PONTIFÍCIA UNIVERSIDADE CATÓLICA DE MINAS GERAIS

Programa de Pós-Graduação em Informática

Marcelo de Sousa Balbino

ECOSS - UM MÉTODO AGNÓSTICO DE EXPLICAÇÕES CONTRAFACTUAIS, SELECIONADAS E SOCIAIS PARA MODELOS DE CLASSIFICAÇÃO

Belo Horizonte

Marcelo de Sousa Balbino

ECOSS - UM MÉTODO AGNÓSTICO DE EXPLICAÇÕES CONTRAFACTUAIS, SELECIONADAS E SOCIAIS PARA MODELOS DE CLASSIFICAÇÃO

Proposta de qualificação apresentada ao Programa de Pós-Graduação em Informática da Pontifícia Universidade Católica de Minas Gerais, como requisito parcial para obtenção do título de Doutor em Informática.

Orientadora: Dra.Cristiane Neri Nobre

Belo Horizonte

RESUMO

Alcançar alta capacidade preditiva nem sempre é suficiente para um modelo de aprendizado de máquina. Em muitos cenários, é necessário entender as decisões do modelo para aumentar a confiança nas previsões e direcionar as ações a serem tomadas a partir delas. Portanto, é essencial prover modelos interpretáveis. No entanto, alguns autores têm apontado a necessidade de melhorias nos métodos atuais de interpretabilidade para fornecer explicações adequadas, principalmente para não especialistas em aprendizado de máquina. A solução é expandir os estudos para além das questões computacionais e entender como as pessoas recebem melhor as explicações. Neste sentido, baseado na literatura, identificamos três aspectos a serem considerados nas explicações: serem contrastivas, selecionadas e sociais. A abordagem contrafactual, contrastiva por natureza, consiste em gerar exemplos que mostrem as pessoas como obter uma predição diferente. Diante disso, o objetivo deste trabalho é apresentar o Método Agnóstico de Explicações Contrafactuais, Selecionadas e Sociais (ECOSS) capaz de gerar explicações locais que considerem os aspectos destacados para uma abordagem orientada ao usuário final. O método utiliza uma estratégia que combina algoritmos genéticos e a abordagem SHAP como elementos centrais. Foram realizados experimentos para comparar o desempenho do ECOSS com outros métodos presentes na literatura. O ECOSS foi ainda aplicado em dois estudos de caso: um modelo de predição de desempenho acadêmico de crianças e adolescentes com TDAH e um modelo de predição de sintomatologias depressivas em crianças e adolescentes. Em ambos os casos, a aplicação do método mostrou-se útil no entendimento do modelo e na geração de conhecimentos sobre os problemas envolvidos. Os experimentos e observações mostraram que o ECOSS apresenta resultados relevantes em comparação com algumas abordagens existentes. Assim, o método proposto apresenta avanços na interpretabilidade, pois oferece explicações orientadas ao usuário final, gerando maior aceitação, confiança e compreensão nas decisões dos modelos. A implementação do método está disponível em `https://github.com/marcelobalbino/ECOSS`.

Palavras-chave: Explicações contrafactuais, Algoritmo genético, Interpretabilidade, SHAP

ABSTRACT

Achieving high predictive capability is not always enough for a machine learning model. In many scenarios, it is necessary to understand the model's decisions to increase confidence in the predictions and direct the actions to be taken from them. Therefore, it is essential to provide interpretable models. However, some authors have pointed out the need for improvements in current interpretability methods to provide adequate explanations, especially for non-specialists in machine learning. The solution is to expand studies beyond computational issues to understand the best way for people to receive explanations. In this sense, based on the literature, we identified three aspects to be considered in the explanations: being contrastive, selected, and social. The counterfactual approach, contrastive in nature, generates examples that show people how to obtain a different prediction. Given this, the objective of this work is to present the Agnostic Method of Counterfactual, Selected, and Social Explanations (ECOSS) capable of generating local explanations that consider the highlighted aspects for an end-user-oriented approach. The method uses a strategy that combines genetic algorithms and the SHAP approach as central elements. Experiments were carried out to compare the performance of ECOSS with other methods present in the literature. The ECOSS was also applied in two case studies: a model to predict the academic performance of children and adolescents with ADHD and a model for predicting depressive symptoms in children and adolescents. In both cases, the method application proved to help understand the model and generate knowledge about the problems involved. The experiments and observations showed that ECOSS presents relevant results compared to some existing approaches. Thus, the proposed method presents advances in interpretability since it offers explanations oriented to the end-user, generating greater acceptance, confidence, and understanding regarding the models' decisions. The method implementation is available at `https://github.com/marcelobalbino/ECOSS`.

Keywords: Counterfactual explanations, Genetic algorithm, Interpretability, SHAP

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

AG – Algoritmo Genético

AM – Aprendizado de Máquina

CDI – *Children's Depression Inventory*

ECOSS – Método Agnóstico de Explicações COntrafactuais, Selecionadas e Sociais

GDPR – *General Data Protection Regulation*

KNN – *K-Nearest Neighbors*

LIME – Local Interpretable Model-agnostic Explanations

LORE – *LOcal Rule-based Explanation*

RNA – Redes Neurais Artificiais

SHAP – *SHapley Additive exPlanations*

SOC – *stochastic optimization counterfactual*

SVM – *Support Vector Machine*

TD – Transtornos Depressivos

TDAH – Transtorno de Déficit de Atenção/Hiperatividade

UFMG – Universidade Federal de Minas Gerais

YSR – *Young Self Report*

SUMÁRIO

1 INTRODUÇÃO

Os sistemas de Aprendizado de Máquina (AM) estão presentes nos mais variados domínios, produtos e serviços com diferentes propósitos. A abrangência desses sistemas estende-se de tarefas cotidianas como sistemas de recomendação de filmes ou assistentes de voz, até domínios extremamente regulamentados envolvendo decisões de grande responsabilidade como aprovação de créditos bancários e hipotecas, justiça criminal e suporte a decisões médicas (ADADI; BERRADA, 2018). Diante da atuação dos modelos AM em problemas diversos e de grande impacto, pesquisas foram direcionadas ao avanço da capacidade preditiva dos modelos. Porém, a melhoria do desempenho preditivo tem sido alcançada às custas de uma maior complexidade e menor transparência nas decisões, características especialmente observadas nos modelos de aprendizados denominados caixa preta (ABDUL et al., 2018; DU; LIU; HU, 2019).

Restrições na interpretabilidade das decisões dos modelos geram inúmeros problemas e dificuldades para os envolvidos. Se a lógica e funcionamento dos modelos estão ocultos para o usuário final, isso impede que um ser humano, especialista ou não, possa verificar, interpretar e entender o raciocínio do sistema e como as decisões são tomadas (MONTAVON et al., 2017). Sendo assim, mesmo que os modelos tenham alto desempenho preditivo, fica evidente o perigo de confiar decisões importantes a sistemas que não podem se explicar, nem tampouco serem explicados por seres humanos, ou seja, é preciso prover a interpretabilidade dos modelos.

O foco na melhoria da interpretabilidade dos modelos de aprendizado de máquina se intensificou recentemente mas existem pontos a serem aprimorados. Segundo Du, Liu e Hu (2019), uma grande limitação dos trabalhos existentes é que as explicações ainda não são adequadas às demandas dos usuários finais. Para Molnar (2020), o formato das explicações atuais seria satisfatório se o público alvo fosse desenvolvedores e pesquisadores, mas é pouco apropriado para leigos em AM.

Karim et al. (2018) também destacam a importância de promover a satisfação do usuário no que se refere às saídas geradas pelos métodos de interpretabilidade. Destaca-se a necessidade de considerar explicações específicas para cada domínio e aproximá-las da forma como os seres humanos interpretam. Miller (2019) aponta a necessidade das explicações serem orientadas ao usuário, o que pode ser obtido considerando-as como um meio de comunicação entre um explicador e receptores de explicação. Sendo assim, apresenta-se a demanda por investir em métodos de interpretabilidade que gerem saídas

com maior capacidade de compreensão do ponto de vista do usuário, atendendo inclusive a não especialistas em AM.

A este respeito, Miller (2019) afirma que as pessoas parecem estar cognitivamente programadas para processar explicações contrastivas, de modo que um não especialista considerará explicações contrastivas mais intuitivas e valiosas. Mittelstadt, Russell e Wachter (2019) esclarecem que explicações contrastivas oferecem diretamente um ponto de dados alternativo: "Se seus dados fossem assim, você teria recebido essa classificação em vez dessa".

Considerando essa ótica de contraste, em muitos problemas de classificação, uma classe remete a algo desejável/positivo e outra indica algo indesejável/negativo. Por exemplo, em um modelo que se deseja prever o risco de uma doença, certamente os indivíduos desejam estar incluídos na classe que corresponde a baixo risco para a doença. Quando a previsão inclui uma dada instância na classe indesejável é natural a pergunta: "O que preciso fazer diferente para obter um resultado favorável da próxima vez?" Tal cenário remete as explicações contrastivas ou casos contrafactuais.

Segundo Stepin et al. (2021), explicações contrafactuais são consideradas contrativas por natureza, porém as explicações contrafactuais especificam mudanças mínimas necessárias na entrada para que uma saída contrastiva seja obtida. No entanto, no contexto da Inteligência Artificial (IA), os termos contrastivo e contrafactual muitas vezes são tratados como similares ou até equivalentes.

Miller (2019) evidencia ainda dois outros elementos para promover explicações orientadas ao usuário. Explicações devem ser selecionadas (provendo formas de o usuário selecionar aspectos relevantes para o seu contexto) e sociais (como parte de uma comunicação explicador/receptor da explicação e gerando conhecimento).

