado contempla três protótipos, indicado nas colunas com seus respectivos pesos na última linha, enquanto as demais linhas são valores dos atributos das instâncias prototípicas.

Diferentemente da explicação contrafactual que constrói uma situação hipotética, o algoritmo ProtoDash encontra as instâncias prototípicas na base de dados selecionada e retorna o índice e o peso de cada um. A partir do índice é possível recuperar os dados originais e apresentá-los, conforme a figura 13.

A forma que o algoritmo fornece explicações locais, fundamenta-se em identificar qual respectivo protótipo para uma dada instância em análise. Este procedimento segue a intuição

[10] https://aix360.readthedocs.io/en/latest/

Figura 13 – ProtoDash

	Protótipo 1	Protótipo 2	Protótipo 3
QTD ALUNOS TIPO ENSINO MEDIO	0.000000	25.000000	0.000000
QTD CLASSES TIPO ENSINO MEDIO	0.000000	3.000000	0.000000
FORMACAO ENSINO MEDIO	0.119403	0.105263	0.041667
JORNADA QTD DISCIPLINAS MAX	6.000000	8.000000	2.000000
SERVIDORES CAT FUNCIONAL O	0.182927	0.241379	0.166667
FORMACAO BACHARELADO/TECNIOLOGO	0.014925	0.052632	0.083333
QTD CLASSES TIPO ENSINO FUNDAMENTAL	44.000000	9.000000	13.000000
JORNADA QTD TOTAL AULAS MEDIA	51.962963	41.368421	42.000000
SERVIDORES TEMPO CARGO C MEDIA	6.597561	9.517241	7.916667
SERVIDORES CAT FUNCIONAL A	0.451220	0.310345	0.750000
QTD FORMACAO CONTINUADA	7.000000	1.000000	1.000000
ANA MATEMATICA	4.000000	4.000000	4.000000
QTD ALUNOS TIPO ENSINO FUNDAMENTAL	1327.000000	84.000000	334.000000
JORNADA QTD DISCIPLINAS MEDIA	3.259259	3.684211	2.000000
COORDENADORES QTD 5 ANOS	12.000000	7.000000	8.000000
Pesos dos Protótipos	0.610000	0.240000	0.150000

Fonte: Elaborado pelo autor

de que o exemplo selecionado para explicar é similar e representativo para indicar o motivo da decisão do algoritmo.

Enquanto alguns métodos fornecem recursos visuais para auxiliar a interpretação, esta abordagem de explicar baseado em instâncias fornece o índice, peso e uma tabela com os valores dos atributos. É necessário esforço para interpretar e entender o que cada exemplo representa, quais as diferenças entre os protótipos, quais atributos variaram, relações entre eles, entre outras análises. O exemplo apresentado é um recorte dos principais atributos, porém a base de treinamento é composta por mais atributos, o que tornaria ainda mais complexa a análise posterior.

A metodologia para criar interpretações a partir de instâncias representativas da base de dados podem ser realizadas com outros algoritmos de seleção de protótipos. Por exemplo, o algoritmo clássico *k-medoids*, baseado no *k-means*, mas ao invés de encontrar pontos no espaço de busca, retorna instâncias existentes. Porém o algoritmo *ProtoDash*, além de avanços na forma de selecionar os dados, apresenta um método com objetivos voltados para interpretabilidade, para isto são incluídas, além dos protótipos, instâncias críticas, ou *outliers*, e o peso associados a cada um desses. O peso demonstra a importância e representativa daquele protótipo na base de dados, eliminando a necessidade de separar o resultado em protótipos e

críticas, pois o peso já fornece um indicativo relacionado a isto.

3.2.4 Extração de Regras

Uma das abordagens comuns em modelos transparentes são as explicações por meio de regras, como *Rules List* ou Árvore de Decisões que criam estruturas baseada em regras. As técnicas *post-hoc* de XAI buscam extrair regras de modelos que não possuem esse formato em sua concepção original. Esse tipo de explicação costuma ser proposto devido a sua natureza de descrever as decisões do algoritmo, sugerindo um formato próximo do que os humanos estão adaptados, isto é, através de condições para alcançar determinado resultado. Os experimentos realizados com algoritmos de explicabilidade desta natureza permitem avaliar os resultados gerados.

3.2.4.1 Anchors

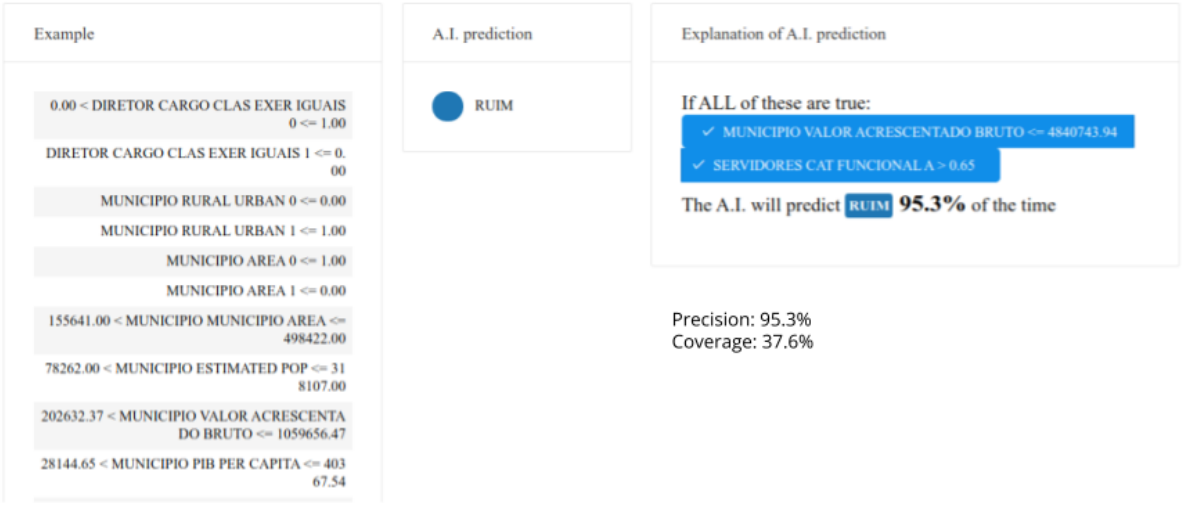
Com escopo direcionado para gerar explicações locais, o método Anchors visa construir poucas regras para justificar determinada decisão do modelo. As explicações geradas também são aproximações locais, porém apresentam melhorias para determinar a vizinhança, em que o formato de regras fica menos suscetível a falhas comparado *surrogates models* lineares. O principal motivo é por capturar relações não-lineares dos dados em seu entorno. As poucas regras geradas ou âncoras, como denomina os autores, são uma forma de criar explicações mais simples, porém durante este processo o algoritmo também dedica-se a maximizar a precisão e a cobertura das regras propostas. A figura 14 apresenta o resultado gerado pela técnica.

O algoritmo Anchors[11] constrói algumas regras que devem ser atendidas para gerar a predição e fornece as informações sobre a precisão e cobertura dessas regras, dado que por ser algo gerado localmente, não necessariamente contemplará a base toda. O resultado é apresentado na figura 14, neste exemplo, duas regras foram estabelecidas para explicar porque a instância analisada teve o desempenho escolar ruim, segundo o modelo.

O método empenha-se em encontrar um conjunto pequeno de regras que são as âncoras para o modelo inferir uma saída. Entretanto não há garantia que um número tão pequeno de clausulas será encontrado, pois algumas regiões podem ser mais complexas e demandar de mais informações. Essa dificuldade também pode ser dada pelo parâmetro do limiar da precisão,

[11] https://github.com/marcotcr/anchor

Figura 14 – Anchors



Fonte: Elaborado pelo autor

o qual o algoritmo busca encontrar regras que satisfaçam essa condição. Assim como outros parâmetros do método podem influenciar, por exemplo o modo de discretizar, a sensibilidade da perturbação, hiper-parâmetros da construção da âncoras, entre outros.

A forma que é proposta a elaboração das explicações por meio de regras, segue uma maneira bem intuitiva por utilizar o conceito de regras, facilmente compreensíveis por pessoas não técnicas. Além disso, objetiva por uma representação esparsa, que foca a análise nas principais informações.

O método apresenta algumas vantagens em comparação a outros da literatura, por limitar bem o espaço ao entorno da vizinhança e apresentar o escopo que as suas explicações contemplam. Essas vantagens são decorrentes da forma não linear para capturar relações complexas, e por fornecer no resultado a cobertura, precisão e indicar para quais instâncias são válidas.

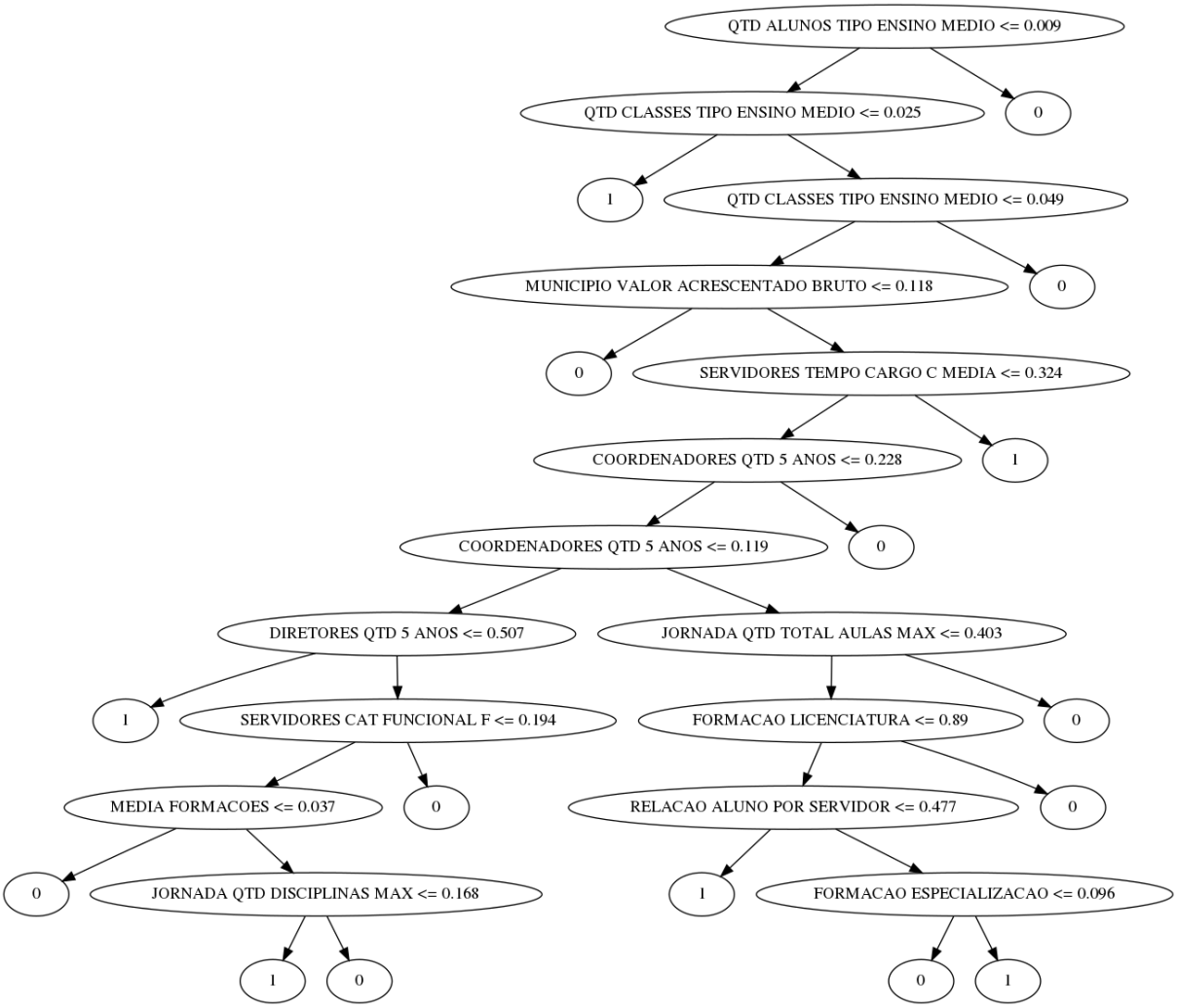
3.2.4.2 Model Extraction

Para obter explicações globais, o método *Model Extraction*[12], também chamado de *Decision Tree Extraction*, produz uma árvore de decisão para simular o comportamento do modelo treinado originalmente com os dados educacionais.

A árvore de decisão construída pelo algoritmo obtém pontuação de 86% da métrica F1 em relação com a saída do modelo original. Essa informação fornecida como um dos resultados

[12] https://github.com/obastani/dtextract

Figura 15 – Model Extraction



Fonte: Elaborado pelo autor

do algoritmo é importante para medir o quanto o *surrogate model* é fiel ao comportamento original.

O autor utiliza estrutura do modelo transparente para interpretar o modelo black-box, portanto as explicações das decisões do modelo são explicadas pelas ramificações e as regras para percorrer o caminho até encontrar o alvo final. No topo da árvore estão os atributos mais importantes, alguns são os mesmos encontrados em outros métodos de interpretabilidade embasados na relevância dos atributos. Ainda que as árvores sejam consideradas interpretáveis, seu tamanho é fundamental para que seja viável analisá-la. A técnica *Model Extraction* foca em controlar a profundida para evitar o aumento da complexidade.

A figura 15 apresenta adaptações na saída do método original para tornar mais fácil a visualização da árvore comparada a impressão dos resultados fornecidas pelo autor. Na implementação de origem, o resultado é exibido em texto, as ramificações são representadas

através da indentação, isto é, dos espaços em branco. Além disso, ao invés de usar os nomes das colunas nos nós, são expostos os índices das posições. Os ajustes realizados para construir a figura foram uma etapa necessária para torná-la mais interpretável. Outros métodos analisados estão mais consolidados na utilização prática, fornecendo interfaces simples para aplicação da técnica e na entrega dos resultados. Nota-se então, que há um emprego muito menor deste método em casos práticos, ainda que seja um trabalho com respaldo na literatura.

O *Model Extraction* faz parte de uma classe de interpretabilidade por meio de *surrogate model* que é bem conhecida e pode ser implementada de diversas formas, inclusive apenas treinando um modelo diretamente com as predições geradas pela versão *black-box*. Embora o *Model Extraction* não esteja amplamente difundido em casos práticos, esse tipo de aplicação costuma ser recorrente. As técnicas desta natureza herdam o grau de interpretabilidade do modelo transparente e tem a facilidade de poder ser comparado através da relação entre as saídas produzidas e a originais.

3.3 DISCUSSÃO DOS RESULTADOS

A análise dos métodos através de aplicações práticas em um mesmo problema facilita na comparação dos resultados, ainda que limitada a um contexto específico, o qual poderia ser expandido em vários problemas simultâneos para fornecer um panorama ainda maior. Nos experimentos realizados, é perceptível a diversidade no formato das interpretações, em que um método de abordagem do mesmo tipo pode calcular e expor os resultados de uma maneira bem diferente. Os formatos apresentados são análogos à forma humana de expressar-se, isto é, diferentes modos de fazer algo. Portanto, definir a melhor abordagem depende do público alvo, dos *insights* ou tipos de explicações desejadas, do escopo da interpretação (global ou local), entre outros fatores que podem influenciar a escolha da solução ideal.

A escolha de uma técnica para interpretar o modelo não necessariamente deve excluir outra, pois as abordagens podem ser combinadas para apresentar diferentes visões do problema ou modelo. Contudo, deve-se ter cuidado para não escolher métodos com mesmo objetivo que entregam resultados diferentes, por exemplo, a relevância dos atributos no mesmo escopo. Ainda que seja possível combinar métodos, o SHAP ganha um destaque na literatura por contemplar explicações locais e globais, mantendo a consistência nos formatos entregues.

As abordagens de explicações locais por meio de exemplo, em especial as contrafactuais, tem bastante destaque e com o passar do tempo surgem novos métodos na literatura alterando

a forma de encontrar as modificações, mas com o mesmo propósito. Esse formato de explicar serve também como uma sugestão de tomada de decisão, a depender do cenário. Comparando com a abordagem global analisada, a forma de interpretar é mais complexa por precisar explorar e ter conhecimento no domínio e técnico para descobrir o que o protótipo representa. Este é um desafio comum de problemas não supervisionados.

Surrogates models globais, como o objetivo do Model Extraction de construir uma árvore de decisão para representar as decisões do modelo *black-box*, podem não conseguir replicar o comportamento de modelos complexos. A tentativa baseada na simplificação pode ficar aquém da performance desejada. Outro cenário provável é a criação de árvores muito grandes para alcançar uma boa taxa de semelhança, entretanto a própria árvore que é um modelo interpretável ficar mais complexa que o desejado. Um outro método que não necessariamente encontrará poucas regras para explicar a decisão, é o Anchors. Neste caso, voltado para o escopo local, o método busca criar regras esparsas para indicar apenas o que é necessário. A saída do algoritmo é uma das mais intuitivas, por ser bem clara e direta. Uma vantagem do método também é o fato de indicar a precisão e cobertura de suas regras, útil para aumentar a confiança e entender o quanto elas generalizam.

A relevância dos atributos é uma forma comum de explicar e é intuitiva. A regressão linear é um exemplo de um modelo clássico transparente que através do coeficiente, de modo similar ao impacto, explica sua decisão. O SHAP encontra o impacto causado por cada atributo em cada instância de forma não linear. Além de ser local e global, o método segue boa fundamentação teórica inspirado na teoria dos jogos, também fornece boas visualizações de dados e bibliotecas de fácil utilização para as principais linguagens de programação da área. Esses são alguns dos motivos que colocam o SHAP como o método mais popular de XAI. Uma das limitações envolve o tempo para interpretar alguns tipos de classificadores. Muitas vezes é preferível escolher o LIME para realizar as explicações locais de SVMs ou Redes Neurais, por exemplo, devido ao custo computacional.