Sendo assim, dado um modelo de classificação f e uma instância de interesse/original x, apresenta-se o problema de explicar ao usuário final a decisão $f(x) = y$. Para esse propósito, desenvolveu-se o Método Agnóstico de Explicações COntrafactuais, Selecionadas e Sociais (ECOSS)* cujo objetivo é gerar explicações que consideram os aspectos destacados por Miller (2019) para uma abordagem orientada ao usuário final, ou seja explicações contrastivas, selecionadas e sociais. O método gera k exemplos contrafactuais $c_1, c_2, ..., c_k$ tão próximos quanto possível de x no qual $f(c_i) = y'$, onde $y' \neq y$. De cada exemplo contrafactual é extraída uma explicação local que consiste nas diferenças entre x e c_i, ou seja, o conjunto de alterações necessário nos atributos de x que o faria mudar da classe y para y'. Ressalta-se que o exemplo contrafactual c_i não precisa estar na base de dados original.

*Um método de explicação agnóstico é definido como aquele que é independente do tipo do modelo original (CARVALHO; PEREIRA; CARDOSO, 2019).

Segundo Kment (2006), gerar os casos contrafactuais pode ser descrito como um problema de otimização. De fato, é necessário minimizar as mudanças na instância original que sejam capazes de inverter a saída da predição. Soma-se ao referido problema, o fato de que os modelos de AM costumam trabalhar com bases de dados de elevado número de atributos, incluindo atributos contínuos que podem assumir infinitos valores. Logo o espaço de busca para encontrar a solução contrafactual tende a ser extenso.

Embora não constituam a única opção, o AG está entre as técnicas que melhor lidam com uma grande quantidade de atributos e é capaz de pesquisar simultaneamente em uma ampla amostragem da superfície de custo (HAUPT; HAUPT, 2004). Abordagens evolucionárias têm sido usadas com sucesso em diferentes problemas relacionados a AM (CHEN et al., 2015; GARCÍA; HERRERA, 2009; RAYMER et al., 2000; TSAI; EBERLE; CHU, 2013; KIM, 2006), inclusive em problemas similares ao tratado nesta pesquisa (DERRAC; GARCÍA; HERRERA, 2012). Em especial, Guidotti et al. (2019) obtiveram resultados satisfatórios na geração de contrafactuais por meio de AG. Diante de tal cenário, é relevante avaliar e explorar a aplicação de AG na solução desse problema. No método proposto, a cada geração, o AG busca gerar contrafactuais que se aproximem da classe desejável, mantendo a similaridade em relação à instância original. Ainda que os exemplos contrafactuais sejam gerados a partir de modificações na instância original, se as alterações forem mínimas, é possível manter a similaridade com a instância original.

Vale salientar que a abordagem contrafactual independe de uma classe desejável e outra indesejável, ainda que este tenha sido o cenário que inicialmente motivou a proposta aqui apresentada. As únicas restrições do ECOSS é que sua aplicação é direcionada a dados tabulares e problemas de classificação com resposta binária.

Um dos pontos fundamentais da solução proposta refere-se a forma como o AG avalia ao longo das gerações se os exemplos produzidos estão se aproximando da classe desejada. Para isso, usou-se o método *SHapley Additive exPlanations* (SHAP) (LUNDBERG; LEE, 2017), no qual o *predicted output value* calculado pelo método vale de pontuação para essa avaliação. O SHAP calcula o impacto/peso de cada atributo na predição de cada instância. O *predicted output value* corresponde a soma dos impactos de todos atributos para uma dada instância. O *base value*, também calculado pelo método, corresponde a fronteira entre as classe. Logo, a ideia é gerar instâncias que "cruzem" esta fronteira.

Foram realizados experimentos com o ECOSS para avaliar seu desempenho comparado aos métodos de explicação contrafactual de Wachter, Mittelstadt e Russell (2017) e Guidotti et al. (2019), no intuito de entender sua capacidade e eventuais limitações.

O trabalho buscou ainda atender duas demandas relacionadas à extração de conhecimento em bases de dados obtidas em parceria com a Universidade Federal de Minas

Gerais (UFMG):

- Por meio da base de dados disponibilizada pelo Departamento de Pediatria da universidade busca-se predizer o desempenho acadêmico de crianças e adolescentes com Transtorno de Déficit de Atenção/Hiperatividade (TDAH);
- O Programa de Pós-Graduação em Psicologia (Cognição e Comportamento) da mesma universidade concedeu os dados de crianças e adolescentes para predição de sintomatologias depressivas.

Para atender as referidas demandas foram desenvolvidos modelos de classificação e aplicou-se o ECOSS no intuito de aprofundar no entendimento das decisões de cada modelo. Por meio das explicações individuais, somadas a algumas implementações adicionais, foi possível entender decisões individuais, identificar os atributos de maior impacto nas decisões e observar o comportamento destes atributos. Para os envolvidos no problema, os resultados obtidos servem de ferramenta de auxílio na tomada de decisão em cenários de grande relevância. Para o desenvolvimento deste trabalho, tais cenários foram fundamentais para a experimentação e evolução do método proposto.

1.1 Problema

Diante do cenário atual dos métodos de interpretabilidade, seus avanços e limitações, surge a seguinte questão de pesquisa:

Dado um modelo de classificação f e uma instância de interesse/original x, é possível explicar ao usuário final, especialistas ou não em AM, a decisão $f(x) = y$, de modo que o mesmo compreenda tal decisão, independente do algoritmo aplicado ao modelo?

1.2 Hipóteses

Este trabalho é baseado fundamentalmente nas seguintes hipóteses:

- Hipótese 1: é possível incluir recursos em um método de interpretabilidade agnóstico que o permita atender as diretrizes indicadas por Miller (2019) para uma explicação orientada ao usuário final;
- Hipótese 2: uma das diretrizes de Miller (2019) é a utilização de explicações contrafactuais devido sua maior capacidade de comunicação com o usuário. O contrafactual deve ser gerado com o mínimo de mudanças na instância original. Diante disso, a segunda hipótese deste trabalho é que a utilização conjunta de AG e a abordagem SHAP é capaz de gerar contrafactuais tão próximos quanto possível da instância original.

1.3 Objetivos

1.3.1 Objetivo geral

O objetivo geral deste trabalho é desenvolver um método agnóstico de interpretabilidade capaz de gerar explicações locais que consideram os aspectos destacados por Miller (2019) para uma abordagem orientada ao usuário final, ou seja, explicações contrastivas, selecionadas e sociais.

1.3.2 Objetivos específicos

Os objetivos específicos deste trabalho são:

- Comparar o método desenvolvido neste trabalho com outros métodos presentes na literatura;
- Aplicar o método desenvolvido nos modelos de classificação do desempenho acadêmico de estudantes com TDAH e de predição de depressão em crianças e adolescentes, de maneira a oferecer aos envolvidos nos problemas uma ferramenta de apoio a tomada de decisão e extração de conhecimento.

1.4 Contribuições da tese

Os atuais métodos de interpretabilidade têm apresentado limitações em relação a compreensão das explicações geradas para os usuários finais, especialmente para não especialistas em AM. Sendo assim, uma das mais significativas contribuições deste trabalho refere-se ao fato de o método de explicações proposto apresentar uma abordagem orientada ao usuário, questão que notadamente carece de aprimoramento considerando o cenário atual. Entende-se que o método é capaz de elevar a capacidade de compreensão dos usuários sobre às decisões dos modelos, ao incluir elementos alinhados às diretrizes apontadas por Miller (2019) sobre a forma como as pessoas melhor recebem explicações.

Os resultados obtidos nos experimentos mostraram que o método apresenta avanços relevantes quando comparado aos métodos encontrados na literatura. Além disso, ser um método agnóstico torna-o amplamente aplicável.

Ressalta-se ainda os resultados obtidos por meio da aplicação do método nas pesquisas desenvolvidas com a instituição parceira como um suporte importante para os grupos envolvidos nos referidos problemas.

Sobre os modelos de classificação e explicação gerados a partir da base de dados de estudantes com TDAH, os resultados obtidos podem direcionar as ações dos pais,

educadores e demais profissionais (psicólogos, psiquiatras e neurologistas), na busca por melhor desempenho acadêmico para os discentes com o transtorno.

Da mesma forma, no problema de classificação de sintomatologias depressivas, o modelo gerado mostrou-se capaz de subsidiar o diagnóstico e apontar ações para auxiliar indivíduos com depressão. Logo, diante da gravidade e a quantidade de pessoas acometidas pelo transtorno, destaca-se a importância de uma ferramenta capaz de auxiliar na identificação e encaminhamento do tratamento de crianças e adolescentes com depressão, sendo esta outra contribuição a ser destacada nesta pesquisa.

1.5 Organização da tese

Este trabalho apresenta a seguinte estrutura: no Capítulo 2 (página 15), são apresentados os principais conceitos relacionados a interpretabilidade, explicações e a abordagem SHAP. O Capítulo 3 (página 38) refere-se aos trabalhos relacionados, apresentando outros métodos de explicações contrafactuais e suas similaridades e diferenças em relação ao método proposto neste trabalho. O Capítulo 4 (página 44) detalha o método ECOSS proposto. O Capítulo 5 (página 52) apresenta os materiais e métodos utilizados neste trabalho. No Capítulo 6 (página 63), são apresentados experimentos que permitem comparar o desempenho do ECOSS em relação a geração de contrafactuais com outros métodos encontrados na literatura. No Capítulo 7 (página 67), encontram-se resultados da aplicação do método em alguns estudos de caso. No Capítulo 8 (página 81), são expostas as considerações finais e trabalhos futuros. Finalmente, no Capítulo 9 (página 83), é apresentado o cronograma com as etapas concluídas e aquelas ainda a serem desenvolvidas no trabalho.