Assim como outros métodos de XAI, o SHAP necessita de esforço dos usuários que consomem a informação para interpretar o que a explicação diz. Por exemplo, em comparação com outros métodos de relevância do atributo, sua exibição dos resultados é bem complexa.

4 UM NOVO MÉTODO BASEADO EM EXPLICAÇÕES TEXTUAIS

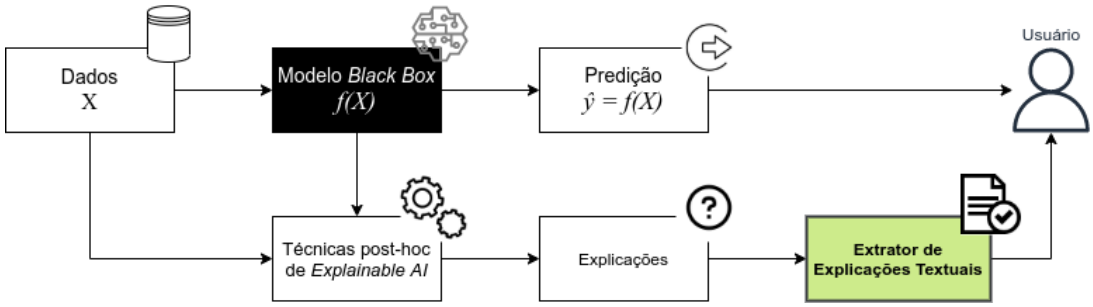
4.1 CAMADA DE EXPLICAÇÕES TEXTUAIS

Ainda que as técnicas de Explainable AI proporcionem maior entendimento a respeito dos mecanismos dos modelos, ainda há um grau elevado na complexidade das explicações para pessoas não técnicas (DHANORKAR et al., 2021). Arrieta (2020) apresenta uma dimensão na análise de XAI que considera o público alvo que consome as explicações fornecidas. Este público é composto por diversos perfis, variando sua capacidade de analisar a complexidade e seu conhecimento no domínio.

De acordo com a definição da área, o objetivo é apresentar conceitos em uma linguagem compreensível para humanos (GUNNING, 2017). Entretanto, com um público alvo variado para interpretar a explicação, um resultado complexo gerado por determinado algoritmo pode ser entendível para um cientista de dados, que está acostumado a lidar com análise de gráficos, dados tabulares, distribuições, entre outros, porém não ser muito intuitivo para um usuário final da aplicação, agente regulatório, executivo de negócios ou outra pessoa interessada.

Para tornar os resultados realmente compreensíveis para humanos, este trabalho propõe uma camada na interpretabilidade voltada para extração das explicações na linguagem natural, sendo apresentado no formato textual. A finalidade é automatizar a interpretação das técnicas de XAI para gerar explicações mais simples de entender e que enderecem as principais informações do algoritmo de explicabilidade utilizado. A figura 16 ilustra o diagrama para gerar explicações textuais. O fluxo proposto generaliza a interpretação post-hoc aplicada, porém, para sua concretização, cada técnica deve ter suas especificidades consideradas. Este trabalho implementa a explicação textual voltada para o algoritmo SHAP, por ser um dos mais importantes da literatura.

Figura 16 – Extrator de Explicações Textuais



Fonte: Elaborado pelo autor

4.2 EXTRAÇÃO DE EXPLICAÇÕES TEXTUAIS DO SHAP

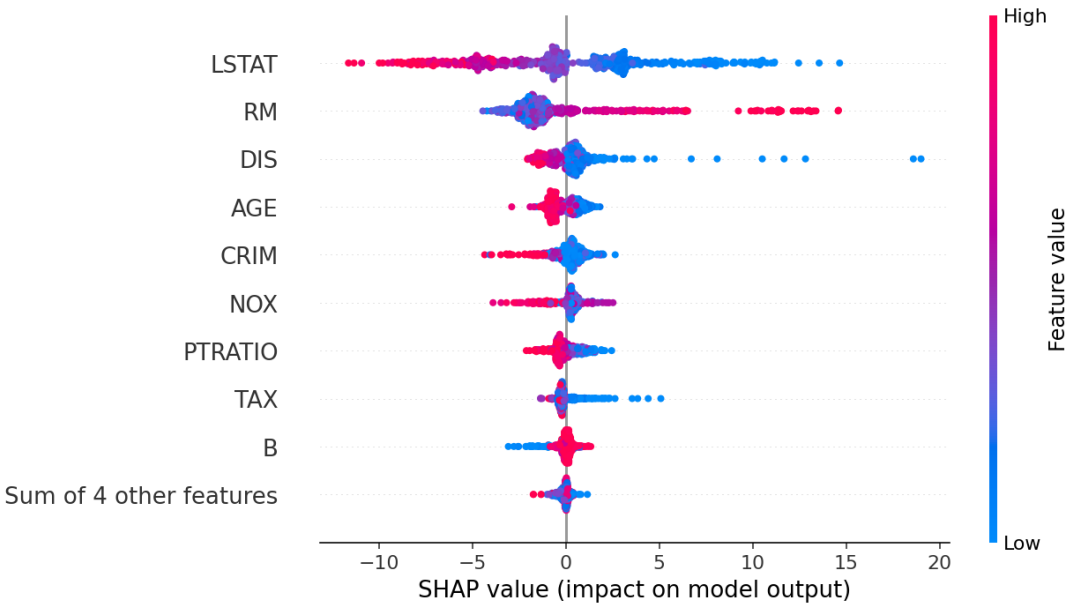
A técnica SHAP fornece a contribuição dos atributos no resultado da predição, a solução desenvolvida por (LUNDBERG; LEE, 2017) apresenta tanto explicações locais como globais, ambas com gráficos para ilustrar o resultado. O gráfico de resumo para interpretação global é composto pela distribuição dos valores dos atributos através da coloração dos pontos, impacto do valor SHAP no eixo X e a ordem de importância (considerando valores absolutos) por meio da ordenação dos atributos no eixo Y. A figura 17 apresenta um exemplo. O método faz parte do estado da arte e é extremamente útil para entender o comportamento de um modelo black-box. Porém devido a complexidade para analisá-lo, a técnica proposta é capaz de extrair as principais informações que a técnica SHAP captura e apresentá-las em forma textual, reduzindo a complexidade para extrair o conhecimento.

A extração das explicações é realizada de acordo com as seguintes etapas:

4.2.1 Criação dos Scores

A primeira etapa consiste em realizar a conversão do valor SHAP original para o conceito de Score. Esta transformação muda a escala do resultado, porém mantém a mesma distribuição e cria uma forma mais simples de interpretar, podendo ser considerada como a contribuição percentual na predição do modelo. Para calcular o score é necessário construir uma matriz de valores absolutos A, a partir da matriz do SHAP, representada por X, como segue a equação 4.1. Em seguida, os valores absolutos de cada instância (ou linha) são somados junto ao valor base do SHAP (b), formando o vetor I, de acordo com a equação 4.2. Por fim, o score é

Figura 17 – SHAP summary



Fonte: Elaborado pelo autor

computado através da equação 4.3, que divide cada valor do SHAP original pelo valor máximo obtido em I.

$$A = |X| \tag{4.1}$$

$$I = \sum_{j=0}^{M} a_{i,j} + b \quad \text{para } i \text{ em } 1, 2, ..., N \tag{4.2}$$

$$S = \frac{1}{\max(I)} X \tag{4.3}$$

onde:

S = matriz de *Scores* extraídos do SHAP

X = matriz com valores SHAP originais

$\max(I)$ = valor máximo do somatório das contribuições dos atributos de cada instância

4.2.2 Relevância do atributo

A importância global de um atributo é definida pela média dos valores absolutos de seus scores, dado que o score já mede uma contribuição percentual. Os valores absolutos são

considerados para que o sinal não crie algum viés na importância geral, podendo, por exemplo, anular os valores quando somados. Então, a importância é medida pela magnitude que impacta de forma geral. O cálculo segue a equação 4.4.

$$F = \frac{1}{N}\sum_{i=0}^{N} a_{i,j} \quad \text{para } j \text{ em } 1, 2, ..., M \tag{4.4}$$

4.2.3 Seleção dos atributos relevantes

Para identificar quais são os atributos relevantes, as importâncias calculadas anteriormente são ordenadas de forma decrescente pela contribuição (equação 4.5). Em seguida, é realizada a soma cumulativa desses valores para representar o ganho na importância a medida que novos atributos são adicionados, de acordo com a equação 4.6. O resultado é transformado em uma matriz T, construída com o índice e o valor acumulado, definido na equação 4.7. Por último, para encontrar o ponto de parada que separa de um lado os atributos mais relevantes para alcançar uma cobertura alta no impacto total e do outro lado os atributos com grau de relevância muito menor, o algoritmo Kneedle (SATOPAA et al., 2011) é aplicado na matriz T. Este algoritmo identifica o momento que adicionar um atributo não traz um benefício tão grande, conhecido por detectar "joelhos" e "cotovelos" em curvas, que representam as curvas côncavas e convexas, respectivamente.

$$\hat{F} = sort(F) \quad \forall i, j \in N \wedge i \leq j \rightarrow \hat{f}_i \geq \hat{f}_j \tag{4.5}$$

$$C = \sum_{i=0}^{j} \hat{f}_i \quad \text{para } j \text{ em } 1, 2, ..., M \tag{4.6}$$

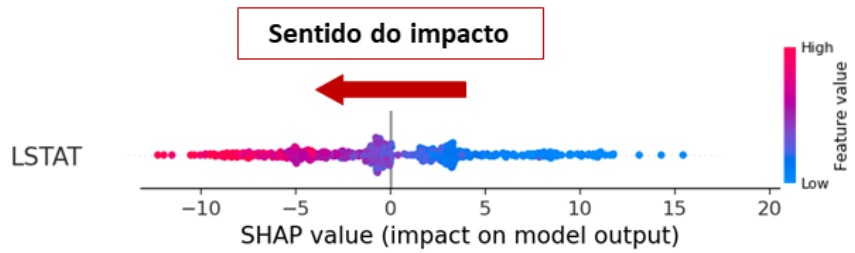
$$T = (j, c_j) \quad \text{para } j \text{ em } 1, 2, ..., M \tag{4.7}$$

4.2.4 Sentido do impacto

A importância global do atributo apresenta um valor absoluto da contribuição, entretanto o valor original SHAP e, consequentemente, o score, são medidas que podem ser negativas. Para extrair a informação sobre qual sentido um atributo tende a impactar na classe alvo, é calculada a correlação de Pearson (PEARSON, 1895) entre os valores do score e os dados do

atributo. As correlações positivas indicam que a medida que os valores do atributo crescem, o impacto na saída do modelo induz para classe positiva, enquanto correlação negativa, indica o oposto. Para facilitar a compreensão é realizada uma discretização do valor da correlação para forma textual, mapeando para relação muito fraca, fraca, moderada, alta e muito alta, de acordo com os limites de 0 a 0.2, de 0.2 a 0.4, de 0.4 a 0.7 e acima de 0.9, respectivamente. A explicação textual do sentido é formada pela extração do sinal e da intensidade da correlação. A figura 18 apresenta como uma informação do sentido pode ser notada em um atributo de acordo com a coloração das instâncias e o eixo X com o valor SHAP.

Figura 18 – Análise do sentido do impacto do SHAP



Fonte: Elaborado pelo autor

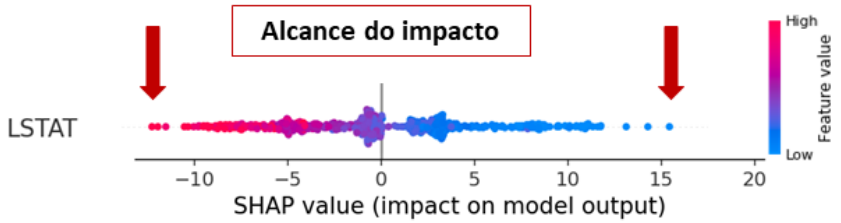
4.2.5 Detalhes por atributo

No detalhamento do impacto causado pelo atributo, além de especificar os valores globais de importância e o sentido, outras informações obtidas do SHAP que também são úteis para aprofundar o entendimento do comportamento são:

4.2.5.1 Alcance do impacto

Indica o valor mínimo e o máximo do impacto causado. Em geral, estes representam a maior contribuição para classe negativa e a maior para classe positiva, respectivamente. A figura 19 exemplifica a informação em uma análise de um atributo individualmente no gráfico do SHAP.

Figura 19 – Análise do alcance do valor SHAP

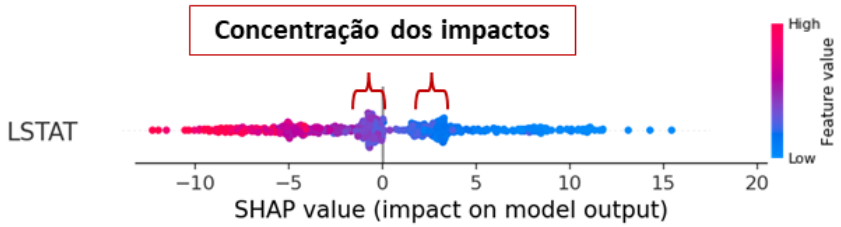


Fonte: Elaborado pelo autor

4.2.5.2 Concentração do impacto

Uma das coisas que chamam atenção no gráfico de resumo do SHAP são as densidades dos impactos em determinadas regiões, formando espécie de bolas, indicando um volume alto. O atributo pode ter um alcance alto no impacto, mas a concentração elevada em valores baixos, ou outras análises podem ser extraídas. Dessa forma, o método proposto também tem como objetivo obter informações a respeito desta distribuição. A figura 20 indica esse tipo de análise que pode ser realizada a partir do gráfico dos valores SHAP de um atributo.

Figura 20 – Análise da concentração do valor SHAP



Fonte: Elaborado pelo autor

Para coletar a informação da concentração, é construído um histograma dos scores do atributo. Com foco em coletar informações para cada uma das classes, é realizado uma separação entre a parte positiva e negativa dos scores. Entretanto, o primeiro passo é identificar o tamanho do intervalo, definido na equação 4.8, que será utilizado para dividir os dados e computar a frequência em cada intervalos. Com o tamanho do intervalo definido, é construído um histograma que separa os scores negativos e os scores positivos. Para explicação textual, é extraído o maior volume em cada uma das partes, indicando percentual de instâncias presentes e o intervalo do impacto/score.

$$h = \left\lceil \frac{\max s - \min s}{k} \right\rceil \tag{4.8}$$

onde:

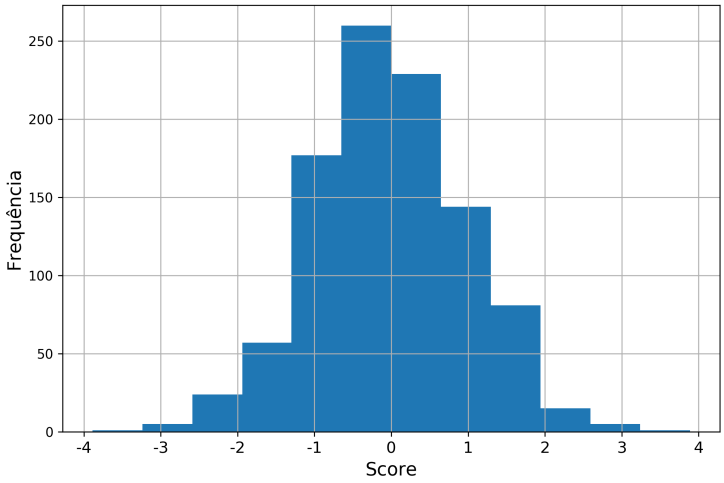
$h =$ tamanho do intervalo

$s =$ são os scores do atributo

$k =$ é a quantidade de intervalos

A figura 21 apresenta o resultado do histograma gerado a partir da separação dos valores negativos e positivos. É possível notar que os tamanhos dos intervalos do histograma são iguais para ambos sentidos dos scores, os positivos e os negativos.

Figura 21 – Exemplo de histograma gerado com a separação dos valores positivos e negativos



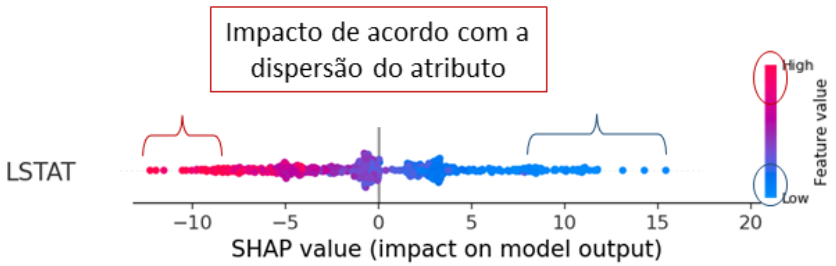
Fonte: Elaborado pelo autor

4.2.5.3 Distribuição dos valores do atributo

Outra informação importante que o gráfico de resumo do SHAP fornece é a distribuição dos dados dos atributos, representadas pela cor dos pontos, indicando os menores e maiores valores. A dispersão dos valores de entrada do atributo em conjunto com o resultado do SHAP permite analisar de forma não-linear o comportamento aprendido pelo modelo. A figura 22 apresenta uma análise que são selecionados os maiores impactos e as faixas que os valores de entrada pertenciam.

Para coletar informações do impacto do atributo de acordo com seu valor, o método proposto utiliza a discretização baseada em quantis para definir intervalos, diferentemente da abordagem anterior que utilizava intervalos fixos. Esta abordagem aproxima-se da forma que as colorações são realizadas no método original e auxilia na interpretação da faixa, visto que

Figura 22 – Análise da dispersão do valor SHAP



Fonte: Elaborado pelo autor

consiste em indicar a posição e não apenas a informação bruta. Após a separação dos atributos em faixas por decis, é calculada a média do score em seu respectivo grupo. Antes de extrair a explicação, é aplicado uma etapa para juntar faixas similares, isto é, as que estão em sequência e apresentam score médio próximos, considerando então comportamentos semelhantes. Por fim, a explicação textual indicará as faixas que representam maior impacto negativo e positivo.

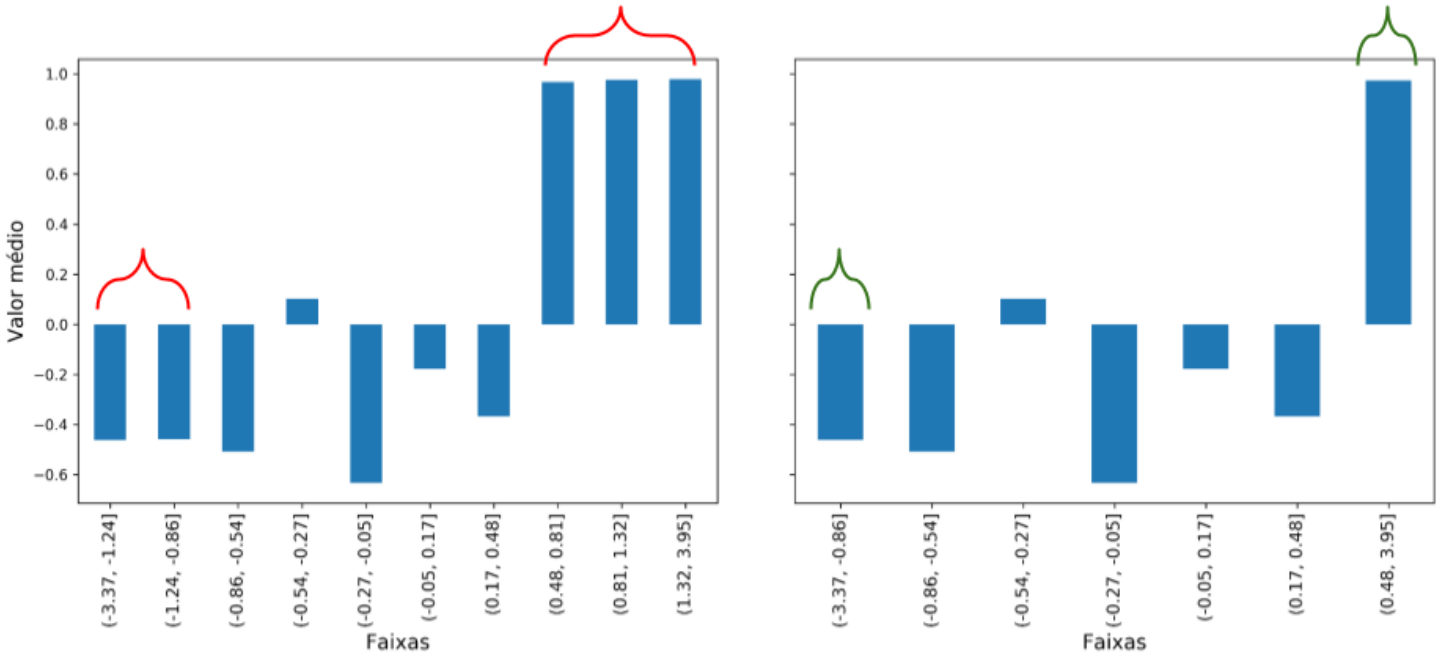
Para exemplificar a junção, a figura 23 apresenta o resultado da metodologia. Neste caso, as duas primeiras faixas e as três últimas são agrupadas formando dois intervalos novos. É possível juntar as faixas devido aos valores médios próximos entre elas e pela forma subsequente que elas encontram-se. Em outras palavras, esses intervalos dos quantis em sequência apresentam mudança no valor de entrado no atributo, mas a contribuição na predição é similar. O eixo X descreve os intervalos, que variam em cada barra, e os intervalos novos são formados pelo mínimo e o máximo das faixas agregadas.

4.2.5.4 Apresentação das explicações

Por fim, todas informações apresentadas anteriormente são geradas no formato textual e unidas posteriormente. Os procedimentos anteriores detalham as etapas necessárias para extrair elementos do SHAP. As principais informações do gráfico com valores SHAP foram resumidas por meio da distribuição, tanto no âmbito geral como também para detalhar o comportamento de atributos específicos. Esse resumo é mapeado para linguagem natural encaixando as informações em moldes textuais pré-estabelecidos com detalhes que dão suporte ao entendimento do comportamento.

A leitura e análise de apenas um atributo conta com diversas informações, então a geração de todos esses detalhes para cada atributo da base de dados pode gerar um trabalho exaustivo para analisar devido ao volume de textos. Com objetivo de permitir uma análise sob demanda,

Figura 23 – Exemplo da junção das faixas similares. Gráfico à esquerda com dados originais e gráfico à direita com resultado após transformação



Fonte: Elaborado pelo autor

foi construída uma aplicação interativa que encapsula o método proposto para explicações textuais e permite que o usuário selecione os detalhes que deseja, além de explorar recursos visuais para auxiliar no entendimento.

As entradas necessárias para construir as explicações são a matriz com valores SHAP, seu valor base (*base value*) e os dados que serão interpretados. A aplicação foi desenvolvida com Streamlit, um *framework* para desenvolvimento de aplicações de dados em Python, linguagem o qual o método proposto foi implementado. As principais bibliotecas que foram utilizados foram Pandas e Numpy, voltadas para manipulação dos dados, Kneedle para detectar os atributos mais relevantes e o pacote padrão do SHAP fornecido pelos autores.

4.3 RESULTADOS

Nesta seção, são apresentados os resultados dos experimentos da aplicação do método proposto, o Textual SHAP, em uma base de dados educacionais. Além do resultado final do método, algumas construções intermediárias são detalhadas para elucidar a lógica para alcançar determinada explicação.

Primeiro, antes de exibir as explicações, é importante visualizar os valores SHAP e o Scores

gerados, apresentado na figura 24a e 24b, respectivamente. A principal mudança está na escala, mantendo os dados com a mesma distribuição, entretanto o Score tem um valor mais amigável de interpretar. A figura também servirá para entender o resultado textual gerado e mapear na representação original.

4.3.1 Interpretação da importância global

Os primeiros atributos mais importantes, baseado na média dos scores absolutos, são selecionadas para compor a explicação. O modelo para predizer o desempenho escolar, considera que os 5 atributos mais importantes com suas respectivas importâncias são: *QTD ALUNOS TIPO ENSINO MEDIO (23,6%), QTD CLASSES TIPO ENSINO MEDIO (10,2%), FORMACAO ENSINO MEDIO (7%), JORNADA QTD DISCIPLINAS MAX (4,2%) e SERVIDORES CAT FUNCIONAL O (3,1%).* A figura 25 apresenta a importância dos 10 primeiros.

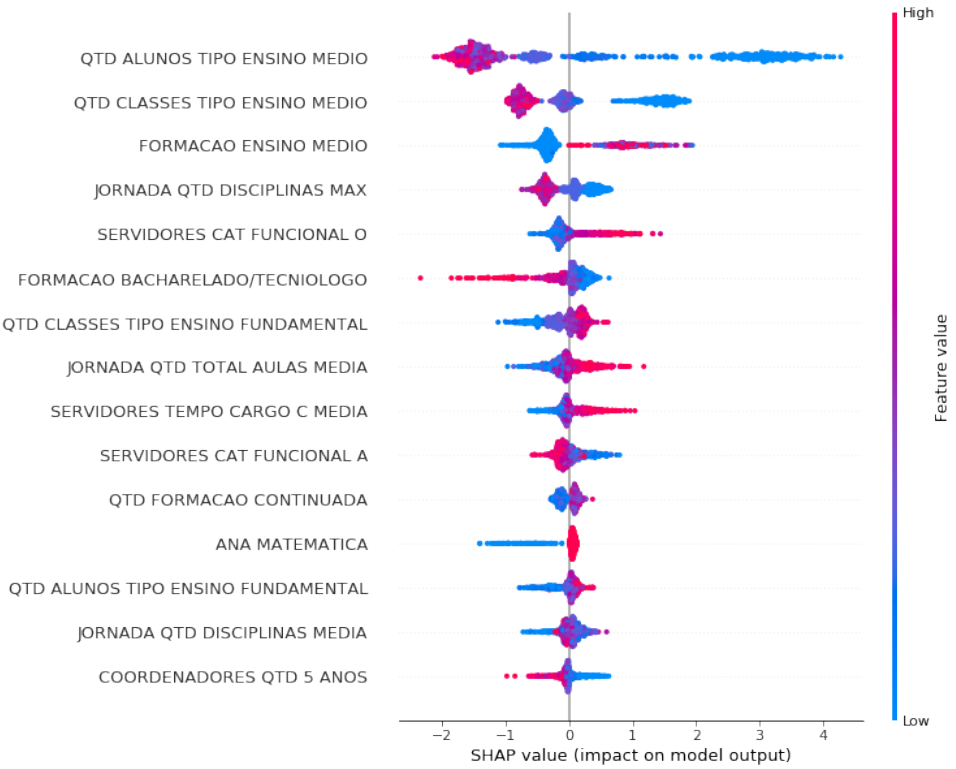
4.3.2 Seleção dos atributos mais relevantes

Muitos modelos de aprendizagem de máquina atribuem peso ou importância muito baixa para algumas variáveis do conjunto de entrada, outra característica comum é concentrar a relevância em poucas características da base. A seleção dos atributos relevantes através do algoritmo Kneedle (SATOPAA et al., 2011) aplicada nas contribuições acumuladas dos atributos ajuda a entender mais sobre o comportamento do modelo. A figura 25 apresenta o ponto de corte detectado, identificando 19 atributos como mais relevantes, o qual após este número o crescimento da importância torna-se mais suave e no final quase nulo. Convertendo para o formato textual, o método constrói a explicação de forma automatizada preenchendo a quantidade de atributos e os percentuais, para o modelo analisado o resultado é: *19 atributos (32,76%), dos 58 da base, apresentam contribuições relevantes, representando 78% da cobertura da contribuição média total. Os outros incrementam muito pouco no resultado final.*

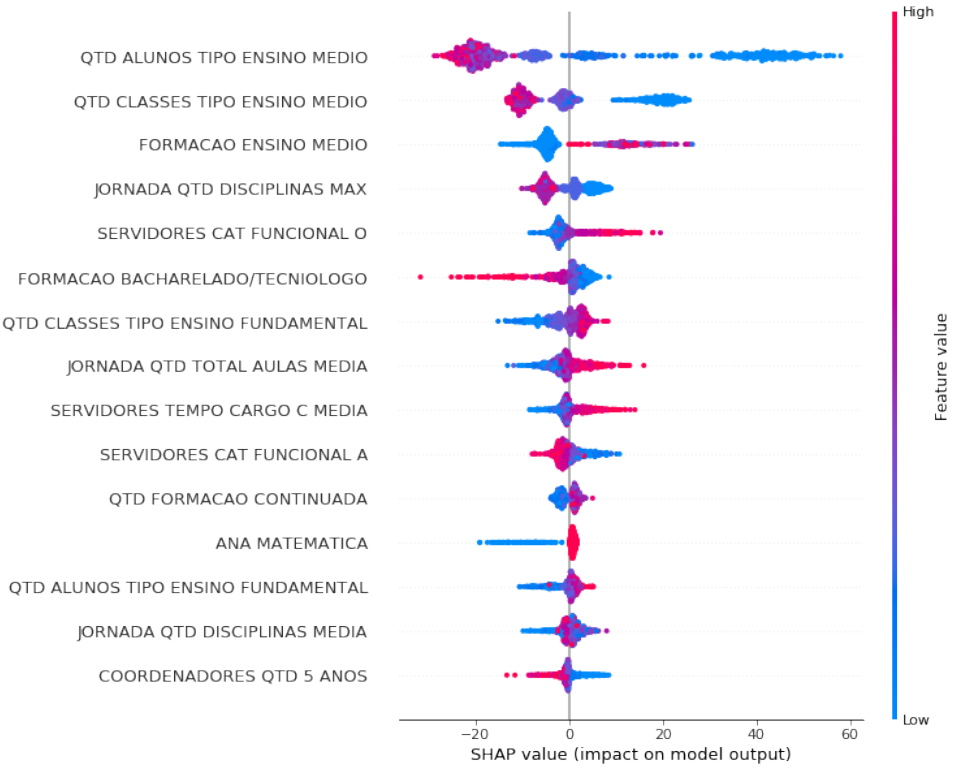
Com objetivo de obter informações dos sentidos que os atributos tendem a impactar, para cada um desses é calculada a correlação de Pearson entre seus valores de entrada e o score obtido. Na figura 27 é apresentado o gráfico de dispersão da variável mais importante do modelo, *QTD ALUNOS TIPO ENSINO MEDIO*, e os scores produzidos para cada instância do conjunto de dados utilizado. O coeficiente de correlação obtido neste exemplo foi $-0,75$. O gráfico conta também com uma regressão linear, que mostra a tendência do impacto ser

Figura 24 – Comparação entre os valores SHAP original e os Scores

(a) Summary plot do SHAP

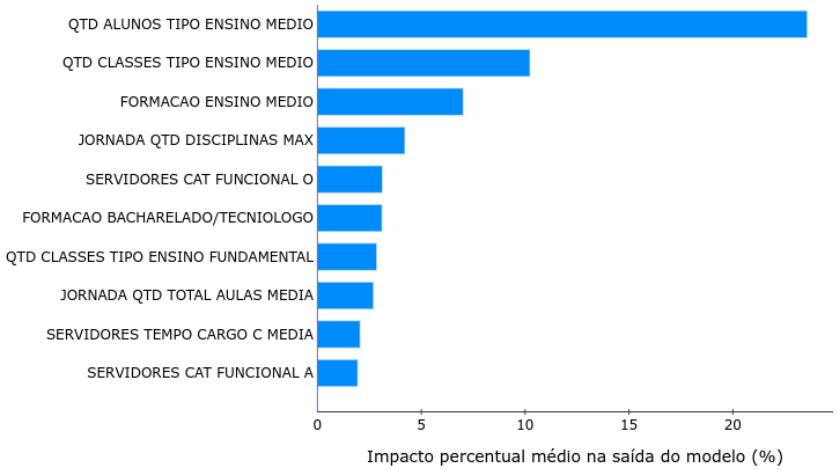


(b) Summary plot do Score



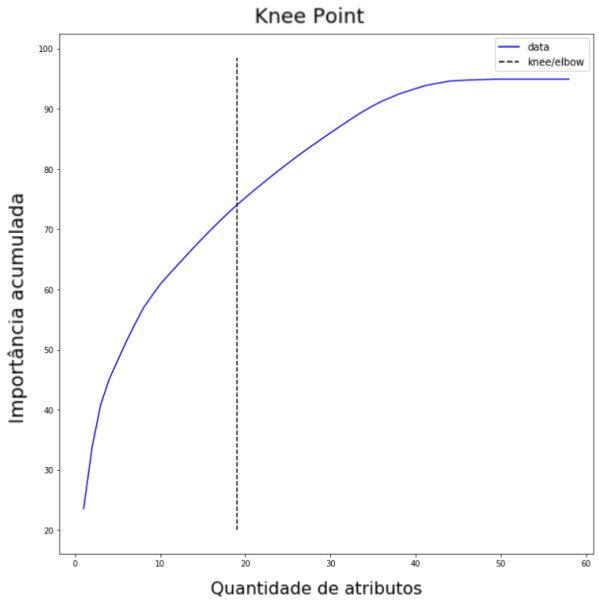
Fonte: Elaborado pelo autor

Figura 25 – Atributos com maior impacto percentual médio na predição



Fonte: Elaborado pelo autor

Figura 26 – Detecção do ponto de corte dos atributos relevantes com algoritmo Kneedle



Fonte: Elaborado pelo autor

negativo, no eixo Y, à medida que os valores da característica aumentam, no eixo X. E com a coloração dos pontos baseado na classe, é possível perceber que há uma propensão para determinada classe a partir de um limiar, ainda que não seja uma relação totalmente linear, o que pode ser notado dado a magnitude moderada da correlação (0,75).

Figura 27 – Dispersão entre os valores do atributo e seu Score

Fonte: Elaborado pelo autor

4.3.3 Sentido dos impactos dos atributos

A partir da identificação do sentido do impacto, os atributos são separados de acordo com a classe de tendência. São apresentados os 5 mais importantes de cada classe, com seu respectivo score médio e intensidade da relação linear.

Classe Positiva (atributos que tem a tendência de impactar de forma positiva na classe alvo a medida que os valores de entrada aumentam):

- *FORMACAO ENSINO MEDIO*: 7.0% de relevância, com relação linear moderada positiva;
- *SERVIDORES CAT FUNCIONAL O*: 3.1% de relevância, com relação linear alta positiva;
- *QTD CLASSES TIPO ENSINO FUNDAMENTAL*: 2.9% de relevância, com relação linear alta positiva;
- *JORNADA QTD TOTAL AULAS MEDIA*: 2.7% de relevância, com relação linear alta positiva;
- *SERVIDORES TEMPO CARGO C MEDIA*: 2.1% de relevância, com relação linear alta positiva;

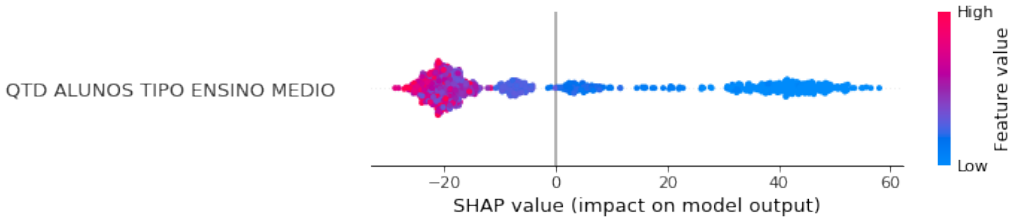
Classe Negativa (atributos que tem a tendência de impactar de forma negativa na classe alvo a medida que os valores de entrada aumentam):

- *QTD ALUNOS TIPO ENSINO MEDIO*: 23.6% de relevância, com relação linear alta negativa;
- *QTD CLASSES TIPO ENSINO MEDIO*: 10.2% de relevância, com relação linear alta negativa;
- *JORNADA QTD DISCIPLINAS MAX*: 4.2% de relevância, com relação linear alta negativa;
- *FORMACAO BACHARELADO/TECNOLOGO*: 3.1% de relevância, com relação linear alta negativa;
- *SERVIDORES CAT FUNCIONAL A*: 1.9% de relevância, com relação linear alta negativa;

4.3.4 Interpretação detalhada do atributo

As explicações anteriores demonstraram um aspecto geral do modelo, suas principais características. Entretanto, é importante aprofundar no comportamento dos atributos individualmente através de uma análise de cada linha do gráfico de resumo do SHAP, neste caso, o gráfico utilizando os scores. Para exemplificar os resultados das análises individuais da metodologia descrito na seção 4.2, o atributo detalhado será o *QTD ALUNOS TIPO ENSINO MEDIO*, selecionado por ser o mais relevante do modelo (23,6%), ilustrado na figura 28.