2 REFERENCIAL TEÓRICO

2.1 Interpretabilidade

Tem se tornado cada vez mais comum a utilização de sistemas de aprendizado de máquina em problemas de alto risco que impactam profundamente na vida humana e na sociedade. Os modelos têm prestado suporte a questões onde uma decisão correta é essencial (RUDIN, 2019). A partir deste cenário, houve um movimento da comunidade científica em busca de modelos cada vez melhores do ponto de vista de desempenho preditivo. Muitas pesquisas foram direcionadas nesse sentido e resultados relevantes foram obtidos, especialmente por meio dos chamados métodos de caixa preta (ABDUL et al., 2018).

Porém, a melhoria do desempenho preditivo tem sido alcançada às custas de uma maior complexidade e menor transparência nas decisões (DU; LIU; HU, 2019). Rudin (2019) e El Shawi et al. (2019) também evidenciam esse *trade off* entre desempenho e transparência e destacam como limitações na interpretabilidade das decisões prejudicam a confiança no modelo e sua consequente utilização na prática.

Além disso, em muitos domínios ter um modelo com alto desempenho preditivo é apenas uma solução parcial do problema. Em cenários de alto risco, como o médico, por exemplo, o tomador de decisão se sente inseguro sem uma explicação dos resultados do modelo (Tjoa; Guan, 2020). Logo, é preciso apresentar elementos que permitam desenvolvedores e usuários entenderem melhor as decisões do sistema, elevando a confiança nos resultados e, quando necessário, permitindo perceber decisões incorretas.

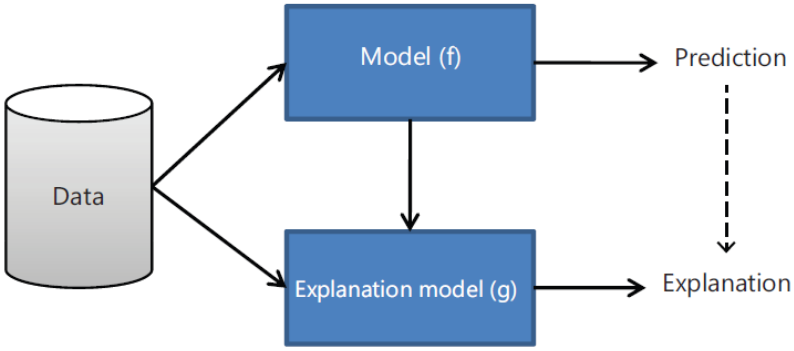
Diante deste contexto, surgem até mesmo questionamentos em relação a medidas de qualidade dos modelos de aprendizado de máquina que se baseiam apenas no desempenho preditivo. Por exemplo, um modelo usado em um sistema bancário que sugere pelo empréstimo ou não para um determinado cliente deve ter o objetivo de reduzir o índice de inadimplência e não discriminar contra qualquer pessoa com base na raça ou lugar onde vive. No geral, as técnicas de aprendizado de máquina preocupam-se apenas em otimizar a métrica de inadimplência do empréstimo e não se importam com outros fatores como discriminação, por exemplo. Logo, é necessário incluir outros fatores importantes na avaliação das técnicas de aprendizado de máquina, sendo a interpretabilidade um deles (KARIM et al., 2018).

Por conseguinte, a *eXplainable Artificial Intelligence* (XAI), campo de estudo com

foco na interpretabilidade dos sistemas de AM, tem ganhado a atenção da comunidade científica. O objetivo é contribuir com a criação de métodos que permitam interpretar os modelos, preservando altos níveis de desempenho preditivo (ADADI; BERRADA, 2018; MILLER, 2019).

Conforme ilustrado na Figura 1, os métodos de interpretabilidade introduzem uma nova perspectiva para soluções de AM adicionando um modelo de explicação(g) ao modelo de previsão original(f). O objetivo desses métodos é apresentar uma aproximação interpretável do modelo original (LUNDBERG; LEE, 2017).

Figura 1 – Como um modelo de explicação é usado na interpretação da previsão.



Fonte: Extraído de Mokhtari, Higdon e Başar (2019).

2.1.1 Definição e terminologia

Dentro do contexto relacionado a transparência dos modelos de AM, emergem termos como interpretabilidade, explicabilidade, justificabilidade e explicação, cuja utilização gera alguma discussão.

Segundo Biran e Cotton (2017), o nível de interpretabilidade de um modelo está relacionado ao grau em que um observador pode compreender suas decisões. A explicação é justamente um modo pelo qual um observador pode obter tal compreensão. As outras possibilidades são os modelos intrinsicamente interpretáveis (Seção 2.1.5.1, página 21) ou por meio de introspecção. Logo, o aprendizado de máquina interpretável engloba métodos e modelos que fazem o comportamento e as previsões de sistemas de aprendizado de máquina compreensível para os humanos (MOLNAR, 2020).

O mais comum na comunidade de AM é que os termos interpretabilidade e explicabilidade sejam usados de forma intercambiável (SILVA et al., 2018; ADADI; BERRADA, 2018; MILLER, 2019; MOLNAR, 2020). No entanto, o termo interpretável é mais usado do que explicável (CARVALHO; PEREIRA; CARDOSO, 2019).

Por outro lado, alguns autores como Gilpin et al. (2018), apresentam uma distinção entre interpretabilidade e explicabilidade. Para os autores, a *interpretabilidade* está relacionada a descrição das decisões de um modelo de uma forma compreensível para os humanos. Assim, é preciso considerar a cognição, conhecimento, preconceitos do usuário e usar um vocabulário que seja significativo para o mesmo. Já a *explicabilidade* é análogo a explicar o rastreio de execução de um programa. Por exemplo, no caso de uma rede neural, é necessário resumir as razões que levaram a um determinado comportamento da rede. As explicações sobre as operações de redes profundas podem, por exemplo, enfocar no processamento ou na representação dos dados dentro da mesma. Os autores consideram que a interpretabilidade traz a compreensão e explicabilidade gera a confiança nas decisões do modelo. Dentro desta visão, modelos explicáveis são interpretáveis por padrão, mas o inverso nem sempre é verdadeiro.

Finalmente, o termo justificabilidade refere-se a validação de um especialista de domínio em relação as decisões do modelo. Em outras palavras, um modelo é justificável quando está alinhado com conhecimento de domínio existente (MARTENS et al., 2011).

Neste trabalho, foram consideradas as definições que a literatura mostrou ser mais comum na comunidade de AM em relação aos termos interpretabilidade e explicabilidade, entendendo-os como equivalentes. Sendo o termo interpretável (e suas variações) o mais frequentemente usado, este foi empregado ao longo do texto.

2.1.2 Histórico da interpretabilidade

Ao longo dos anos, a relevância do tema interpretabilidade para a comunidade acadêmica passou por diferentes estágios, de maneira a existir registros descontinuados de interesse nas explicações de sistemas inteligentes (CARVALHO; PEREIRA; CARDOSO, 2019).

Na década de 1970, a necessidade de explicações esteve ligada ao uso dos sistemas especialistas. Ainda que, no geral, as heurísticas utilizadas nesses sistemas tenham boa performance, podem haver casos onde o sistema produza resultados incorretos. Naquele momento, a necessidade de explicações geralmente estavam relacionadas a suspeita e identificação de defeitos nos sistemas especialistas (SWARTOUT, 1983).

A partir da década de 80, a compreensão das Redes Neurais Artificiais (RNA) também passou a receber alguma atenção. Surgiram alguns questionamentos sobre a importância de aumentar a capacidade de explicação das RNA para expandir sua aplicação. Nessa linha, por exemplo, Andrews, Diederich e Tickle (1995) apresentaram mecanismos para extração de regras para elevar a compreensão das redes.

Nos anos 2000, houve um certo interesse em entender o quanto a transparência

dos sistemas de recomendação influenciava a confiança dos usuários. Esses são sistemas que visam prever a afinidade de uma pessoa por itens ou informações, como um filme, por exemplo. A partir dos registros de interesse de um indivíduo e de comunidade de pessoas, o sistema procura indicar itens que mais se aproximam do perfil desse indivíduo. Percebeu-se a transparência como um fator importante para aceitação das recomendações realizadas por esses sistemas (HERLOCKER; KONSTAN; RIEDL, 2000).

Em meados de 2010, ocorreu uma redução no ritmo do progresso para resolver os problemas relacionados a interpretabilidade, já que a prioridade das pesquisas estava focada na melhoria do poder preditivo dos algoritmos e modelos (CARVALHO; PEREIRA; CARDOSO, 2019).

Finalmente, aproximadamente a partir de 2015, a expansão dos sistemas de aprendizado de máquina para muitos domínios, bem como o uso de algoritmos mais complexos e não transparentes e até mesmo questões legais relacionadas ao "direito de explicação" proveniente da Lei de Proteção de Dados trouxeram novamente o interesse na necessidade de uma melhor compreensão dos resultados dos referidos sistemas (ABDUL et al., 2018; CARVALHO; PEREIRA; CARDOSO, 2019).

No entanto, segundo Carvalho, Pereira e Cardoso (2019), ainda que as pesquisas relacionadas a interpretabilidade tenham se intensificado recentemente, estas ainda constituem um subconjunto relativamente pequeno de toda a pesquisa em AM quando comparado ao interesse no desenvolvimento de técnicas e modelos de AM, bem como a busca pela melhoria do desempenho preditivo dos modelos. Logo, é necessário elevar o conhecimento científico nessa área.