Figura 28 – Gráfico de resumo do SHAP com Scores do atributo analisado



Fonte: Elaborado pelo autor

4.3.5 Importância e alcance

A primeira explicação é direcionada para indicar a importância no atributo no modelo. A explicação textual é: *O atributo QTD ALUNOS TIPO ENSINO MEDIO é o 1º mais importante*

da base. Entre os 100% de impacto que as informações podem ter, esse atributo contribui com 23.59%.

Depois de indicar o valor absoluto médio da contribuição, o sentido de impacto é declarado através da seguinte explicação automatizada: *À medida que os valores de QTD ALUNOS TIPO ENSINO MEDIO aumentam, seu impacto no modelo tem uma tendência alta de contribuir para um desempenho escolar ruim.*

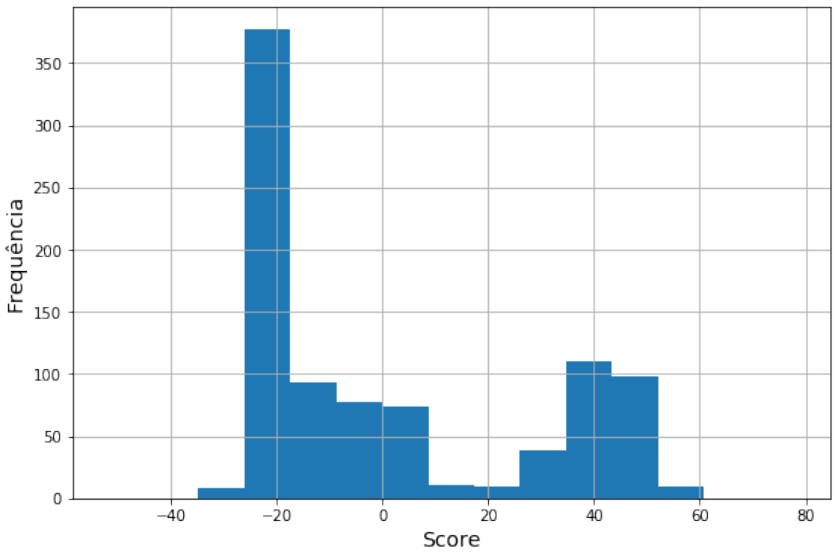
Além de saber a média absoluta do impacto, é importante entender os limites que a contribuição alcança no resultado final. A explicação textual com valor mínimo e máximo é: *O impacto que QTD ALUNOS TIPO ENSINO MEDIO causa para um desempenho escolar ruim é de até 28,8%, e para um desempenho escolar bom chega a 58,04%.*

4.3.6 Concentrações do impacto

No exemplo analisado existe uma divergência grande entre a maior contribuição para o desempenho escolar ruim e bom, 8.07% e 18.98%, respectivamente. Esses valores são um indicativo que as distribuições dos impactos não são simétricas para ambas as classes. Para aumentar o nível de detalhamento da distribuição, calcular a concentração do impacto por faixas dos scores auxiliará a perceber onde há maior densidade de instâncias e qual valor médio naquela região. A figura 29 apresenta o histograma dos scores considerando a separação entre valores positivos e negativos, para evitar enviesar a análise, porém mantendo o tamanho das faixas. Com objetivo de obter explicações do comportamento de forma que simplifique para os usuários finais, são extraídos do histograma gerado os valores máximos para cada classe.

A explicação textual formada pela análise da concentração é a seguinte: *O impacto mais comum gerado por QTD ALUNOS TIPO ENSINO MEDIO para um desempenho escolar bom varia de 34,72% a 43,4%, isso ocorre em 12% dos casos. E para um desempenho escolar ruim varia de 26,04% a 17,36%, isso ocorre em 42% dos casos.* A formação da sentença é dada pelo seguinte padrão, onde os termos maiúsculos são preenchidos pelos valores coletados: *O impacto mais comum gerado por ATRIBUTO para CLASSE varia de VALOR MÍNIMO DA FAIXA% a VALOR MÁXIMO DA FAIXA%, isso ocorre em VOLUME% dos casos.*

Figura 29 – Histograma dos scores do atributo analisado



Fonte: Elaborado pelo autor

4.3.7 Distribuição dos atributos

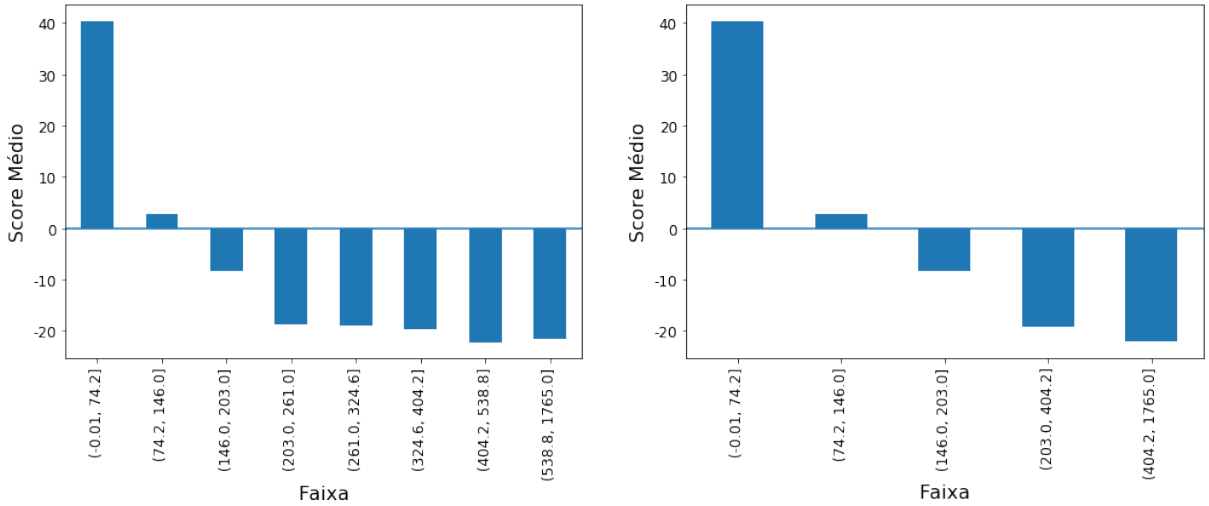
Para explicar quais valores do atributo impactam mais no modelo, de forma que possibilite uma comparação entre os valores exatos e sua relação com de variabilidade da amostra, é utilizada a discretização baseada nos quantis. A separação é realizada em 10 faixas, equivalente aos decis da amostra. Os quantis permitem mencionar a posição dos dados, o que seria algo similar as cores no gráfico, o qual os primeiros quantis fazem analogia aos pontos azuis, já os quantis mais altos são referentes aos pontos vermelhos. No final, essas posições são convertidas para informações discretas que simplificam o entendimento da magnitude dos dados.

A figura 30 apresenta os valores médio por cada faixa. À esquerda, as faixas iniciais geradas através dos quantis, enquanto à direita, o resultado após a junção das faixas similares. Nota-se que a partir da segunda faixa da primeira segmentação, os valores apresentam médias similares mesmo com aumento do atributo, o que também pode ser notado através da figura 27. Devido à similaridade, estas são unidas para formar um novo intervalo, conforme apresentado na figura à direita. Para gerar a explicação das faixas que mais impactam em cada classe, são considerados as maiores médias para o sentido positivo e negativo. Ainda que neste caso só exista uma de cada tipo, não é sempre que esse comportamento ocorre.

Sobre a similaridade, vale ressaltar que sem a agregação, a escolha do maior valor conduziria para um intervalo que cobriria um percentual de 10% dos dados, devido ao decil utilizado, porém reduziria o espaço que demonstra que o mesmo impacto seria causado e poderia levar

a uma interpretação errada.

Figura 30 – Score médio por faixa dos quantis. À esquerda, discretização original, à direita intervalos após união por similaridade



Fonte: Elaborado pelo autor

A última conversão na faixa selecionada ocorre para torná-la uma representação nominal, em outras palavras, as posições são transformadas para uma forma de contextualizar o resultado. Por exemplo, quando a maior média do impacto ocorre entre os percentis de 80% e 100%, uma simplificação é afirmar que ocorreu quando os valores do atributo estavam muito altos. As conversões são realizadas conforme a 4.9, onde a e b, representam o percentual mínimo e o máximo do intervalo, respectivamente. Posteriormente a faixa nominal encontrada é inserida no texto.

$$faixa = \begin{cases} \text{muito altos,} & \text{para } a \geq 70 \wedge b \geq 90 \\ \text{altos,} & \text{para } a \geq 70 \\ \text{acima da média,} & \text{para } a \geq 50 \wedge b \geq 70 \\ \text{um pouco acima da média,} & \text{para } a \geq 50 \wedge b < 70 \\ \text{medianos,} & \text{para } a \geq 40 \wedge b \geq 60 \\ \text{um pouco abaixo da média,} & \text{para } a \geq 30 \wedge b \leq 50 \\ \text{abaixo da média,} & \text{para } a \leq 30 \wedge b \leq 50 \\ \text{baixos,} & \text{para } b \leq 30 \\ \text{muito baixos,} & \text{para } a \leq 10 \wedge b \leq 30 \end{cases} \tag{4.9}$$

A sentença para gerar a explicação é formada pela substituição das descrições maiúsculas pelos valores coletados. O molde é dado por: *Os maiores impactos para CLASSE, ocorreram quando os valores de ATRIBUTO eram REPRESENTAÇÃO NOMINAL DA FAIXA (valores entre VALOR MÍNIMO DA FAIXA e VALOR MÁXIMO DA FAIXA), sua influência média no desempenho foi IMPACTO MÉDIO%.*

Aplicando no exemplo, é gerado o seguinte resultado:

Os maiores impactos para o desempenho ruim, ocorreram quando os valores de QTD ALUNOS TIPO ENSINO MEDIO eram muito altos (valores entre 404 e 1765), sua influência média no desempenho foi 21.92%

Os maiores impactos para o desempenho bom, ocorreram quando os valores de QTD ALUNOS TIPO ENSINO MEDIO eram muito baixos (valores entre 0 e 74), sua influência média no desempenho foi 40.41%

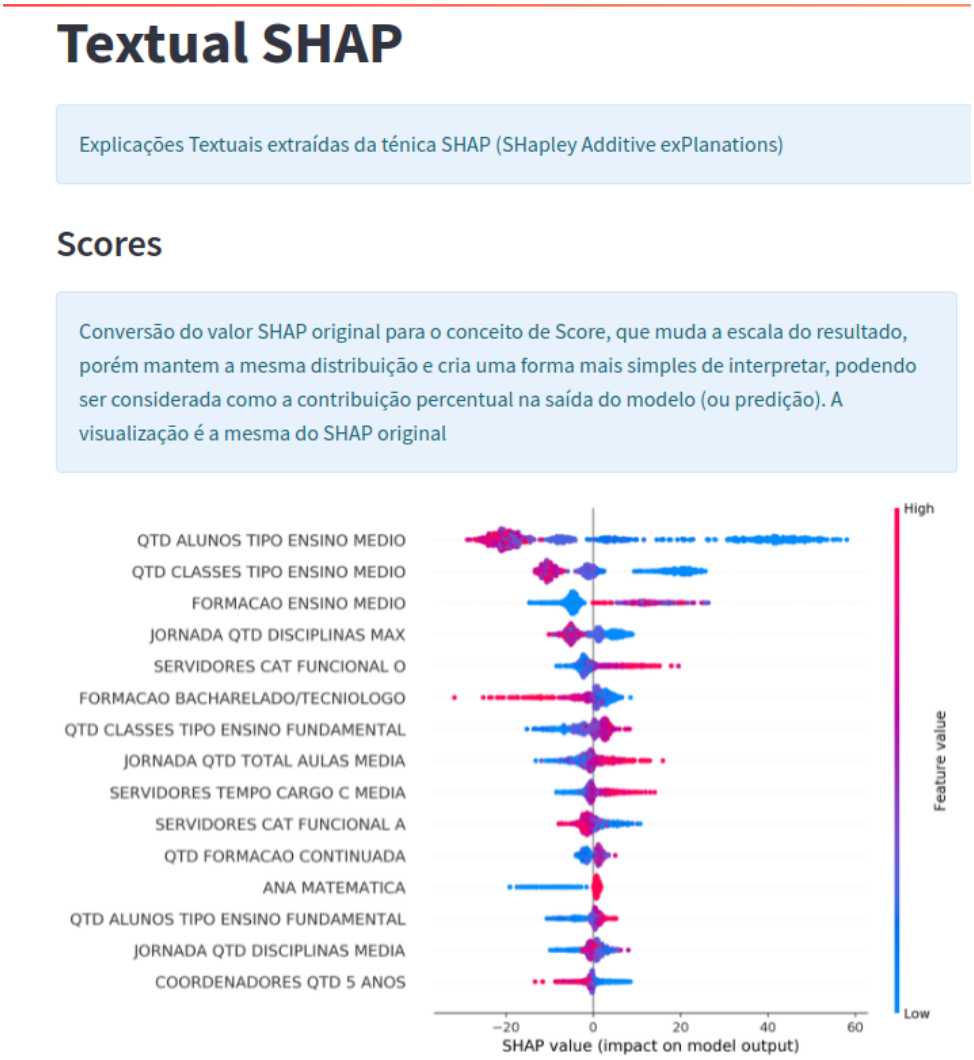
4.4 APLICAÇÃO INTERATIVA COM RESULTADOS CONSOLIDADOS

As etapas anteriores detalham o resultado dos principais passos para extração das explicações a partir do SHAP. Por último, é necessário expressar em conjunto no formato textual para o usuário final. Os resultados do método textual para interpretar o modelo de aprendizagem de máquina para o contexto educacional são apresentados por meio de imagens obtidas da aplicação desenvolvida. A ferramenta apresenta o gráfico e as explicações, além disso clarifica o significado dos impactos por meio do score, conforme a figura 31 e a figura 32, respectivamente.

Em um dos estágios do método proposto é computado a quantidade de variáveis com impacto relevante. A ferramenta apresenta um componente para exibir os nomes desses atributos, de modo que possibilita escolher quando quer ver e quando deseja manter ocultado, como segue a figura 33. Em bases com muitos atributos a lista será grande e saber todos os detalhes pode não ser desejado, a utilização desse componente é útil para usabilidade.

Para concluir, a ferramenta conta com um componente de seleção de atributo para construir as explicações deste, detalhando o comportamento sob demanda, exemplo exibido na figura 34. Do modo que são extraídas várias leituras do SHAP, com objetivo de indicar a importância, sentido, variação, densidade e dispersão, exibir nesse formato textual para todos os atributos vai resultar em um texto enorme, principalmente em bases de dados com muitos atributos. Em função disso, a seleção evita esse comportamento e foco em detalhar o que é desejado.

Figura 31 – Aplicação



Fonte: Elaborado pelo autor

As primeiras leituras podem ser mais demoradas, no entanto com o decorrer da análise de novos atributos, a leitura passa a ser mais rápida, dado que as explicações seguem moldes que definem sua estrutura, então as análises em sequência tendem a ser mais diretas, localizando rapidamente os valores e entendendo seu significado.

Informações gerais:

Principais atributos (top 5):

(*Relevância média do atributo*)

- QTD ALUNOS TIPO ENSINO MEDIO (23.6%)
- QTD CLASSES TIPO ENSINO MEDIO (10.2%)
- FORMACAO ENSINO MEDIO (7.0%)
- JORNADA QTD DISCIPLINAS MAX (4.2%)
- SERVIDORES CAT FUNCIONAL O (3.1%)

Principais atributos por classe

Classe Positiva:

(atributos que tem a tendência de impactar de forma positiva na classe alvo a medida que os valores do atributo aumentam)

- FORMACAO ENSINO MEDIO: 7.0% de relevância, com relação linear moderada positiva
- SERVIDORES CAT FUNCIONAL O: 3.1% de relevância, com relação linear alta positiva
- QTD CLASSES TIPO ENSINO FUNDAMENTAL: 2.9% de relevância, com relação linear alta positiva
- JORNADA QTD TOTAL AULAS MEDIA: 2.7% de relevância, com relação linear alta positiva
- SERVIDORES TEMPO CARGO C MEDIA: 2.1% de relevância, com relação linear alta positiva

Classe Negativa:

(atributos que tem a tendência de impactar de forma negativa na classe alvo a medida que os valores do atributo aumentam)

- QTD ALUNOS TIPO ENSINO MEDIO: 23.6% de relevância, com relação linear alta negativa
- QTD CLASSES TIPO ENSINO MEDIO: 10.2% de relevância, com relação linear alta negativa
- JORNADA QTD DISCIPLINAS MAX: 4.2% de relevância, com relação linear alta negativa
- FORMACAO BACHARELADO/TECNIOLOGO: 3.1% de relevância, com relação linear alta negativa
- SERVIDORES CAT FUNCIONAL A: 1.9% de relevância, com relação linear alta negativa

Figura 32 – Aplicação
Fonte: Elaborado pelo autor

Figura 33 – Aplicação

Cobertura dos principais atributos

19 atributos (32.76%), dos 58 da base, apresentam contribuições relevantes, representando 77.97% da cobertura da contribuição média total. Os outros incrementam muito pouco no resultado final

Mostrar 19 atributos mais relevantes +

Mostrar 39 atributos mais irrelevantes −

- DIRETOR ANOS TRAB CARGO E
- FORMACAO DOUTORADO
- MUNICIPIO VALOR ACRESCENTADO BRUTO
- MEDIA ALUNOS SALA
- QTD TOTAL ALUNOS

Fonte: Elaborado pelo autor

Detalhes por atributo:

Selecionar atributo

QTD ALUNOS TIPO ENSINO MEDIO

QTD ALUNOS TIPO ENSINO MEDIO

Importância:

O atributo QTD ALUNOS TIPO ENSINO MEDIO é o 1º mais importante da base. Entre os 100% de impacto que as informações podem ter, esse atributo contribui com **23.59%**

Sentido do impacto:

A medida que os valores de QTD ALUNOS TIPO ENSINO MEDIO aumentam, seu impacto no modelo tem uma tendência alta de contribuir para um desempenho escolar ruim

Variação:

O impacto que QTD ALUNOS TIPO ENSINO MEDIO causa para um desempenho escolar ruim é de até 28.8%, e para um desempenho desempenho escolar bom chega a 58.04%

Impacto mais comuns:

O impacto mais comum gerado por QTD ALUNOS TIPO ENSINO MEDIO para um desempenho escolar bom varia de 34.72% a 43.4%, isso ocorre em 12% dos casos

O impacto mais comum gerado por QTD ALUNOS TIPO ENSINO MEDIO para um desempenho escolar ruim varia de 26.04% a 17.36%, isso ocorre em 42% dos casos

Quando ocorrem os maiores impactos:

Os maiores impactos para o desempenho ruim, ocorreram quando os valores de QTD ALUNOS TIPO ENSINO MEDIO eram muito altos (valores entre 404 e 1765), sua influência média no desempenho foi 21.92%

Os maiores impactos para o desempenho bom, ocorreram quando os valores de QTD ALUNOS TIPO ENSINO MEDIO eram muito baixos (valores entre 0 e 74), sua influência média no desempenho foi 40.41%

Fonte: Elaborado pelo autor

4.5 AVALIAÇÃO

Uma das principais carências da área de Explainable AI é percepção no humano no centro fundamental para o desenvolvimento e o avanço das interpretabilidades (SCHOONDERWOERD et al., 2021; EHSAN; RIEDL, 2020). Alguns métodos propostos apresentam um formalismo matemático, mas falta a validação da compreensibilidade da proposta, visto que esse é o objetivo principal da área. Por essa razão, esse trabalho propõe uma metodologia para extrair explicações em um formato textual para torná-las mais simples para interlocutores não técnicos e aplica uma avaliação da percepção dos usuários. Nesta seção são apresentados o processo de coleta de avaliações com usuários reais e os resultados encontrados.

A pesquisa com usuários segue uma abordagem de XAI centrada nos humanos inspirada em alguns trabalhos da literatura (LU et al., 2019; LIAO; GRUEN; MILLER, 2020). O objetivo principal foi validar a hipótese que a interpretação fornecida pela técnica SHAP apresentava uma alta complexidade para pessoas não técnicas, mesmo possuindo conhecimento no domínio. Por outro lado, as explicações em um formato de linguagem natural seriam mais simples e mais bem compreendidas.

4.5.1 Processo de avaliação

A avaliação realizada neste trabalho foi por meio de um questionário desenvolvidos pelo *Google Forms*[1] e enviado para o público da área de educação, visto que se tratava do modelo de predição de desempenho escolar. O formulário tem a seguinte estrutura:

1. Apresentação do escopo da pesquisa
2. Contextualização sobre a predição do desempenho escolar
3. Perguntas para identificar o público da pesquisa
    a) Faixa etária
    b) Nível de escolaridade
    c) Familiarização com Inteligência Artificial
    d) Área de atuação

[1] https://www.google.com/intl/pt-BR/forms/about/

4. Apresentação do dicionário dos dados
5. Apresentação do SHAP
6. Apresentação das Explicações Textuais
7. Perguntas sobre as formas de interpretar
    a) Qual explicação ajudou a entender melhor o modelo de IA?
    b) Qual a complexidade de cada uma das técnicas?
    c) Qual foi o grau de entendimento de cada uma das técnicas?
    d) Espaço aberto para eventuais comentários

Na pesquisa foram exibidos o gráfico do SHAP e as Explicações Textuais obtidas com método proposto, Textual SHAP. Para ambos, foram inseridas informações básicas de como realizar a leitura dos resultados fornecidos pelas técnicas. Após apresentar ambas as técnicas, foram realizadas as perguntas sobre qual ajudou mais a entender o modelo (gráfico, texto, ambas ou nenhuma), e por meio da escala Likert, perguntas voltadas para a complexidade e o entendimento. O formulário completo está disponível no apêndice A.

4.5.2 Resultados da avaliação

A pesquisa foi elaborada com 30 participantes distintos que responderam o questionário online. Os participantes foram encontrados por meio de redes sociais em grupos no contexto educacional. Esta seção apresenta os resultados da pesquisa. Os gráficos com detalhes de todas as respostas estão no apêndice B.

Algumas perguntas do formulário são voltadas para validar o público alvo. A primeira análise apresenta a área de atuação, o qual 93,3% responderam serem de Educação. Dos 30, apenas duas indicaram atuar em outra área, que foram Saúde e Finanças. Dado que a área de educação pode envolver assuntos multidisciplinares, o que não exclui essas pessoas de algum contato com a área. Por representar um número pequeno também, não há motivos para descartar.

A tabela 3 apresenta as faixas etárias dos participantes da pesquisa e a tabela 4 as respostas para o nível de escolaridade. Ambas as informações serviram como um *proxy* ou representante para a experiência na área, visto que uma idade pode ter correlação com tempo de experiência

profissional e o nível de estudo também é um fator determinante para o conhecimento no domínio. Mais de 80% dos participantes têm idade superior a 25 anos, 70% têm pós-graduação completa. De todos os níveis de escolaridade, os mais baixos encontrado nas respostas foram referentes ao ensino superior incompleto, com 13,3%, que, de acordo com os meios de divulgação, representam os estudantes de graduação na área de educação.

Tabela 3 – Faixa etária dos participantes da pesquisa

Faixa etária	Quantidade	Percentual
Abaixo de 18	0	0%
18-25	5	16,67%
26-35	3	10%
36-46	13	43,33%
46-55	8	26,67%
Acima de 55	1	3,33%

Fonte: Elaborada pelo autor

Tabela 4 – Escolaridade dos participantes da pesquisa

Escolaridade	Quantidade	Percentual
Ensino Médio - Incompleto	0	0%
Ensino Médio - Completo	0	0%
Superior - Incompleto	4	13,33%
Superior - Completo	3	10%
Pós-graduação - Incompleto	2	6,67%
Pós-graduação - Completo	21	70%

Fonte: Elaborada pelo autor

Bem como o conhecimento no domínio, a ausência do entendimento técnico em IA é um ponto fundamental na validação dos consumidores finais das explicações. Uma das perguntas foi elaborada seguindo o intuito de representar a escala Likert no grau de compreensão, mapeando os valores para conceitos que ajudam a responder. Dessa forma, não houve um participante com conhecimento elevado no tema, considerando as respostas "Costuma utilizar para tomar decisões" e "Desenvolve soluções de IA", equivalente aos valores 4 e 5 na escala Likert, respectivamente. Portanto, as informações coletadas validam o público da pesquisa para as hipóteses levantadas.

A primeira pergunta sobre a avaliação das técnicas, "*Qual das explicações ajudou a entender melhor o resultado do modelo de Inteligência Artificial?*", foi empreendida para extrair

de forma direta a preferência em relação ao esclarecimento fornecido pelas técnicas. O resultado detalhado apresentado na tabela 5. A principal escolha foi para as explicações textuais, (46,67%), em segundo lugar a opção que ambas serviram da mesma forma (33,33%). Apenas 20% escolheram a opção do gráfico, e ninguém informou que nenhuma explicação ajudou de alguma forma.

Tabela 5 – Resultado da pergunta sobre a escolha da explicação que mais ajudou a entender o modelo de IA

Opção	Quantidade	Percentual
Opção 1: Gráfico	6	20%
Opção 2: Explicação Textual	**14**	**46,67%**
Ambas serviram da mesma forma	10	33,33%
Nenhuma ajudou	0	0%

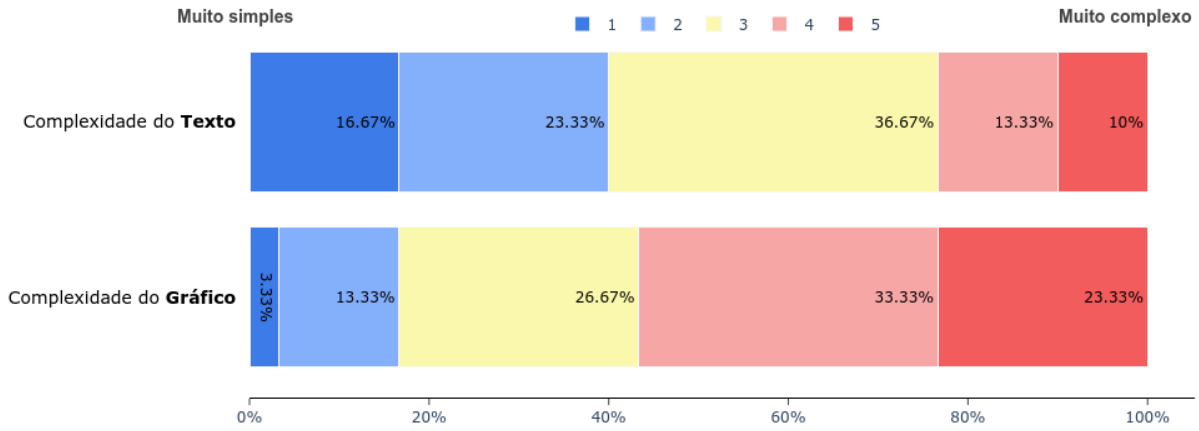
Fonte: Elaborada pelo autor

A maioria das respostas (80%) contam com a explicação textual de alguma forma, seja individualmente (46,67%) ou com ambas abordagens (33,33%). Essa quantidade expressiva serve para mostrar que é possível transmitir o conteúdo do gráfico em um formato textual. O método proposto não determina a exclusão do gráfico precisamente, mas sim sugere um formato que pode ser uma alternativa ou complemento para facilitar a interpretação. Então, faz sentido as explicações complementarem-se e juntas apresentar uma versão melhor. Contudo, a pergunta tem caráter comparativo e não sobre aumentar a qualidade quando utilizadas simultaneamente. As perguntas na sequência sobre cada uma das técnicas fornecem mais detalhes sobre a percepção das técnicas.

Na análise individual das técnicas, a primeira pergunta foi relacionada a complexidade. A figura 35 apresenta a distribuição dos resultados obtidos em cada uma das técnicas por meio do gráfico de barras empilhadas na horizontal. A escala vai de "muito simples" até "muito complexo", as extremidades apresentam uma saturação maior nas cores, que variam entre azuis para os resultados positivos, vermelho para parte negativa e a cor amarela para indicar resultados medianos. Quanto menor o valor, melhor o resultado, visto que quanto menos complexo, melhor para o usuário.

A partir da figura 35 é possível notar que nas explicações textuais há uma concentração maior na faixa mais simples, representada pelos valores 1 e 2, totalizando 40%. Enquanto por meio do gráfico do SHAP, nessas mesmas faixas o percentual de respostas foi 16,67%. Para verificar se realmente há uma diferença estatisticamente significante, foi realizado o teste de Wilcoxon, um teste estatístico não paramétrico para amostras dependentes, visto que o

Figura 35 – Avaliação da complexidade das explicações



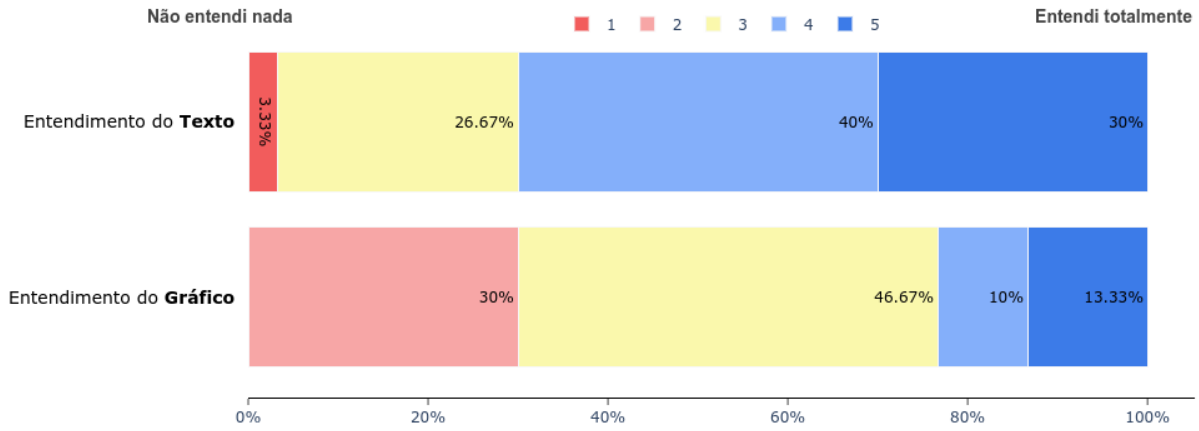
Fonte: Elaborado pelo autor

objetivo é realizar uma comparação pareada de dados qualitativos ordinais. A hipótese nula é que as medianas das amostras são iguais, enquanto hipótese alternativa é que a mediana da explicação textual é menor que a mediana do gráfico, ou seja, um teste unilateral. Com nível de significância de 0,05, o teste rejeita a hipótese nula com p-valor de 0,002, sendo forte evidência para aceitar a hipótese alternativa. Por meio do teste é plausível afirmar que as explicações são menos complexas que o gráfico fornecido pelo SHAP.

A segunda análise individual é voltada para o grau de entendimento das explicações. O resultado das distribuições das respostas é apresentado na figura 36. O gráfico é construído de forma similar ao da complexidade (figura 35), exceto pelo sentido das respostas. Neste caso, a escala varia entre a falta de compreensão das explicações até o entendimento total, consequentemente, valores mais altos representam resultados melhores, por este motivo a inversão das cores.

Analisando a figura 36, nota-se uma grande diferença entre a compreensão das explicações, o qual o formato textual obteve uma nota maior em comparação com a interpretação do gráfico. Enquanto com o gráfico do SHAP 23,33% entenderem bem ou totalmente, por meio de textos esse entendimento foi bem maior, com 70%. De forma similar ao exemplo anterior, foi realizado o teste estatístico de Wilcoxon pareado das duas amostras com as respostas sobre o grau de entendimento para comparar se a mediana da explicação textual é menor que com

Figura 36 – Avaliação do entendimento das explicações



Fonte: Elaborado pelo autor

o gráfico. O resultado do teste unilateral foi o p-valor de 0,001, rejeitando a hipótese nula (medianas iguais), com nível de significância de 0,05%, e aceitando a hipótese alternativa que a explicação textual foi mais compreendida que o gráfico.

5 CONSIDERAÇÕES FINAIS

A necessidade de *Explainable AI* aumentou nos últimos anos com objetivo de atender princípios de responsabilidade nas soluções de IA tais como justiça, privacidade, transparência, prestação de contas, segurança e ética. As técnicas desenvolvidas até a atualidade para interpretação dos modelos de aprendizagem de máquina, proporcionaram um aumento na confiança e robustez, principalmente em modelos black-box, principal foco da área. Apesar dos avanços, ainda existem lacunas para serem exploradas neste campo de pesquisa relativamente novo, entre elas está a qualidade da explicação do ponto de vista centrado nos humanos.

Este trabalho realizou uma ampla investigação do estágio atual de Explainable AI, através da análise da literatura e comparação prática entre os principais métodos existentes. Com base nos pontos fracos da área, de uma perspectiva centrada nos humanos, foi proposto uma técnica para gerar explicações textuais a partir de um dos métodos do estado-da-arte.

5.1 CONTRIBUIÇÕES

Após o levantamento das principais técnicas, foi proposta uma organização de acordo com as principais taxonomias da área e os tipo da abordagem para construção das explicações. Seguindo esta organização, foram realizados experimentos para comparar os métodos. Esta comparação é uma das principais contribuições do trabalho. O estudo foi direcionado para interpretação de modelos agnósticos de forma post-hoc para classificação binária com dados tabulares. Diferente de outros trabalhos da literatura, são apresentados resultados práticos em uma mesma base de dados e modelo treinado previamente, de forma a facilitar a comparação das explicações geradas. A seleção dos métodos contempla o escopo local e global para cada possibilidade de abordagem da explicação, baseado na organização construída neste trabalho, que difere dos trabalhos da literatura, nos quais os métodos criam suas explicações por visualização, relevância dos atributos, baseado em instâncias ou extração de regras.

Os resultados da comparação mostram a diversidade nas formas de explicar, incluindo a possibilidade de combinar técnicas para alcançar diferentes pontos de vista a partir de determinado modelo. É ressaltada também a vantagem do algoritmo SHAP frente a seus concorrentes, por apresentar explicações locais e globais, além de conter boas visualizações e boas propriedades teóricas para indicar a relevância dos atributos.

Apesar do SHAP apresentar muitas vantagens e ser uma referência na área, o método apresenta uma alta complexidade para ser analisado, principalmente da ótica de pessoas sem conhecimento em ciência de dados e inteligência artificial. Esse não é um aspecto exclusivo do SHAP, mas sim uma carência de traduzir para usuários leigos que área enfrenta. Como resultado da análise, foi levantada a necessidade de adaptar a linguagem utilizada. Neste trabalho foi proposto o método Textual SHAP para analisar os valores SHAP capturar as principais informações do gráfico para apresentar em um formato textual.