2.1.3 A função da interpretabilidade

A restrição de entendimento das decisões dos modelos gera inúmeros problemas e dificuldades para os desenvolvedores, usuários e outros interessados em suas decisões. Neste sentido, podem-se destacar questões relacionadas a:

- *Confiança*: segundo Ribeiro, Singh e Guestrin (2016), "se os usuários não confiarem em um modelo ou previsão, eles não o usarão". Assim, ao estabelecer a interpretabilidade por meio de uma explicação, os usuários ganham mais confiança no modelo e ficam mais propensos a aceitá-lo e usá-lo (HONEGGER, 2018);
- *Segurança*: para o desenvolvedor, a falta de transparência dos modelos dificulta a percepção de decisões equivocadas por parte dos sistemas. Quando os dados fornecidos ao modelo possuem distribuições tendenciosas, esses padrões são aprendidos e previsões orientadas para um dado comportamento são retornadas e nem sempre notadas (HONEGGER, 2018). A interpretabilidade ajuda os desenvolvedores a

verificar e melhorar o modelo, aumentando sua segurança (MOLNAR, 2020);

- *Justiça*: para os usuários, grupos minoritários podem ser discriminados, injustiças podem não ser percebidas e as pessoas prejudicadas ainda permanecem com poucos recursos para argumentar. Além disso, na maioria das vezes, as entidades detentoras dos sistemas não conseguem explicar como as decisões foram tomadas devido à opacidade desses sistemas (CARVALHO; PEREIRA; CARDOSO, 2019). Isso tem levado entidades ao redor do mundo a criar algum tipo de regulamentação, como o artigo 13 do *General Data Protection Regulation* (GDPR) proposto pela União Europeia. Segundo o regulamento, em modelos que usam dados pessoais dos usuários para tomar decisões, por exemplo, para obter um empréstimo bancário ou decidir se a pessoa terá um dado tratamento específico ou não, o usuário tem direito a uma explicação que o permita entender porque o modelo chegou a uma determinada decisão (KARIM et al., 2018). Diante disso, a interpretabilidade incorpora não só um papel social, mas também uma exigência legal na tentativa de proteger os usuários submetidos as decisões dos sistemas;
- *Conhecimento*: existe um interesse crescente da comunidade acadêmica e industrial em interpretar modelos de aprendizado de máquina. Isso se deve ao fato desses sistemas terem grande capacidade de obter *insights* que podem elevar o nível de entendimento em relação ao problema e desencadear descobertas científicas e industrias (DU; LIU; HU, 2019). É da natureza dos sistemas de aprendizado de máquina a descoberta de conhecimento não óbvio. No entanto, muitas vezes esse conhecimento fica oculto em virtude da falta de transparência do modelo. Por exemplo, um sistema médico capaz de identificar se determinado paciente tem potencial de desenvolver uma doença tem grande relevância. Mas se o mesmo sistema for capaz de gerar explicações que apontem as variáveis que elevam esse potencial, isso pode suscitar novas pesquisas e descobertas.

Ainda no papel da interpretabilidade para geração de conhecimento, Karim et al. (2018) afirmam que técnicas de aprendizado de máquina são frequentemente usadas em campos científicos e, desta forma, devem não somente classificar ou prever mas também responder às perguntas “como” e “por que” para serem coerentes com os objetivos da ciência. Os autores destacam a importância de considerar os modelos de aprendizado de máquina e a inteligência artificial como formas de elevar a compreensão humana em relação a diferentes problemas do mundo real e não simplesmente confiar nos modelos. Isso significa dizer que, em muitos casos, o conhecimento por trás da previsão é tão ou mais valioso que ela própria.

2.1.4 Interpretabilidade local e global: dimensões da interpretabilidade

Os métodos de interpretabilidade podem se propor a gerar explicações a respeito de diferentes aspectos do processo de previsão: globalmente ou localmente.

A *interpretabilidade global* refere-se a uma compreensão holística de como o modelo toma decisões, significa apresentar um entendimento geral dos resultados da previsão com base nas características, respondendo à pergunta de como o modelo treinado faz previsões (MOLNAR, 2020).

Já a *interpretabilidade local* tem o objetivo de apresentar a explicação de como o modelo realizou a previsão para uma instância em particular. Uma alternativa é criar uma aproximação de uma pequena região de interesse do modelo caixa preta usando um modelo interpretável mais simples, ou seja, um modelo substituto (RÜPING, 2006).

Esses dois tipos de interpretabilidades trazem diferentes benefícios. A interpretabilidade global tem a capacidade de iluminar os mecanismos internos de trabalho dos modelos de aprendizado de máquina e, portanto, podem aumentar sua transparência. Já a interpretabilidade local ajudará a descobrir as relações causais entre uma entrada específica e sua previsão correspondente no modelo. Enfim, a interpretabilidade global ajuda os usuários a confiarem no modelo e a interpretabilidade local aumenta a confiança em uma dada previsão (DU; LIU; HU, 2019).

As explicações locais tendem a ser mais fiéis que explicações globais, uma vez que localmente a previsão pode depender apenas linearmente de alguns atributos, em vez de ter uma dependência complexa entre eles. Já a interpretabilidade global é mais difícil de ser alcançada uma vez que os modelos, especialmente os mais complexos, realizam diversas associações, relacionando atributos, pesos e parâmetros para obter as previsões (MOLNAR, 2020). Por exemplo, pode ser muito difícil descrever o mapeamento geral realizado por uma rede neural e por isso alguns trabalhos preferem explicações locais da rede (KARIM et al., 2018).

Ribeiro, Singh e Guestrin (2016) destacam que existe um *trade-off* entre fidelidade e interpretabilidade, uma vez que para uma explicação ser completamente fiel, essa precisaria ser uma descrição completa do próprio modelo, o que compromete sua interpretabilidade. Os autores complementam que para uma explicação ser significativa deve ser pelo menos localmente fiel, ou seja, deve corresponder a como o modelo se comporta na vizinhança da instância que está sendo prevista.

2.1.5 Classificações dos métodos de interpretabilidade

Existem na literatura algumas formas de categorizar os métodos a partir de diferentes critérios. Esta seção apresenta algumas dessas classificações.

2.1.5.1 Modelos intrinsicamente interpretáveis e métodos post hoc

Esta classificação é baseada na forma como a interpretabilidade é obtida (MOLNAR, 2020).

A interpretabilidade intrínseca está presente em modelos autoexplicativos que incorporam interpretabilidade diretamente às suas estruturas, ou seja, naturalmente interpretáveis (DU; LIU; HU, 2019; CARVALHO; PEREIRA; CARDOSO, 2019). Essa categoria inclui árvores de decisão, modelos baseado em regras, modelos lineares e outros. Segundo Karim et al. (2018), mesmo os modelos lineares em casos de alta dimensão na base de dados, que é o cenário mais comum no mundo real, são pouco compreensíveis e por vezes ainda pouco precisos.

Por outro lado, quando é necessária a criação de um segundo modelo para fornecer explicações para o modelo principal classifica-se a interpretabilidade como *post hoc*. Por definição, métodos de interpretabilidade *post hoc* são aplicados após o treinamento do modelo (CARVALHO; PEREIRA; CARDOSO, 2019).

Segundo Du, Liu e Hu (2019), em linhas gerais, a principal diferença entre esses dois grupos está no *trade-off* entre a acurácia do modelo e a fidelidade da explicação. Modelos inerentemente interpretáveis são capazes de fornecer explicações precisas e não distorcidas, mas podem sacrificar o desempenho da previsão em algumas situações. Os modelos *post hoc* nem sempre garantem uma explicação fiel do comportamento do modelo de previsão, mas não interferem na acurácia do mesmo.

2.1.5.2 *Pre-model*, *in-model* e *post-model*

Trata-se de outra forma de agrupar os métodos em relação ao momento em que a interpretabilidade é aplicada: se antes, durante ou depois da construção do modelo aprendizado de máquina. Segundo este critério, a interpretabilidade pode ser (CARVALHO; PEREIRA; CARDOSO, 2019):

- *Pre-model*: técnicas de interpretabilidade deste tipo são independentes do modelo, pois estão relacionadas apenas a técnicas de análise dos próprios dados. Incluem técnicas de estatística descritiva clássica, Análise de Componentes Principais e métodos de agrupamento, como o *k-means*, por exemplo. Essas são técnicas que permitem explorar e visualizar graficamente os dados permitindo o bom entendimento dos

mesmos;

- *In-model*: refere-se aos modelos de aprendizado de máquina que possuem interpretabilidade inerente, sendo assim chamados de intrinsecamente interpretáveis;
- *Post-model*: refere-se a prover a interpretabilidade após a construção do modelo de previsão.

Esta classificação de certa forma se sobrepõe a apresentada na Seção 2.1.5.1, página 21. Alguns autores como Du, Liu e Hu (2019) consideram apenas os modelos como intrinsicamente interpretáveis e *post hoc*. Outros, como Carvalho, Pereira e Cardoso (2019), dividem em *pre-model*, *in-model* e *post-model*, no qual o *in-model* corresponde ao intrinsicamente interpretável e o *post-model* refere-se ao *post hoc*.