O método proposto foi analisado por 30 participantes em uma pesquisa conduzida com pessoas sem conhecimento avançado em inteligência artificial e com conhecimento no domínio educacional, dado que o modelo analisado era para predição do desempenho escolar. A pesquisa que permitiu concluir que há vantagens no formato textual de explicar em comparação com o gráfico das explicações globais do SHAP para um público não técnico. Além de maior preferência, os resultados indicaram que a forma proposta conseguiu ser mais bem compreendidas e foram menos complexas que o gráfico da técnica original.

5.2 LIMITAÇÕES

O método proposto demonstrou ser menos complexo que o gráfico original do SHAP, contudo alguns participantes da pesquisa responderam achar complexo o que foi apresentado. Então, a técnica deve ser aprimorada, simplificar os termos utilizados e tornar mais fácil de capturar a essência dos resultados. Outro ponto relacionado está no tamanho dos textos, devido a especificidade de cada atributo, a explicação fica ainda maior.

Na questão técnica da geração das explicações também existem limitações a respeito dos cálculos para extrair informações. As principais questões são a forma de capturar a tendência linear por meio do coeficiente de Pearson, que pode falhar em alguns casos e o número não condizer exatamente com a tendência. Outras informações sobre a tendência podem ser úteis também, como um crescimento exponencial, saturação a partir de determinado valor e os casos não-lineares com comportamento bem definido, por exemplo impacto alto nas extremidades do atributo e baixo nos valores medianos. Uma outra simplificação está na captura das faixas, principalmente na concentração, o qual são utilizados uma quantidade fixa de faixas.

5.3 TRABALHOS FUTUROS

As explicações textuais demonstraram um grande potencial para os usuários leigos, entretanto o método desenvolvido apresenta algumas limitações que podem ser melhor investigadas e aprimoradas em trabalhos futuros. Bem como outras possibilidades surgem a partir da técnica proposta que não se encerram nas limitações descritas, tal como a generalização para outros métodos de relevância de atributos, experimentação em novas bases, adaptações para funcionar com múltiplas classes e em casos de regressão. Realizar uma pesquisa de campo de forma qualitativa com público alvo também pode ser útil para entender o impacto das explicações, avaliar sua qualidade e levantar ideias para aperfeiçoar a técnica.

No que diz respeito à análise comparativa, a avaliação pode ser estendida para outras bases de dados, além de incluir novos métodos. Por fim, realizar a avaliação dessas técnicas com pessoas para entender os ganhos do ponto de vista do usuário e quais cenários são mais propícios para determinadas técnicas.

REFERÊNCIAS

ADADI, A.; BERRADA, M. Peeking inside the black-box: a survey on explainable artificial intelligence (xai). *IEEE access*, IEEE, v. 6, p. 52138–52160, 2018.

ANGWIN, J.; LARSON, J.; MATTU, S.; KIRCHNER, L. *Machine bias: There's software used across the country to predict future criminals. And it's biased against blacks. ProPublica, May 23*. 2016.

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)*, Wiley Online Library, v. 82, n. 4, p. 1059–1086, 2020.

ARRIETA, A. B.; DÍAZ-RODRÍGUEZ, N.; SER, J. D.; BENNETOT, A.; TABIK, S.; BARBADO, A.; GARCÍA, S.; GIL-LÓPEZ, S.; MOLINA, D.; BENJAMINS, R. et al. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information Fusion*, Elsevier, v. 58, p. 82–115, 2020.

BAKER, R.; ISOTANI, S.; CARVALHO, A. Mineraçao de dados educacionais: Oportunidades para o Brasil. *Revista Brasileira de Informática na Educação*, v. 19, n. 02, p. 03, 2011.

BAROCAS, S.; FRIEDLER, S.; HARDT, M.; KROLL, J.; VENKA-TASUBRAMANIAN, S.; WALLACH, H. *The FAT-ML Workshop Series on Fairness, Accountability, and Transparency in Machine Learning*. [S.l.]: Accessed: Jun, 2018.

BASTANI, O.; KIM, C.; BASTANI, H. Interpreting blackbox models via model extraction. *arXiv preprint arXiv:1705.08504*, 2017.

BELLE, V.; PAPANTONIS, I. Principles and practice of explainable machine learning. *arXiv preprint arXiv:2009.11698*, 2020.

BENJAMINS, R.; BARBADO, A.; SIERRA, D. Responsible ai by design in practice. *arXiv preprint arXiv:1909.12838*, 2019.

BIECEK, P. Dalex: explainers for complex predictive models in r. *The Journal of Machine Learning Research*, JMLR. org, v. 19, n. 1, p. 3245–3249, 2018.

BREIMAN, L. Random forests. *Machine learning*, Springer, v. 45, n. 1, p. 5–32, 2001.

CARVALHO, D. V.; PEREIRA, E. M.; CARDOSO, J. S. Machine learning interpretability: A survey on methods and metrics. *Electronics*, Multidisciplinary Digital Publishing Institute, v. 8, n. 8, p. 832, 2019.

CONFALONIERI, R.; COBA, L.; WAGNER, B.; BESOLD, T. R. A historical perspective of explainable artificial intelligence. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, Wiley Online Library, v. 11, n. 1, p. e1391, 2021.

COOK, R. D. Detection of influential observation in linear regression. *Technometrics*, Taylor & Francis, v. 19, n. 1, p. 15–18, 1977.

CRAVEN, M.; SHAVLIK, J. Extracting tree-structured representations of trained networks. *Advances in neural information processing systems*, v. 8, p. 24–30, 1995.

DEMAJO, L. M.; VELLA, V.; DINGLI, A. Explainable ai for interpretable credit scoring. *arXiv preprint arXiv:2012.03749*, 2020.

DHANORKAR, S.; WOLF, C. T.; QIAN, K.; XU, A.; POPA, L.; LI, Y. Who needs to know what, when?: Broadening the explainable ai (xai) design space by looking at explanations across the ai lifecycle. In: *Designing Interactive Systems Conference 2021*. [S.l.: s.n.], 2021. p. 1591–1602.

DHURANDHAR, A.; CHEN, P.-Y.; LUSS, R.; TU, C.-C.; TING, P.; SHANMUGAM, K.; DAS, P. Explanations based on the missing: Towards contrastive explanations with pertinent negatives. *arXiv preprint arXiv:1802.07623*, 2018.

DOSHI-VELEZ, F.; KIM, B. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, 2017.

DOŠILOVIĆ, F. K.; BRČIĆ, M.; HLUPIĆ, N. Explainable artificial intelligence: A survey. In: IEEE. *2018 41st International convention on information and communication technology, electronics and microelectronics (MIPRO)*. [S.l.], 2018. p. 0210–0215.

EHSAN, U.; RIEDL, M. O. Human-centered explainable ai: towards a reflective sociotechnical approach. In: SPRINGER. *International Conference on Human-Computer Interaction*. [S.l.], 2020. p. 449–466.

FISHER, A.; RUDIN, C.; DOMINICI, F. Model class reliance: Variable importance measures for any machine learning model class, from the" rashomon" perspective. *arXiv preprint arXiv:1801.01489*, v. 68, 2018.

FRIEDMAN, J. H. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, JSTOR, p. 1189–1232, 2001.

GILPIN, L. H.; BAU, D.; YUAN, B. Z.; BAJWA, A.; SPECTER, M.; KAGAL, L. Explaining explanations: An overview of interpretability of machine learning. In: IEEE. *2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)*. [S.l.], 2018. p. 80–89.

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, 2015.

GUIDOTTI, R.; MONREALE, A.; RUGGIERI, S.; PEDRESCHI, D.; TURINI, F.; GIANNOTTI, F. Local rule-based explanations of black box decision systems. *arXiv preprint arXiv:1805.10820*, 2018.

GUIDOTTI, R.; MONREALE, A.; RUGGIERI, S.; TURINI, F.; GIANNOTTI, F.; PEDRESCHI, D. A survey of methods for explaining black box models. *ACM computing surveys (CSUR)*, ACM New York, NY, USA, v. 51, n. 5, p. 1–42, 2018.

GUNNING, D. Explainable artificial intelligence (xai). *Defense Advanced Research Projects Agency (DARPA), nd Web*, v. 2, n. 2, 2017.

GURUMOORTHY, K. S.; DHURANDHAR, A.; CECCHI, G.; AGGARWAL, C. Efficient data representation by selecting prototypes with importance weights. In: IEEE. *2019 IEEE International Conference on Data Mining (ICDM)*. [S.l.], 2019. p. 260–269.

HLEG, A. Ethics guidelines for trustworthy ai. *B-1049 Brussels*, 2019.

ISLAM, S. R.; EBERLE, W.; GHAFOOR, S. K.; AHMED, M. Explainable artificial intelligence approaches: A survey. *arXiv preprint arXiv:2101.09429*, 2021.

JORDAN, M. I.; MITCHELL, T. M. Machine learning: Trends, perspectives, and prospects. *Science*, American Association for the Advancement of Science, v. 349, n. 6245, p. 255–260, 2015.

KE, G.; MENG, Q.; FINLEY, T.; WANG, T.; CHEN, W.; MA, W.; YE, Q.; LIU, T.-Y. Lightgbm: A highly efficient gradient boosting decision tree. In: *Advances in neural information processing systems*. [S.l.: s.n.], 2017. p. 3146–3154.

KIM, B.; KHANNA, R.; KOYEJO, O. O. Examples are not enough, learn to criticize! criticism for interpretability. *Advances in neural information processing systems*, v. 29, 2016.

KNAPIČ, S.; MALHI, A.; SALUJA, R.; FRÄMLING, K. Explainable artificial intelligence for human decision support system in the medical domain. *Machine Learning and Knowledge Extraction*, Multidisciplinary Digital Publishing Institute, v. 3, n. 3, p. 740–770, 2021.

KOH, P. W.; LIANG, P. Understanding black-box predictions via influence functions. In: PMLR. *International Conference on Machine Learning*. [S.l.], 2017. p. 1885–1894.

KRUSKAL, W. H.; WALLIS, W. A. Use of ranks in one-criterion variance analysis. *Journal of the American statistical Association*, Taylor & Francis, v. 47, n. 260, p. 583–621, 1952.

LACRUZ, A. J.; AMÉRICO, B. L.; CARNIEL, F. Indicadores de qualidade na educação: análise discriminante dos desempenhos na prova brasil. *Revista Brasileira de Educação*, SciELO Brasil, v. 24, 2019.

LIAO, Q. V.; GRUEN, D.; MILLER, S. Questioning the ai: informing design practices for explainable ai user experiences. In: *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. [S.l.: s.n.], 2020. p. 1–15.

LIAO, Q. V.; VARSHNEY, K. R. Human-centered explainable ai (xai): From algorithms to user experiences. *arXiv preprint arXiv:2110.10790*, 2021.

LINARDATOS, P.; PAPASTEFANOPOULOS, V.; KOTSIANTIS, S. Explainable ai: A review of machine learning interpretability methods. *Entropy*, Multidisciplinary Digital Publishing Institute, v. 23, n. 1, p. 18, 2021.

LOOVEREN, A. V.; KLAISE, J. Interpretable counterfactual explanations guided by prototypes. *arXiv preprint arXiv:1907.02584*, 2019.

LU, J.; LEE, D.; KIM, T. W.; DANKS, D. Good explanation for algorithmic transparency. *Available at SSRN 3503603*, 2019.

LUNDBERG, S.; LEE, S.-I. A unified approach to interpreting model predictions. *arXiv preprint arXiv:1705.07874*, 2017.

MITCHELL, T. M. et al. *Machine learning*. [S.l.]: McGraw-hill New York, 1997.

MORGENSTERN, O.; NEUMANN, J. V. *Theory of games and economic behavior*. [S.l.]: Princeton university press, 1953.

NETO, M. V. G.; VASCONCELOS, G. C.; ZANCHETTIN, C. Mineração de dados aplicada à predição do desempenho de escolas e técnicas de interpretabilidade dos modelos. In: SBC. *Anais do XXXII Simpósio Brasileiro de Informática na Educação*. [S.l.], 2021. p. 773–782.

PEARSON, K. Vii. note on regression and inheritance in the case of two parents. *proceedings of the royal society of London*, The Royal Society London, v. 58, n. 347-352, p. 240–242, 1895.

PINTO, G. da S.; JÚNIOR, O. F.; COSTA, E.; BARBIRATO, J. C. C.; RODRIGUES, W. R. M. Identificação dos fatores de melhorias no IDEB pelo uso de mineração de dados: Um estudo de caso em escolas municipais de maceió. In: *Simpósio Brasileiro de Informática na Educação-SBIE*. [S.l.: s.n.], 2019. v. 30, n. 1, p. 1828.

QIN, F.; LI, K.; YAN, J. Understanding user trust in artificial intelligence-based educational systems: Evidence from china. *British Journal of Educational Technology*, Wiley Online Library, v. 51, n. 5, p. 1693–1710, 2020.

QUINLAN, J. R. Induction of decision trees. *Machine learning*, Springer, v. 1, n. 1, p. 81–106, 1986.

RIBEIRO, M. T.; SINGH, S.; GUESTRIN, C. "why should i trust you?"explaining the predictions of any classifier. In: *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. [S.l.: s.n.], 2016. p. 1135–1144.

RIBEIRO, M. T.; SINGH, S.; GUESTRIN, C. Anchors: High-precision model-agnostic explanations. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. [S.l.: s.n.], 2018. v. 32, n. 1.

ROSSI, F. Ai ethics for enterprise ai (2019). *URL https://economics. harvard. edu/files/economics/files/rossifrancesca_4-22-19_ai-ethics-for-enterprise-ai_ec3118-hbs. pdf*, 2019.

SATOPAA, V.; ALBRECHT, J.; IRWIN, D.; RAGHAVAN, B. Finding a"kneedle"in a haystack: Detecting knee points in system behavior. In: IEEE. *2011 31st international conference on distributed computing systems workshops*. [S.l.], 2011. p. 166–171.

SCHOONDERWOERD, T. A.; JORRITSMA, W.; NEERINCX, M. A.; BOSCH, K. V. D. Human-centered xai: Developing design patterns for explanations of clinical decision support systems. *International Journal of Human-Computer Studies*, Elsevier, v. 154, p. 102684, 2021.

SCIENCE, C. Passeio no mundo livre. *Chico Science & Nação Zumbi. Afrociberdelia. Rio de Janeiro: Chaos*, v. 1, 1996.

SHAPLEY, L. S. *A value for n-person games, Contributions to the Theory of Games, 2, 307–317*. [S.l.]: Princeton University Press, Princeton, NJ, USA, 1953.

SILVA, M. C. d.; SOUZA, F.; TAVARES, A.; SILVA, J. D. Índice de oportunidades da educação brasileira: Variáveis explicativas de rendimento dos alunos das capitais estaduais e dos estados brasileiros. *Revista Científica Hermes*, v. 20, p. 20, 2018.

TJOA, E.; GUAN, C. A survey on explainable artificial intelligence (xai): Toward medical xai. *IEEE Transactions on Neural Networks and Learning Systems*, IEEE, 2020.

VILONE, G.; LONGO, L. Explainable artificial intelligence: a systematic review. *arXiv preprint arXiv:2006.00093*, 2020.

WACHTER, S.; MITTELSTADT, B.; RUSSELL, C. Counterfactual explanations without opening the black box: Automated decisions and the gdpr. *Harv. JL & Tech.*, HeinOnline, v. 31, p. 841, 2017.

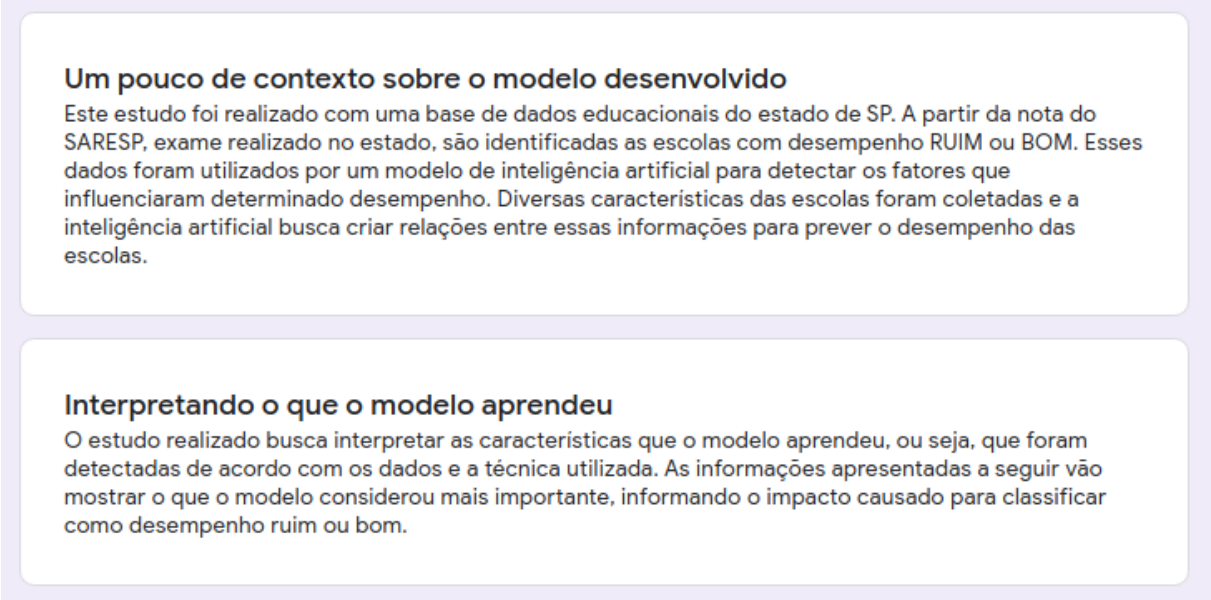
WANG, D.; YANG, Q.; ABDUL, A.; LIM, B. Y. Designing theory-driven user-centric explainable ai. In: *Proceedings of the 2019 CHI conference on human factors in computing systems*. [S.l.: s.n.], 2019. p. 1–15.

WEISBERG, S. *Applied linear regression*. [S.l.]: John Wiley & Sons, 2005. v. 528.

WEXLER, J.; PUSHKARNA, M.; BOLUKBASI, T.; WATTENBERG, M.; VIÉGAS, F.; WILSON, J. The what-if tool: Interactive probing of machine learning models. *IEEE transactions on visualization and computer graphics*, IEEE, v. 26, n. 1, p. 56–65, 2019.

APÊNDICE A – QUESTIONÁRIO DA AVALIAÇÃO

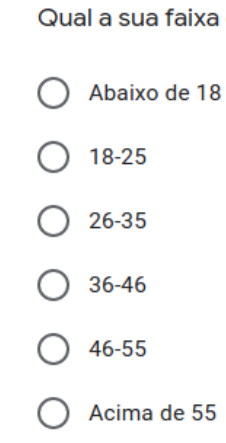
Figura 37 – Contextualização do problema e do modelo de AM



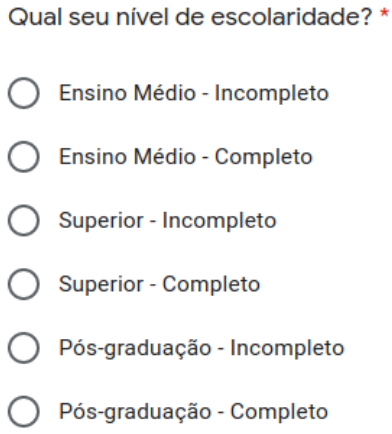
Fonte: Elaborado pelo autor

Figura 38 – Perguntas sobre o perfil do entrevistado

(a) Faixa etária



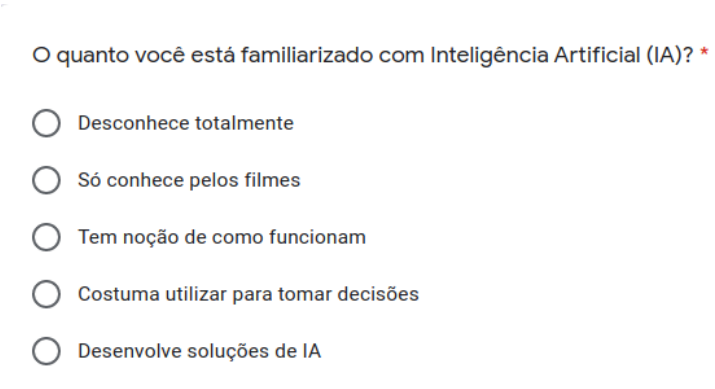
(b) Escolaridade



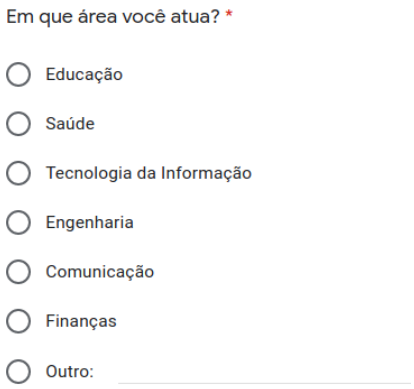
Fonte: Elaborado pelo autor

Figura 39 – Perguntas sobre o perfil do entrevistado

(a) Conhecimento em IA

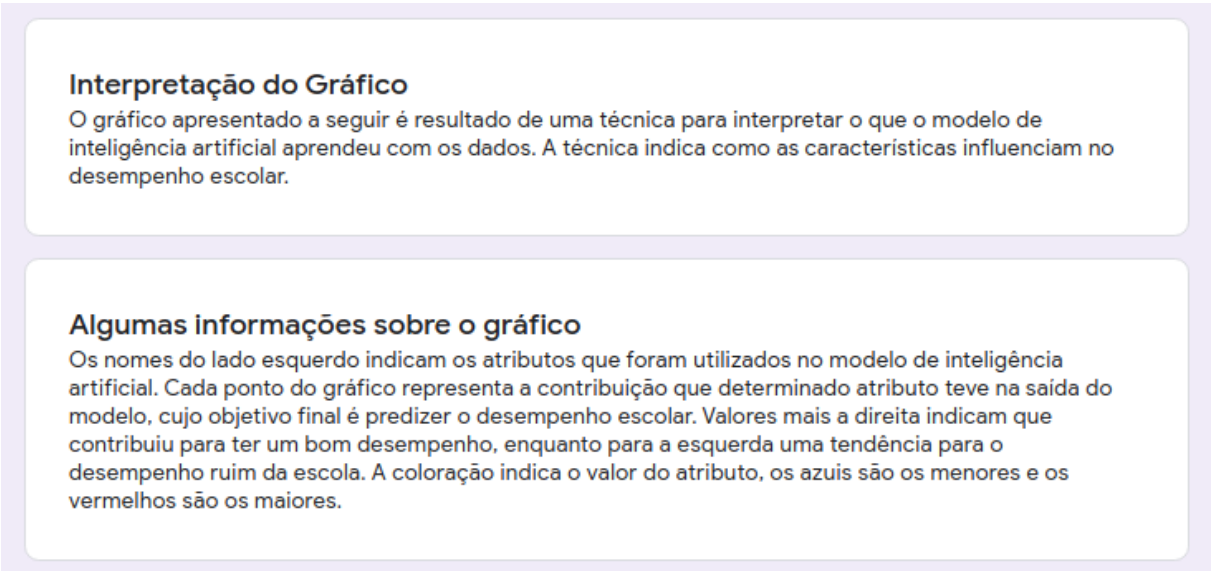


(b) Área de atuação



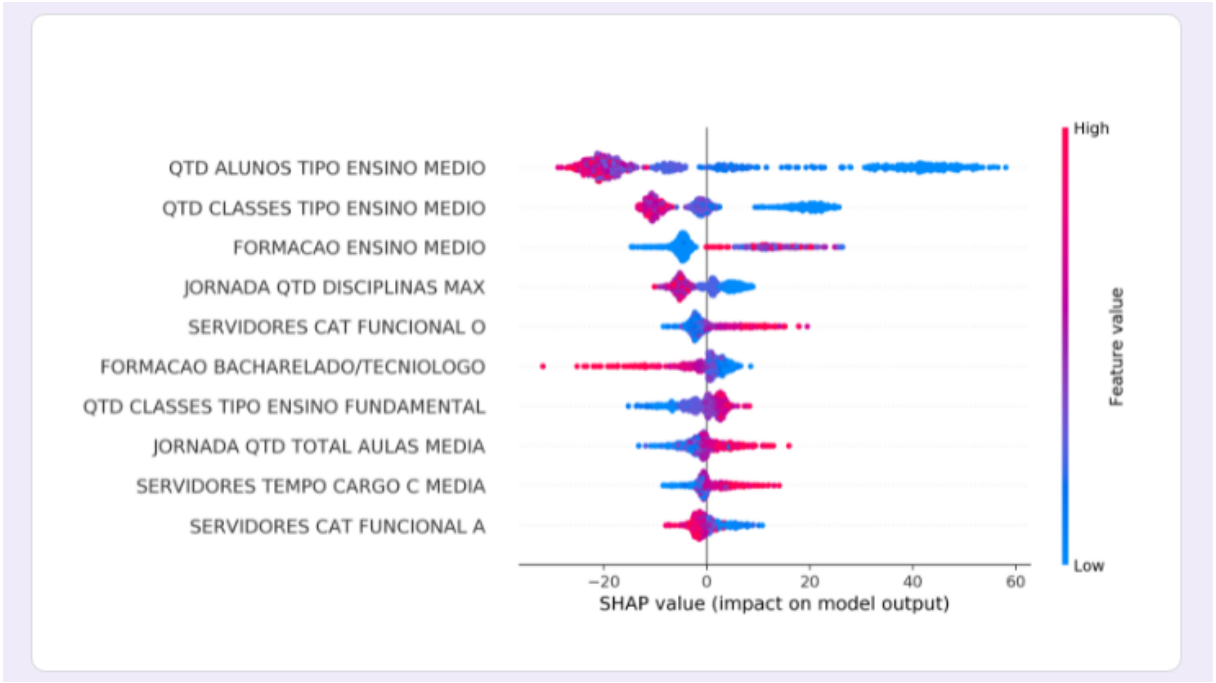
Fonte: Elaborado pelo autor

Figura 40 – Introdução a explicação com gráfico (SHAP)



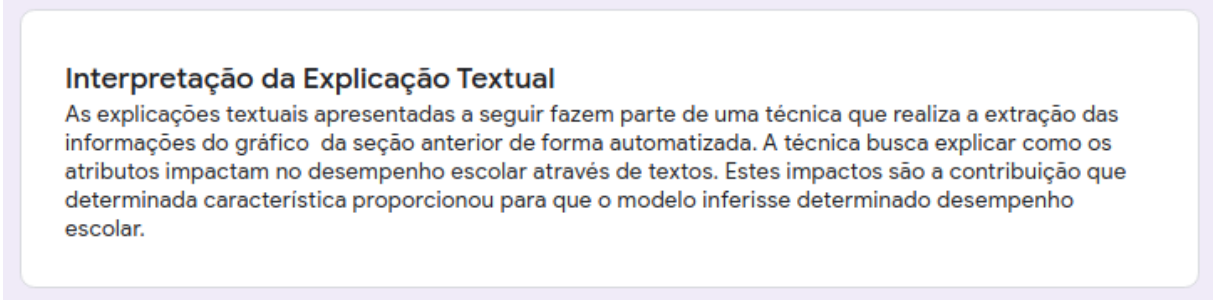
Fonte: Elaborado pelo autor

Figura 41 – Explicação por meio do gráfico



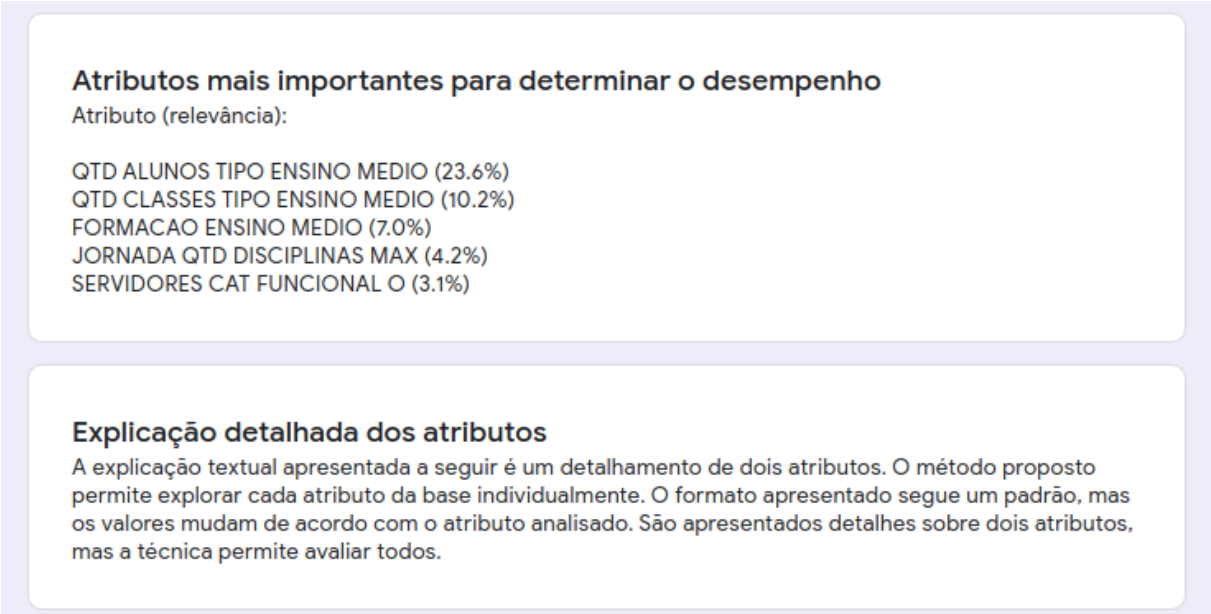
Fonte: Elaborado pelo autor

Figura 42 – Introdução as explicações textuais (Textual SHAP)



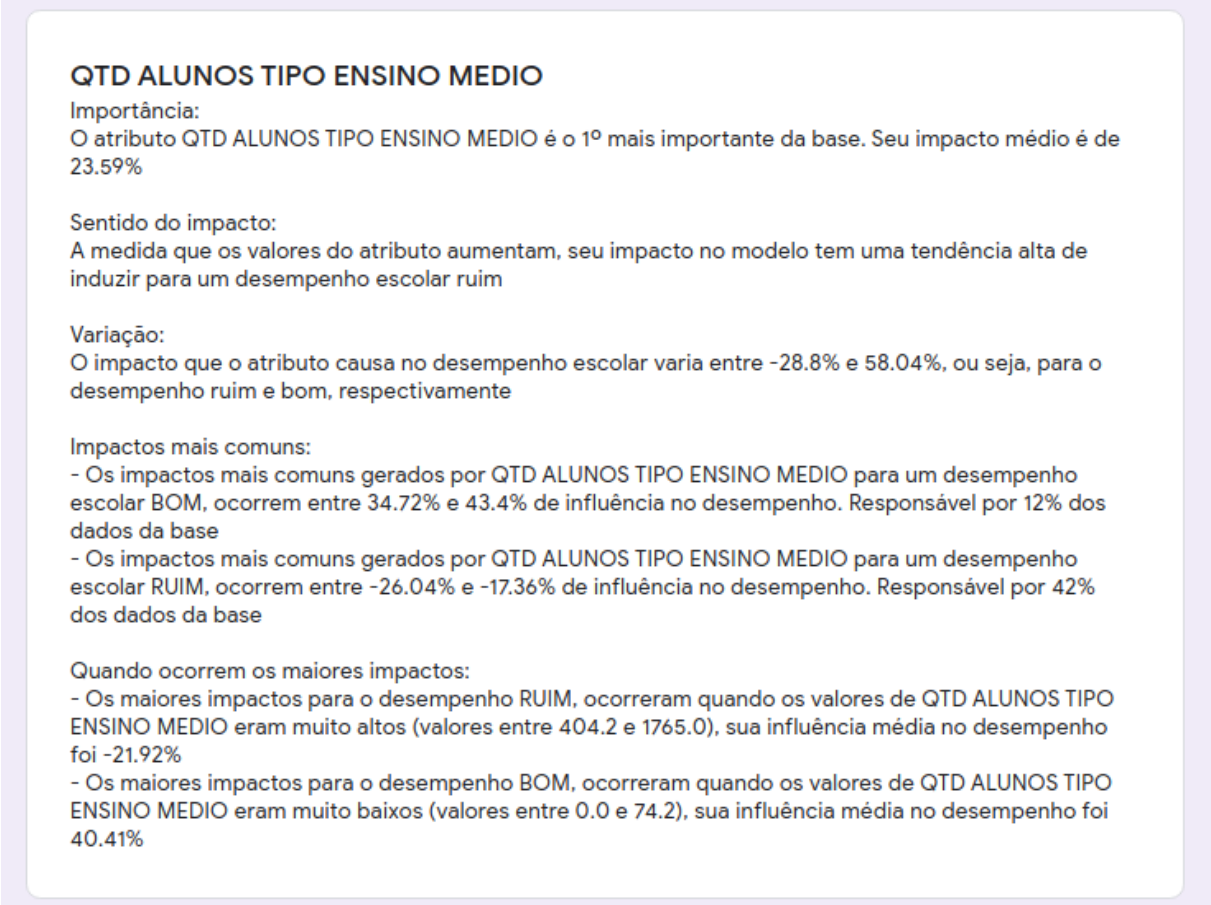
Fonte: Elaborado pelo autor

Figura 43 – Explicação textual



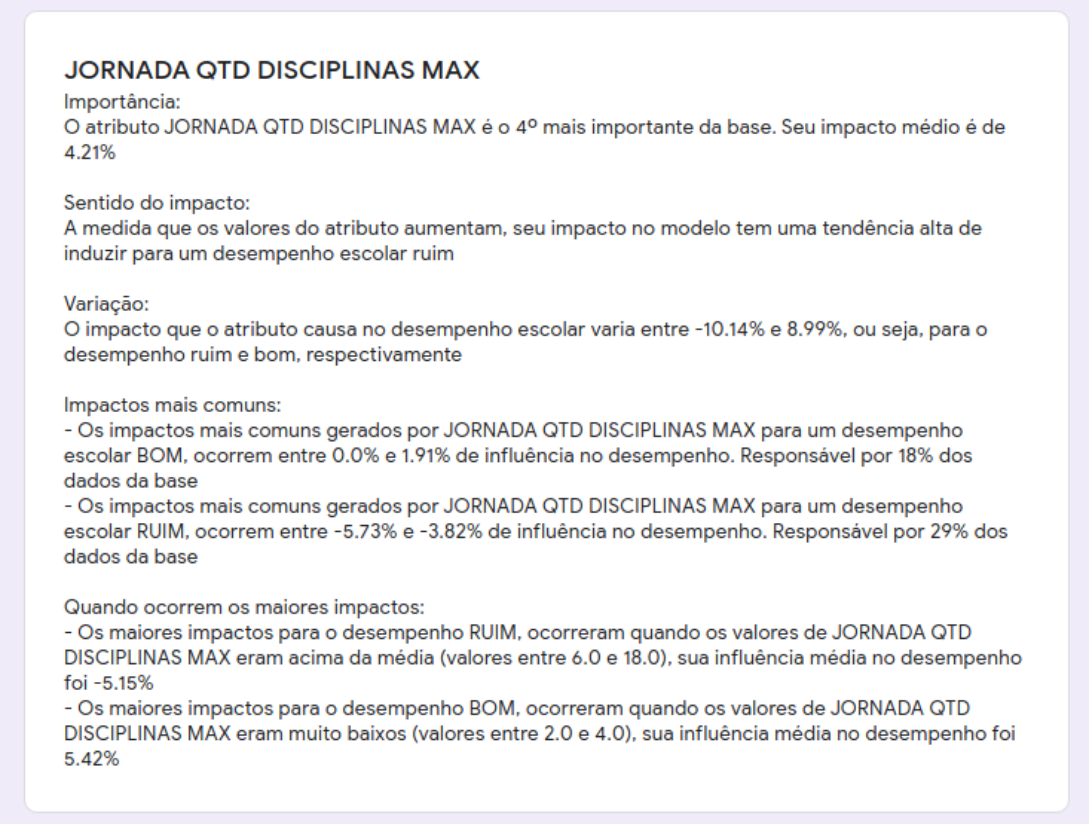
Fonte: Elaborado pelo autor

Figura 44 – Explicação textual



Fonte: Elaborado pelo autor

Figura 45 – Explicação textual



JORNADA QTD DISCIPLINAS MAX

Importância:
O atributo JORNADA QTD DISCIPLINAS MAX é o 4º mais importante da base. Seu impacto médio é de 4.21%