2.1.5.3 Métodos específicos e agnósticos

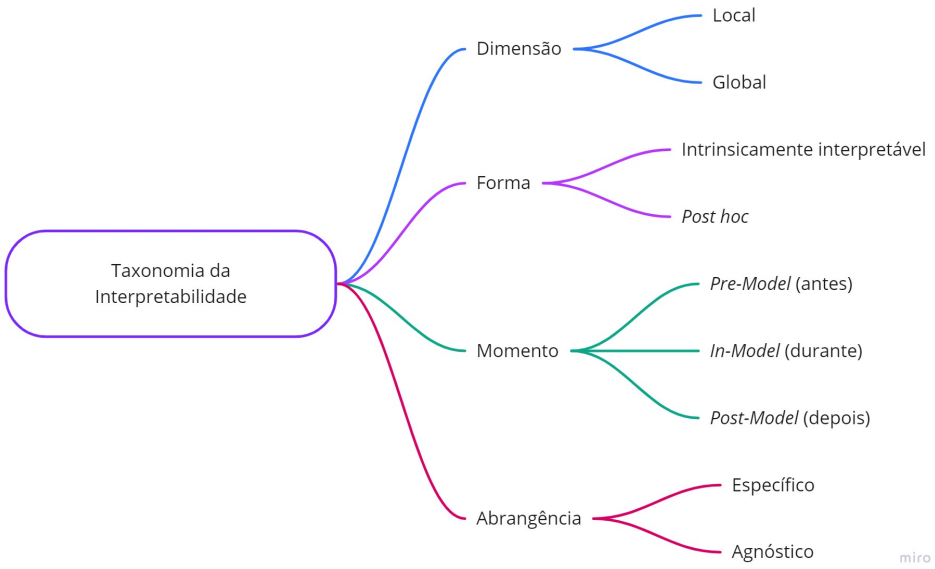
Quanto a sua abrangência os modelos podem ser classificados em específicos ou agnóstico. Um método é chamado de específico quando se restringe-se a uma classe de modelos em particular. Os métodos específicos geralmente geram as explicações examinando estruturas e parâmetros internos do modelo (DU; LIU; HU, 2019; MOLNAR, 2020).

Por outro lado, métodos chamados de agnósticos são independentes de modelo, ou seja, podem ser aplicados a qualquer modelo de aprendizado de máquina. Desta forma, esses métodos não têm acesso a informações relacionadas ao funcionamento interno do modelo de previsão, como pesos ou questões estruturais. Esses são ainda considerados *post hoc*, o que significa que são aplicados após o treinamento do modelo. No geral, a estratégia desses métodos baseia-se na análise dos valores dos atributos de cada par de entrada e saída. Além disso, como são aplicados após o treinamento, esses métodos não interferem no poder preditivo dos modelos (LIPTON, 2018; MOLNAR, 2020).

Du, Liu e Hu (2019) destacam duas vantagens dos métodos agnósticos que favorecem desenvolvimento da interpretabilidade baseada nessa categoria de explicação: 1) são muito abrangentes, uma vez que são independentes do modelo principal; 2) não prejudicam o desempenho da previsão, pois são aplicados após o treinamento do modelo. Por outro lado, os autores destacam que esse tipo de método também inclui alguns riscos, pois não se pode garantir que as explicações refletem fielmente o processo de tomada de decisão de um modelo.

Enfim, é possível perceber na literatura diferentes tipos de classificação para os métodos de interpretabilidade a partir de diferentes perspectivas. A Figura 2 sintetiza os tipos de classificação apresentados neste capítulo.

Figura 2 – Taxonomia da interpretabilidade.



Fonte: Elaborado pelo autor.

2.1.6 Ferramentas de interpretabilidade

Esta seção abrange algumas ferramentas de interpretabilidade local para dados tabulares, tal qual o enfoque do método proposto neste trabalho.

O primeiro método de explicação local para modelos caixa preta foi o ExplainD (*Explain Decision*) desenvolvido por Poulin et al. (2006). O *framework* foi especialmente projetado para modelos baseados em *Naive Bayes*, *Support Vector Machine* (SVM) e Regressão Linear, mas seu princípio é generalizável para qualquer modelo caixa preta. O ExplainD utiliza o conceito de modelos aditivos para ponderar a importância dos atributos de entrada (POULIN et al., 2006; GUIDOTTI et al., 2018). Este é o mesmo conceito que norteia o SHAP.

O ExplainD fornece 5 recursos gráficos de visualização das explicações. Poulin et al. (2006) ilustram tais recursos em um classificador de diagnóstico de doença arterial coronariana obstrutiva (*Coronary Artery Disease - CAD*). No exemplo apresentado, considerou-se um homem de 35 anos com CAD. Os referidos recursos gráficos incluem:

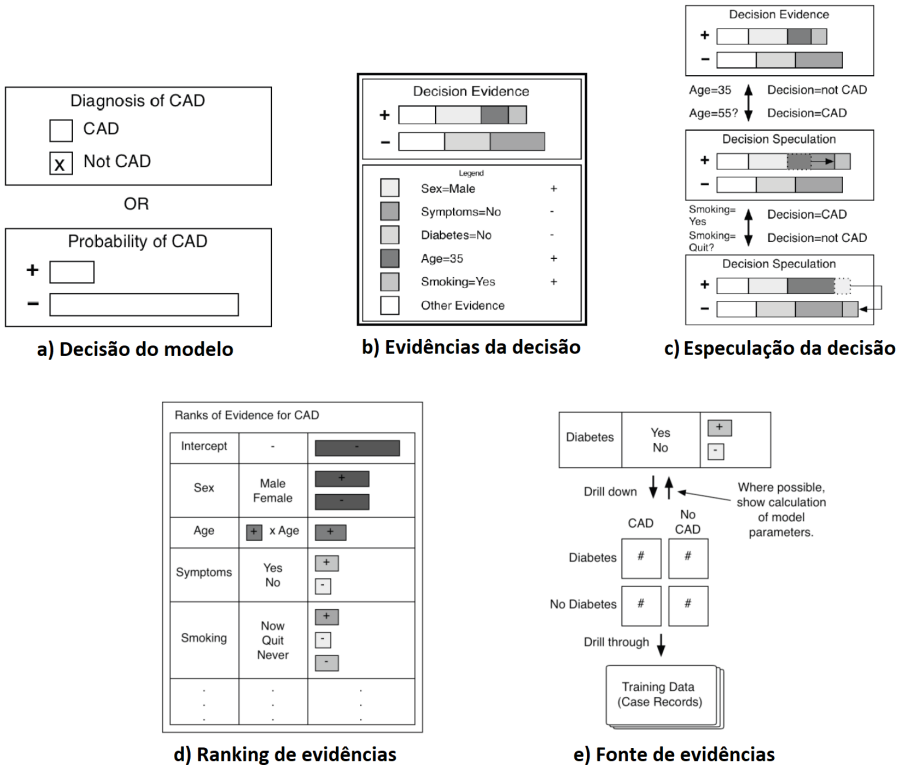
- Decisão: apresenta a classificação predita (Figura 3(a));
- Evidências da decisão: representa a contribuição relativa de cada atributo na decisão (Figura 3(b));
- Especulação da decisão: análise *"What-if?"* que permite especular sobre o efeito que uma mudança nos valores dos atributo teria na decisão (Figura 3(c));
- Ranking de evidências: representação visual que visa apresentar de forma ordenada

os atributos de maior efeito no contexto geral do classificador (Figura 3(d));

- Fonte de evidências: ajuda os usuários a explorar o raciocínio e os dados por trás do classificador. Este recurso visa auditar a relação entre as classes e os atributos. Basicamente, para cada atributo, os dados de treinamento são "fatiados" e um sumário dos dados mostra a quantidade de cada "fatia" em cada classe (Figura 3(e));

O recurso de especulação da decisão presente no ExplainD segue o princípio das explicações contrafactuais, embora no caso apresentado pelos autores, mostre-se como mudar a decisão e em seguida como voltar a decisão original (Figura 3(c)).

Figura 3 – Recursos gráficos de explicação do ExplainD.



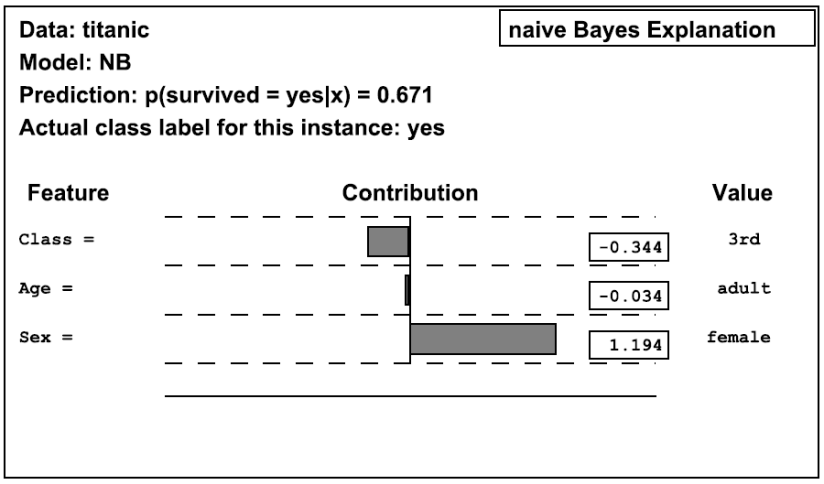
Fonte: Extraído de Poulin et al. (2006).

Strumbelj e Kononenko (2010) desenvolveram um método de explicações locais baseado na teoria dos jogos cooperativos (*coalitional game theory*). A explicação gerada consiste na contribuição dos atributos para a predição. A teoria dos jogos cooperativos é usado justamente no cálculo da contribuição de cada atributo. A intenção é encontrar um valor "justo" para a contribuição de cada atributo. A noção de "justiça" é apoiada no *Shapley value* cujo cálculo é apresentado na Seção 2.2, página 33.