Sentido do impacto:
A medida que os valores do atributo aumentam, seu impacto no modelo tem uma tendência alta de induzir para um desempenho escolar ruim

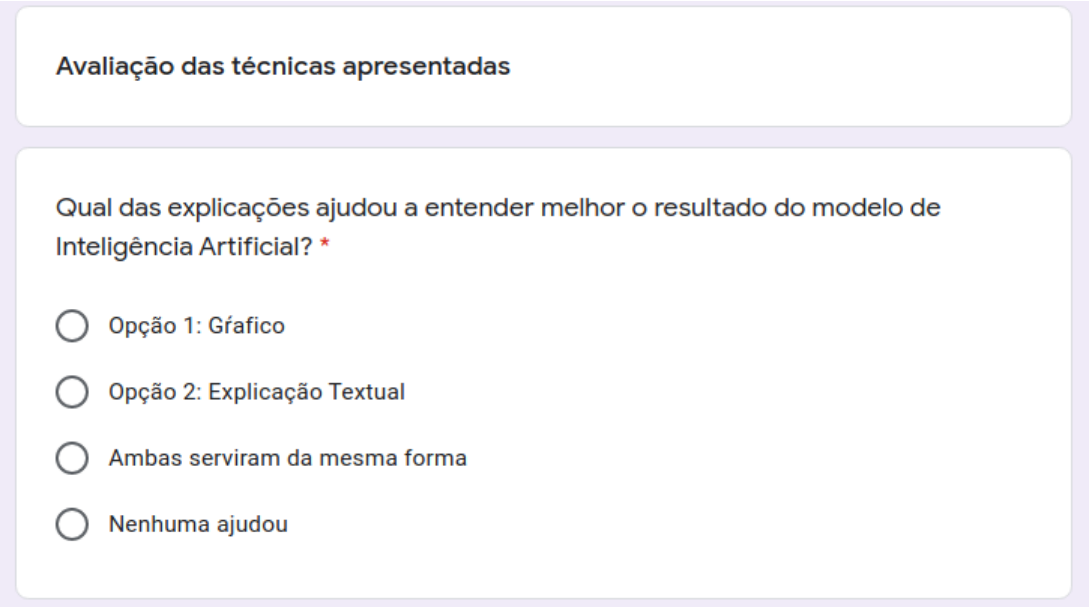
Variação:
O impacto que o atributo causa no desempenho escolar varia entre -10.14% e 8.99%, ou seja, para o desempenho ruim e bom, respectivamente

Impactos mais comuns:
- Os impactos mais comuns gerados por JORNADA QTD DISCIPLINAS MAX para um desempenho escolar BOM, ocorrem entre 0.0% e 1.91% de influência no desempenho. Responsável por 18% dos dados da base
- Os impactos mais comuns gerados por JORNADA QTD DISCIPLINAS MAX para um desempenho escolar RUIM, ocorrem entre -5.73% e -3.82% de influência no desempenho. Responsável por 29% dos dados da base

Quando ocorrem os maiores impactos:
- Os maiores impactos para o desempenho RUIM, ocorreram quando os valores de JORNADA QTD DISCIPLINAS MAX eram acima da média (valores entre 6.0 e 18.0), sua influência média no desempenho foi -5.15%
- Os maiores impactos para o desempenho BOM, ocorreram quando os valores de JORNADA QTD DISCIPLINAS MAX eram muito baixos (valores entre 2.0 e 4.0), sua influência média no desempenho foi 5.42%

Fonte: Elaborado pelo autor

Figura 46 – Avaliação das técnicas apresentadas



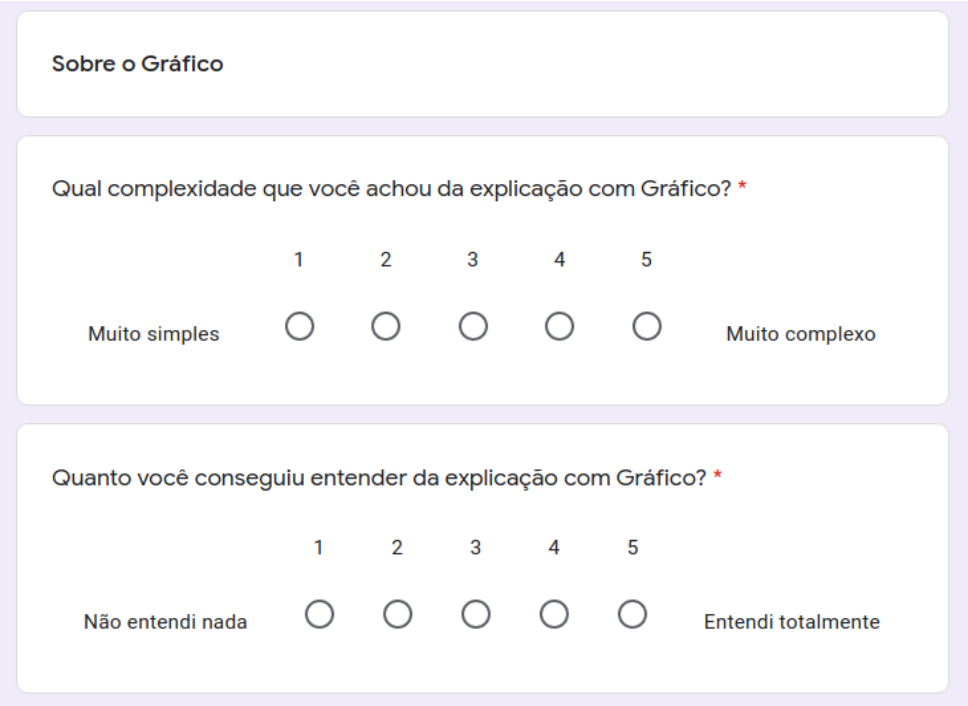
Avaliação das técnicas apresentadas

Qual das explicações ajudou a entender melhor o resultado do modelo de Inteligência Artificial? *

- ◯ Opção 1: Gŕafico
- ◯ Opção 2: Explicação Textual
- ◯ Ambas serviram da mesma forma
- ◯ Nenhuma ajudou

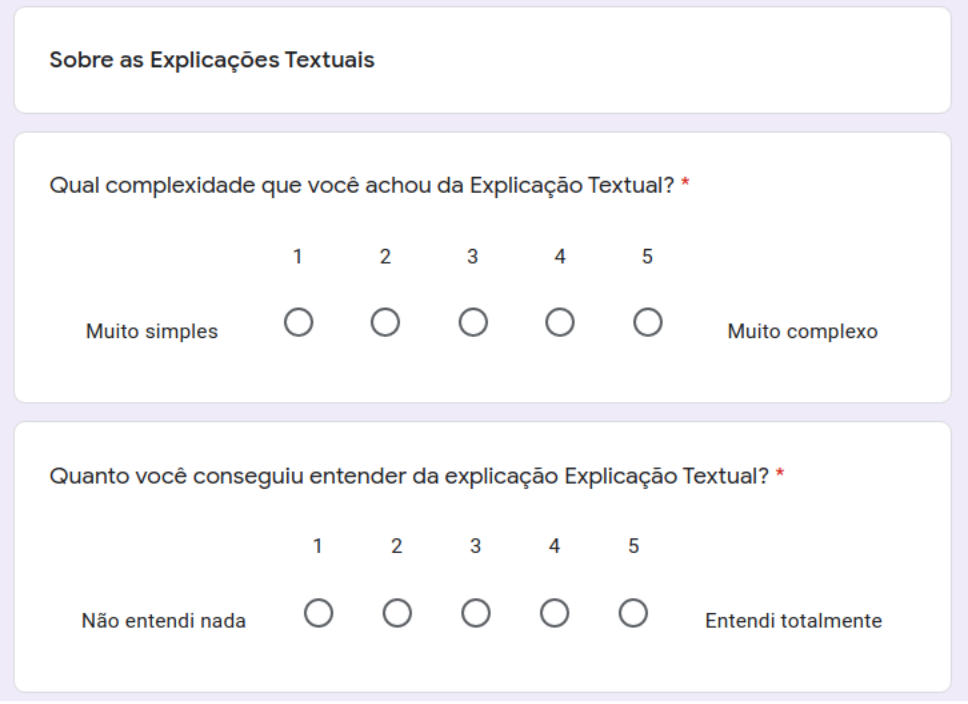
Fonte: Elaborado pelo autor

Figura 47 – Avaliação do gráfico



Fonte: Elaborado pelo autor

Figura 48 – Avaliação da explicação textual



Fonte: Elaborado pelo autor

APÊNDICE B – RESULTADOS DA AVALIAÇÃO

Figura 49 – Idade dos entrevistados

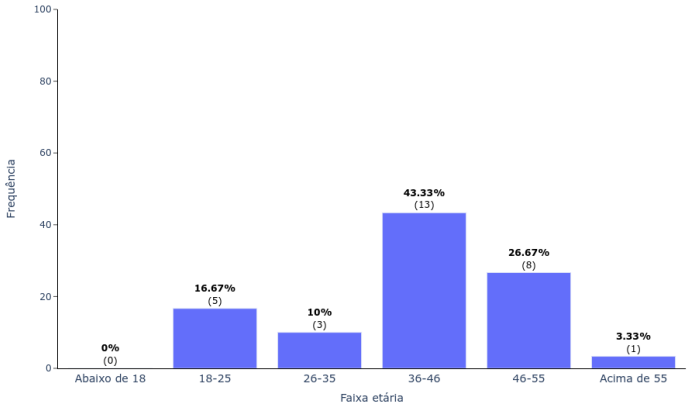
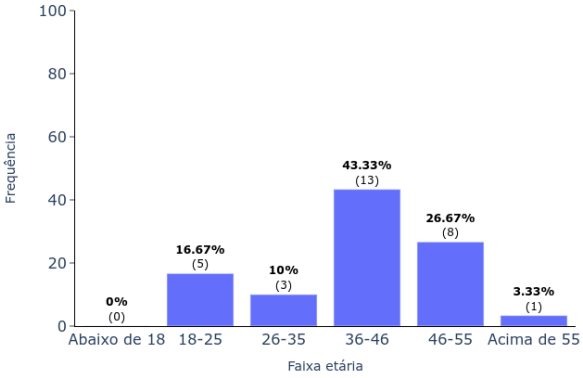


Figura 50 – Perguntas sobre o perfil do entrevistado

(a) Conhecimento em IA



(b) Área de atuação

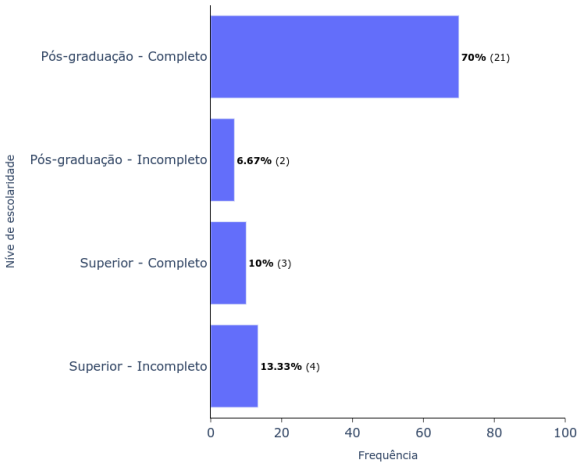
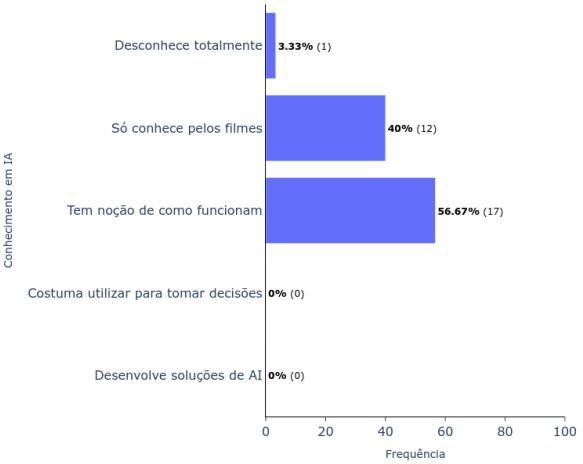


Figura 51 – Perguntas sobre o perfil do entrevistado

(a) Conhecimento em IA



(b) Área de atuação

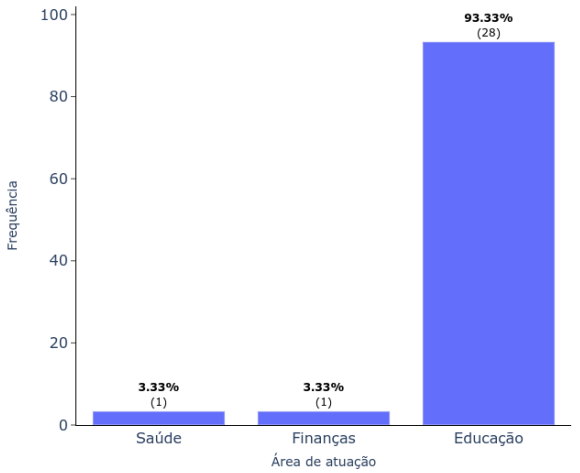


Figura 52 – Abordagem que ajuda a entender melhor o modelo

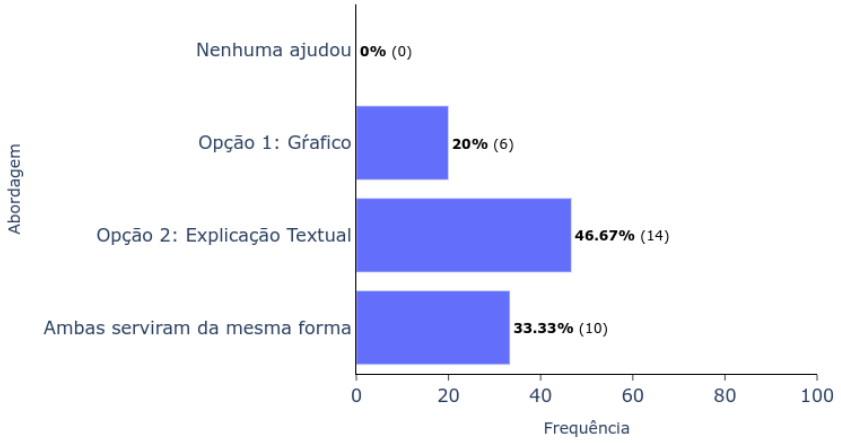
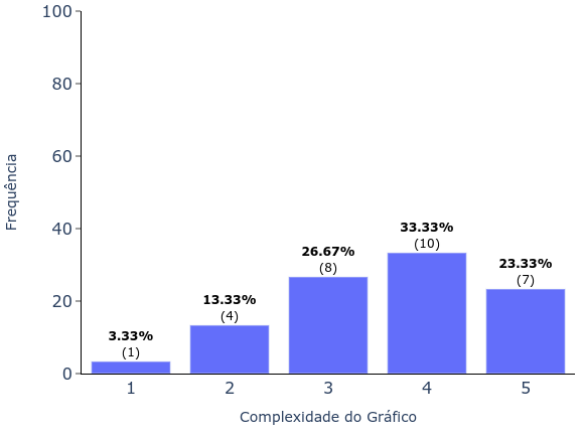


Figura 53 – Avaliação da explicação com gráfico

(a) Complexidade



(b) Entendimento

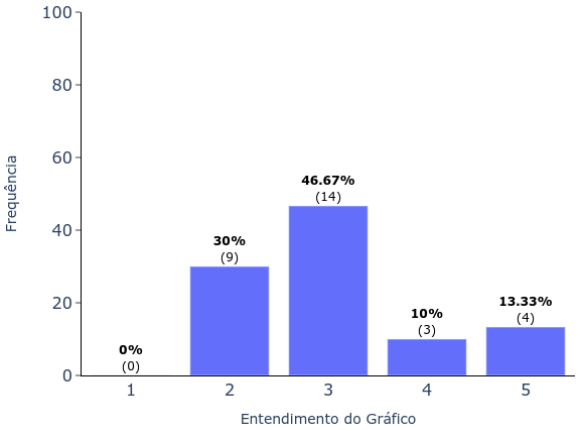
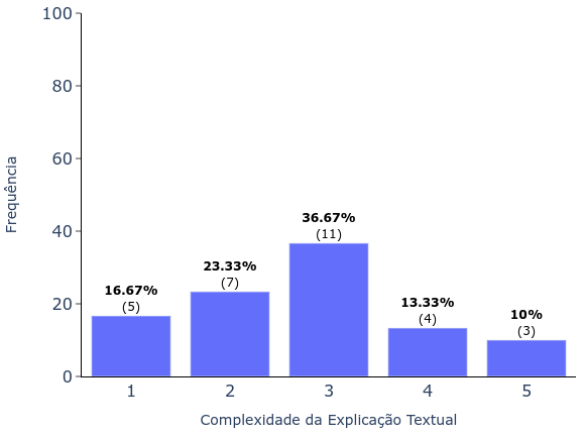


Figura 54 – Avaliação da explicação com texto

(a) Complexidade



(b) Entendimento