A Figura 4 ilustra a saída gerada pelo método de Strumbelj e Kononenko (2010)

para a base de dados Titanic* considerando uma instância predita como "sobrevivente" ao naufrágio.

Figura 4 – Exemplo de explicação gerada pelo método de Strumbelj e Kononenko (2010).



Fonte: Extraído de Strumbelj e Kononenko (2010).

O princípio utilizado pelo método de Strumbelj e Kononenko (2010) baseado na teoria dos jogos cooperativos e do *Shapley value* é semelhante a abordagem SHAP de Lundberg e Lee (2017) (Seção 2.2, página 33).

O Local Interpretable Model-agnostic Explanations (LIME) é uma ferramenta de interpretabilidade capaz de trabalhar com dados tabulares, texto ou imagem cuja abordagem consiste em treinar um modelo substituto local interpretável (por exemplo modelos lineares ou uma árvore de decisão) para explicar predições individuais (RIBEIRO; SINGH; GUESTRIN, 2016). O modelo substituto deve ser uma boa aproximação do modelo caixa preta localmente, mas não precisa ser uma boa aproximação global. Esse tipo de acurácia entre o modelo substituto e o modelo caixa preta é chamada de fidelidade local (MOLNAR, 2020).

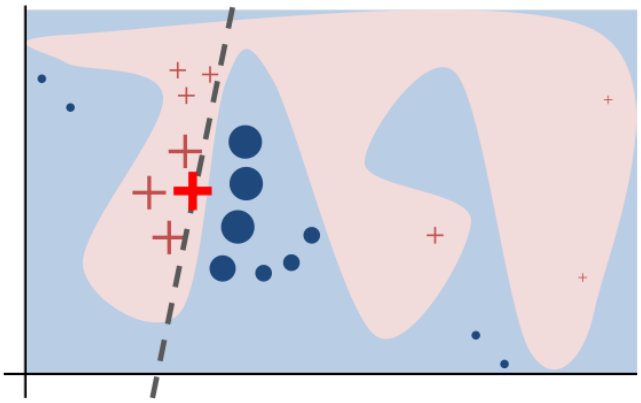
Para criar o modelo substituto local, o LIME gera um novo conjunto de dados, na vizinhança de uma instância original, por meio de perturbações nos valores dos atributos desta instância (RIBEIRO; SINGH; GUESTRIN, 2016). Segundo Ribeiro, Singh e Guestrin (2018), a estratégia de perturbação dos valores é a mais comumente aplicada pelos métodos agnóstico de explicação.

Ribeiro, Singh e Guestrin (2016) ilustram o princípio intuitivo do LIME por meio da Figura 5. O complexo modelo caixa-preta f (desconhecido para o LIME) está representado pelas cores de fundo azul e rosa. A instância que se deseja explicar está retratada pela cruz vermelha em destaque. Em torno desta estão as instâncias geradas pelo LIME,

*Disponível em `https://www.kaggle.com/c/titanic`.

cuja classificação é feita por f, com peso (representado pelo seu tamanho) definido pela proximidade em relação a instância a ser explicada. A linha tracejada representa a explicação que é localmente (mas não globalmente) fiel. Essa é uma interessante ilustração que auxilia no entendimento do problema por trás das explicações locais e globais de uma forma geral.

Figura 5 – Princípio intuitivo do LIME.



Fonte: Extraído de Ribeiro, Singh e Guestrin (2016).

Vale salientar que uma extensão do método (SP-LIME) propõe a compreensão global do modelo por meio da explicação das instâncias mais representativas da base de dados (RIBEIRO; SINGH; GUESTRIN, 2016).

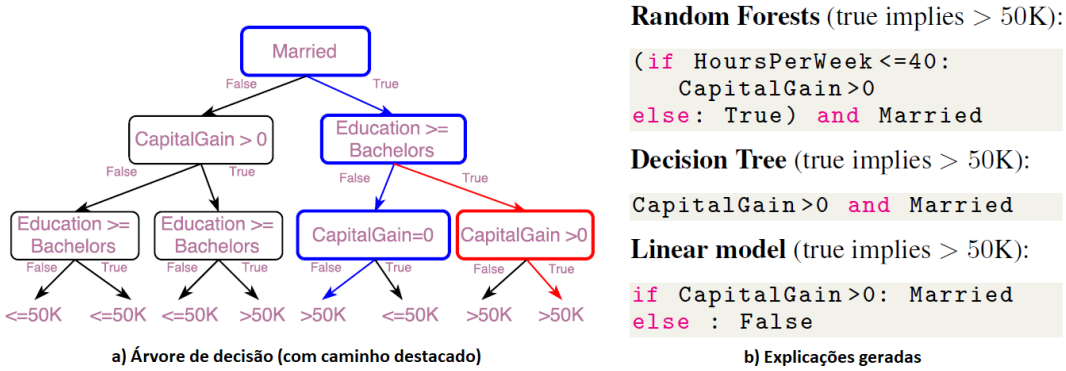
Singh, Ribeiro e Guestrin (2016) propõem o uso de *programas* para explicar o comportamento local de modelos caixa-preta. Os autores apresentam uma linguagem que inclui como recursos constantes booleanas (*True, False*), alguns operadores lógicos (*and, or, not*), a ausência/presença de determinados atributos na instância de entrada, constantes de valor numérico, operadores algébricos $(+, -, *)$, atributos de valor numérico e condições *if-then-else*. Os autores consideram a linguagem como bastante expressiva, mas reconhecem que existem limitações devido à falta de laços, recursão e variáveis. Logo, ainda não se trata de uma linguagem completa.

Para exemplificar a utilização da ferramenta, Singh, Ribeiro e Guestrin (2016) treinaram modelos para a base de dados *Adult*[†] baseados em três algoritmos: Árvore de Decisão, *Random Forest* e regressão logística. Em particular, no caso da árvore de decisão, é possível observar a representação convencional da mesma (Figura 6(a)) e o programa gerado para explicação dos caminhos em destaque (Figura 6(b)).

Os autores apontam uma série de vantagens no uso da explicação em forma de programas. Dentre elas, argumenta-se que:

[†]Trata-se de uma base de dados para previsão de renda baseado em dados do censo. Base disponível em `https://archive.ics.uci.edu/ml/datasets/adult`.

Figura 6 – Exemplo de explicação gerada pelo método de Singh, Ribeiro e Guestrin (2016).



Fonte: Extraído de Singh, Ribeiro e Guestrin (2016).

- As linguagens de programação são capazes de capturar comportamentos complexos com uma sintaxe de alto nível que é sucinta e intuitiva, e há um grupo crescente de usuários já treinados para manipulá-las;
- Qualquer representação interpretável existente na literatura pode ser escrita como um programa, mas além disso, os programas também podem representar combinações de múltiplas dessas representações;
- É possível trocar a expressividade e a compreensibilidade do programa, por exemplo programas simples para novos programadores (ao custo de ser uma aproximação do sistema complexo) ou programas detalhados e mais longos para uma explicação mais precisa do comportamento;
- Pode-se aproveitar pesquisar direcionadas para a análise de programas/softwares na avaliação de vários aspectos, como complexidade, segurança, privacidade e assim por diante.

A proposta do método Anchors, desenvolvido por Ribeiro, Singh e Guestrin (2018), baseia-se nas ideias de Ribeiro, Singh e Guestrin (2016) descrita anteriormente, ou seja, a explicação local é gerada por meio de perturbações em torno da instância que se deseja explicar.

O Anchors é um método agnóstico aplicável a dados tabulares, texto ou imagem baseado em regras do tipo "se-então". A explicação gerada pelo Anchors é uma regra que suficientemente "ancora" a previsão localmente, de forma que mudanças nos valores dos demais atributos da instância não importam. Em outras palavras, instâncias cobertas pela mesma "âncora" (quase) sempre terão a mesma predição (RIBEIRO; SINGH; GUESTRIN, 2018).

A Figura 7 mostra exemplos de aplicação do Anchors em modelos baseados em

Gradient Boosted Trees em três base de dados públicas (*Adult*, *RCDV*[‡] e *Lending*[§]). Uma explicação "âncora" de cada modelo é apresentada na figura (RIBEIRO; SINGH; GUESTRIN, 2018).

Figura 7 – Exemplo de explicação gerada pelo Anchors Ribeiro, Singh e Guestrin (2018).

	If	Predict
adult	No capital gain or loss, never married	$\leq$ 50K
	Country is US, married, work hours $>$ 45	$>$ 50K
rcdv	No priors, no prison violations and crime not against property	Not rearrested
	Male, black, 1 to 5 priors, not married, and crime not against property	Re-arrested
lending	FICO score $\leq$ 649	Bad Loan
	649 $\leq$ FICO score $\leq$ 699 and \$5,400 $\leq$ loan amount $\leq$ \$10,000	Good Loan

Fonte: Extraído de Ribeiro, Singh e Guestrin (2018).

A Tabela 1 apresenta informações relacionadas a disponibilidade dos métodos para os usuários como a possibilidade de acesso ao código fonte, se há pacote disponível e em que linguagem os métodos foram implementados. Para os métodos de Strumbelj e Kononenko (2010) e Singh, Ribeiro e Guestrin (2016) não se localizou o código fonte ou algum pacote para acesso ao método. Os autores do ExplainD afirmam disponibilizar a implementação do mesmo, porém, no endereço indicado no trabalho não foi possível localizar o código fonte. Os autores do LIME e Anchors disponibilizaram seu código fonte. Para esses métodos, existem ainda pacotes para Python por meio dos quais os usuários podem utilizá-los em implementações na mesma linguagem. Ressalta-se que o Anchors, LIME e a proposta de interpretabilidade baseada em programas foram desenvolvidos pelos mesmos autores.

Tabela 1 – Acesso aos métodos de interpretabilidade local com dados tabulares

Nome	Autor	Código fonte	Pacote disponível	Linguagem implementação
ExplainD	Poulin et al. (2006)	_	_	Python
_	Strumbelj e Kononenko (2010)	_	_	_
LIME	Ribeiro, Singh e Guestrin (2016)	`https://github.com/marcotcr/lime`	Sim	Python
_	Singh, Ribeiro e Guestrin (2016)	_	_	_
Anchors	Ribeiro, Singh e Guestrin (2018)	`https://github.com/marcotcr/anchor`	Sim	Python

Fonte: Elaborado pelo autor.

[‡]Base para previsão de risco de reincidência para indivíduos liberados da prisão. Base disponível em `https://www.icpsr.umich.edu/web/NACJD/studies/8987`.

[§]Base de dados para previsão de empréstimo baseado no *FICO credit score* (pontuação usada nas principais instituições de crédito dos EUA para avaliação de risco). Base disponível em `https://www.kaggle.com/wordsforthewise/lending-club`.

O SHAP também é um método de interpretabilidade local. Optamos por descrevê-lo em uma seção a parte (Seção 2.2, página 33) em função de sua importância para o trabalho. Métodos de explicações contrafactuais estão descritos nos Trabalhos Relacionados (Capítulo 3, página 38).

2.1.7 Explicações

A explicação é um modo pelo qual um observador pode obter compreensão. Sendo assim, uma forma de elevar o nível de interpretabilidade dos sistemas é prover explicações apropriadas aos usuários (MILLER, 2019).

No entanto, pesquisas apontam inadequações nos atuais métodos de interpretabilidade no que se refere a capacidade de comunicação com os usuários finais (KARIM et al., 2018; DU; LIU; HU, 2019; MILLER, 2019; MOLNAR, 2020).

Neste sentido, Miller (2019) afirma que se quisermos projetar e implementar agentes inteligentes que sejam realmente capazes de fornecer explicações às pessoas, é fundamental entender como os humanos definem, geram, selecionam, avaliam e apresentam explicações. No entanto, o autor destaca que a maioria das pesquisas e práticas na área parecem usar as intuições dos pesquisadores do que se considera uma "boa" explicação. A solução está em expandir os estudos além das questões computacionais. O autor complementa que nos campos da filosofia, psicologia/ciências cognitivas e psicologia social há um vasto e maduro corpo de trabalhos que estudam exatamente esses tópicos.

Diante disso, embasado em estudos que consideram aspectos computacionais e das ciências sociais, Miller (2019) lista três pontos que deveriam ser considerados para se construir uma IA verdadeiramente explicável: explicações contrastivas, selecionadas e sociais.

2.1.7.1 Tipos de explicação

O tipo de explicação produzida pode ser considerado um outro critério para diferenciar métodos de explicação. Os tipos de explicação mais comuns são:

- *Sumário dos atributos*: alguns métodos geram explicações por meio de estatísticas relacionadas a cada atributo. Nesse tipo de explicação é comum apresentar os atributos mais importantes, segundo algum critério definido pelo método, destacando a relevância dos atributos no processo preditivo do modelo principal (CARVALHO; PEREIRA; CARDOSO, 2019).

  O método agnóstico e global de importância dos atributos trata o modelo original como uma caixa preta de maneira que não inspeciona parâmetros internos do

mesmo. A abordagem geral é obter o quão relevante é o atributo para as previsões do modelo por meio da permutação dos seus valores. Em linhas gerais, usa-se a seguinte estratégia: dado um modelo treinado e um conjunto de teste têm-se um desempenho médio p, que é considerada a acurácia base. Permuta-se os valores de um dado atributo no conjunto de teste e mede-se o valor do desempenho médio com os dados modificados. Repete-se o processo iterativamente para cada um dos n atributos obtendo-se n pontuações de desempenho. Os atributos são ranqueados de acordo com a redução de seu desempenho em relação a acurácia base p. Em outras palavras, se a mudança dos valores de um dado atributo tem alta interferência no desempenho médio do modelo, isso significa que este atributo tem grande importância na previsão (DU; LIU; HU, 2019).

Para as explicações *post hoc* locais, o método de importância dos atributos também pode ser utilizado. Neste caso, as explicações locais têm como objetivo identificar as contribuições de cada atributo de entrada em relação a uma previsão específica. Como métodos locais geralmente atribuem a decisão de um modelo aos atributos de entrada, eles também são chamados de métodos de atribuição (DU; LIU; HU, 2019);

- *Ponto de dados*: também chamados de métodos baseados em exemplos, esses retornam pontos de dados (já existentes ou não) para criar um modelo interpretável. É necessário que os pontos de dados selecionados pelo método sejam significativos e possam ser interpretados. A abordagem contrafactual enquadra-se neste tipo de método (MOLNAR, 2020; CARVALHO; PEREIRA; CARDOSO, 2019);

- *Modelo substituto intrinsecamente interpretável*: a ideia é gerar modelos intrinsecamente interpretáveis que representem uma aproximação global ou local do modelo de aprendizado. Assim, a interpretação do modelo substituto fornecerá *insights* do modelo original (MOLNAR, 2020; CARVALHO; PEREIRA; CARDOSO, 2019);

- *Modelos internos*: a interpretação dos modelos intrinsecamente interpretáveis se enquadra nesta categoria. Os resultados podem ser, por exemplo, pesos utilizados internamente em modelos lineares ou a estrutura de árvore aprendida (os atributos e limites usados para as divisões). Esses também são considerados, por definição, métodos de interpretabilidade específicos (MOLNAR, 2020).

Os métodos de interpretabilidade podem utilizar mais de uma das estratégias acima. Por exemplo, Du, Liu e Hu (2019) apresentam a explicação local baseada em aproximação. Esse baseia-se na ideia de que previsões de aprendizado de máquina na vizinhança de uma determinada entrada pode ser aproximada por um modelo de caixa branca interpretável, ou seja, um modelo substituto. O modelo interpretável pode não

funcionar bem globalmente, mas deve aproximar-se do modelo de caixa preta na vizinhança da entrada original. Em seguida, obtém-se a pontuação da contribuição de cada característica examinando os parâmetros do modelo de caixa branca.

2.1.7.2 Explicações contrafactuais

Teorias contrastivas argumentam que explicações causais inevitavelmente geram interesse em um caso contrafactual, ou seja, uma causa ou evento que não ocorreu (MITTELSTADT; RUSSELL; WACHTER, 2019). Especialmente quando a previsão refere-se a uma classe indesejável, é natural que o indivíduo afetado deseje saber como fazer para mudar essa previsão, o que significa entender "Por que P ao invés de Q" (em que P é o evento alvo e Q é um caso de contraste contrafactual que não ocorreu ou ainda "Por que [classe predita] e não [classe desejada]?").

Segundo Miller (2019), *everyday contrastive explanations* são mais eficientes na busca pela confiança nas decisões de um modelo, pois as pessoas compreendem melhor explicações geradas a partir de fatos específicos (eventos, propriedades, decisões, etc.), ao invés de tentar entender e construir teorias generalizadas.

No AM moderno, o comportamento do algoritmo pode consistir de um grande número de variáveis intrinsecamente conectadas. Diante disso, estabelecer uma linha de explicação que procure transmitir a um leigo esse comportamento para que o mesmo possa raciocinar sobre o funcionamento do algoritmo é uma tarefa extremamente difícil. Essa é uma grande vantagem das explicações contrafactuais, pois é necessário apenas entender o que é diferente entre os dois casos (WACHTER; MITTELSTADT; RUSSELL, 2017).

Wachter, Mittelstadt e Russell (2017) destacam ainda a contribuição das explicações contrafactuais em relação ao "direito de explicação" presente no GDPR. Outras formas de explicação com um caráter mais técnico podem ter pouco valor prático para os titulares dos dados. Por outro lado, explicações contrafactuais contornam o desafio de explicar o funcionamento interno de sistemas complexos de AM, pois fornecem informações que são mais facilmente compreensíveis e na prática mais úteis para entender os motivos de uma decisão, contestá-los e alterar o comportamento futuro para um melhor resultado.

Mittelstadt, Russell e Wachter (2019) observam uma questão preocupante em relação aos métodos contrastivos relacionado ao fato desses retornarem um único ponto de dados. Como muitas vezes existem múltiplas explicações, se o ponto a ser apresentado não corresponde a um factóide de interesse para o usuário, esse não pode ser usado para deduzir conclusões relevantes e neste caso não será útil para o mesmo.

2.1.7.3 Explicações selecionadas e sociais

Já em relação aos aspectos seletivo e social das explicações, Hanson (1965) apresenta a seguinte narrativa que exemplifica a importâncias de tais elementos. Considere o caso de um acidente automobilístico no qual uma pessoa morre. Caso se peça a um médico para explicar a morte ele dirá “múltiplas hemorragias”, um advogado diria “negligência do motorista”, um engenheiro automotivo explica que “o freio não foi capaz de parar a tempo”, um engenheiro civil poderia argumentar “a presença de arbustos altos naquela curva". Desta forma, o autor mostra que mesmo que todas explicações sejam verdadeiras, há um contexto particular que pode tornar uma mais relevante do que outra para cada interessado.

Tipicamente em um determinado evento múltiplas explicações são possíveis atribuindo diferentes causas. Logo, é natural que cada destinatário se interesse pela explicação que considera mais útil para determinado contexto ou propósito (MITTELSTADT; RUSSELL; WACHTER, 2019).

Mittelstadt, Russell e Wachter (2019) alertam ainda que na XAI é comum que as explicações sejam escolhidas com base em atributos chaves ou pesos que influenciaram na predição. No entanto, a relevância dos atributos não deveria considerar apenas pesos estatísticos, mas também os interesses e expectativas do receptor da explicação.

Além disso, raramente as pessoas estão interessadas na cadeia causal completa de um evento. Normalmente as pessoas irão preferir a explicação que transmitir informações úteis para um determinado propósito. Certos vieses cognitivos também podem influenciar na preferência do receptor. Tais aspectos evidenciam a seletividade humana quanto as explicações (MILLER, 2019). Logo, é fundamental estabelecer uma comunicação com o usuário e permitir que interesses e demandas particulares sejam considerados nas explicações.

Quanto ao aspecto social da explicação, este implica em estabelecer um processo interativo entre o explicador e o destinatário da explicação e a transferência de conhecimento. Sendo assim, é necessário que a informação seja adequada às crenças e capacidades de compreensão do destinatário (MILLER, 2019).

Segundo Mittelstadt, Russell e Wachter (2019), no caso dos modelos de AM, essa interação envolve uma mistura de atores humanos (desenvolvedor e o usuário) e automatizados (modelo ou sistema). Para os autores, as explicações são processos iterativos na medida em que devem ser selecionadas e avaliadas com base em pressuposições e crenças compartilhadas. A iteração é essencial para comunicar de forma eficaz aspectos que permitam gerar explicações relevantes para o usuário.

É importante ressaltar que, ao propiciar seletividade e contextualização nas expli-

cações, contribui-se com sua socialização. Primeiro, porque a seletividade é uma forma de estabelecer interação entre o explicador e o receptor. Segundo, não há transferência de conhecimento se a explicação não têm sentido ou relevância para o receptor. Além disso, a própria explicação contrafactual, como resposta para uma *why question* específica, tem um formato que suporta a ideia de explicação como uma conversa.

Enfim, dado que existe mais de uma possível explicação para um determinado evento e cada destinatário pode ter diferentes expectativas e interesses, é importante prover uma forma de comunicação entre métodos de interpretabilidade e seus usuários para que cada contexto e propósito possa ser considerado. Assim, os métodos poderão gerar explicações relevantes e adequadas às capacidades de cada usuário, proporcionando compreensão e conhecimento.

2.2 SHAP

SHAP é uma abordagem proposta por Lundberg e Lee (2017) baseada na teoria dos jogos cooperativos para interpretar modelos de AM. Para gerar o modelo de explicação, o SHAP usa um *additive feature attribution method*, isto é, a saída do modelo é definida como uma adição linear dos atributos de entrada. Os efeitos dos atributos são aditivos, o que significa que não há interações, e a relação é linear, o que significa que um aumento de um atributo em uma unidade pode ser diretamente traduzido em um aumento/diminuição do resultado previsto. O modelo linear nos permite comprimir a relação entre um atributo e o resultado esperado em um único número, ou seja, o peso (efeito ou impacto) estimado (MOLNAR, 2020). Tais métodos se enquadram na classe *post hoc*.

Dado um vetor de atributos $x \in \mathbb{R}^M$, sendo M o número de atributos, $f(x_i)$ representa a predição da *i-ésima* instância com o vetor x_i. A interpretabilidade é dada pela Equação 2.1.

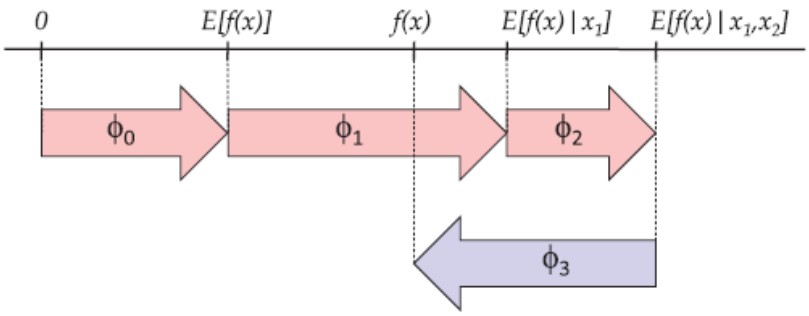
$$f(x_i) = E[f(x)] + \sum_{j=1}^{M} \phi_{i,j} \tag{2.1}$$

sendo $E[f(x)]$ a previsão esperada com base em todo o conjunto de treinamento (valor base) e $\phi_{i,j}$ corresponde ao impacto do j-ésimo atributo para i-ésima instância, que pode ser positivo, negativo ou zero. Assim, $\phi_{i,j}$ representa a extensão em que a j-ésimo atributo distancia a previsão da i-ésima instância do valor médio (ZENG; DAVOODI; TOPALOGLU, 2020). Outra forma de entender o valor base $E[f(z)]$ é percebê-lo como a previsão do modelo, caso não se conhecesse nenhum atributo para uma dada saída $f(x)$ (LUNDBERG; LEE, 2017).

A soma dos efeitos de todos atributos corresponde a uma aproximação da predição

$f(x)$ que se deseja explicar. A Equação 2.1 está ilustrada na Figura 8, onde ϕ_0, ϕ_1, ϕ_2 incrementam o valor da predição, enquanto ϕ_3 decrementa o mesmo valor (LUNDBERG; LEE, 2017; MANGALATHU; HWANG; JEON, 2020).

Figura 8 – Visão geral da abordagem SHAP.



Fonte: Extraído de Mokhtari, Higdon e Başar (2019).

O SHAP *value* (LUNDBERG; LEE, 2017) corresponde a uma forma para calcular os $\phi_{i,j}$s e está, portanto, relacionado à importância dos atributos. O cálculo desses valores requer retreinar o modelo para todos os subconjuntos de atributos $S \subseteq F$, onde F é o conjunto de todos atributos. O valor de importância calculado para cada atributo representa o efeito da presença desse atributo no modelo de previsão. Para calcular o efeito de um atributo j, um modelo $f_{S\cup\{j\}}$ é treinado com a presença de j e outro modelo f_S é treinado sem o mesmo atributo. A diferença das predições dos modelos $f_(x_{S\cup\{j\}}) - f(x_S)$ para uma específica entrada x_S mostra o efeito do atributo j na previsão. Uma vez que o efeito de um atributo j depende dos demais atributos, as diferenças são computadas para todos possíveis subconjuntos $S \subseteq F\backslash\{j\}$. A Equação 2.2 calcula o SHAP *value* como uma média ponderada de todas as possíveis diferenças (MOKHTARI; HIGDON; BAsAR, 2019; ZENG; DAVOODI; TOPALOGLU, 2020):

$$\phi_{i,j} = \sum_{S\subseteq F\backslash\{j\}} \{\frac{|S|!(|F|-|S|-1)!}{|F|!}\{E[f(x_i) \mid x_{i,S\cup\{j\}}] - E[f(x_i) \mid x_{i,S}]\}\} \tag{2.2}$$

Lundberg e Lee (2017) afirmam que a computação exata dos *SHAP values* é um desafio devido a alta carga computacional, já que o número de termos a serem computados cresce exponencialmente com o número de atributos. Desta forma, as implementações do SHAP reduzem essa carga de retreinar o modelo em todas as combinações de S por meio de métodos de aproximação.

Enquanto o *Kernel SHAP* representa um método de aproximação aplicável a qualquer modelo, os autores criaram métodos específicos para determinados tipos como o *Linear SHAP*, *Low-Order SHAP*, *Max SHAP*, *Deep SHAP* e o *SHAP Tree Explainer*.

Por exemplo, o *SHAP Tree Explainer* sugere que a computação exata dos SHAP *values* podem ser feitos em tempo polinomial para modelos baseados em árvore, explorando as informações armazenadas na estrutura da árvore (ZENG; DAVOODI; TOPALOGLU, 2020).

2.3 Algoritmos genéticos

Algoritmo genético é uma técnica heurística de otimização e busca baseada nos princípios da genética e nos mecanismos de seleção natural (HAUPT; HAUPT, 2004).

Segundo Linden (2005), sempre que houver uma necessidade de otimização, um AG pode ser considerado e, sendo assim, sua aplicabilidade é quase infinita. Logo, a questão consiste em tornar o problema em um formato tratável pelos AGs.

AG é uma técnica heurística de otimização global. Sendo assim, estes se opõem aos métodos de gradiente (hill climbing), que seguem a derivada de uma função ficando facilmente retidos em máximos locais (LINDEN, 2008).

selecao é considerado operador genetico? A proposta dos AGs é trabalhar uma população composta de muitos indivíduos para evoluir segundo um critério definido pela função *fitness* (aptidão ou função de avaliação). A população do AG é representada por indivíduos (cromossomos) os quais são submetidos a operadores genéticos de seleção, *crossover* (cruzamento) e mutação ao longo das gerações. Os indivíduos são avaliados pela função *fitness* e conduzidos para as próximas gerações seguindo o princípio da seleção natural (HAUPT; HAUPT, 2004). O Algoritmo 1 apresenta o pseudocódigo do AG básico proposto por Linden (2005).

Algoritmo 1: Basic GA Pseudocode. Extraído de Linden (2005).

```
1  t ← 0;
2  InitializePopulation(P(0));
3  while not Stopping Criteria do
4  |  Fitness(P(t));
5  |  parents ← Selection(P(t));
6  |  children ← Crossover_and_Mutation(parents);
7  |  P(t + 1) ← NewPopulation(children);
8  |  t ← t + 1;
9  end
10 best ← GetBest(P);
```

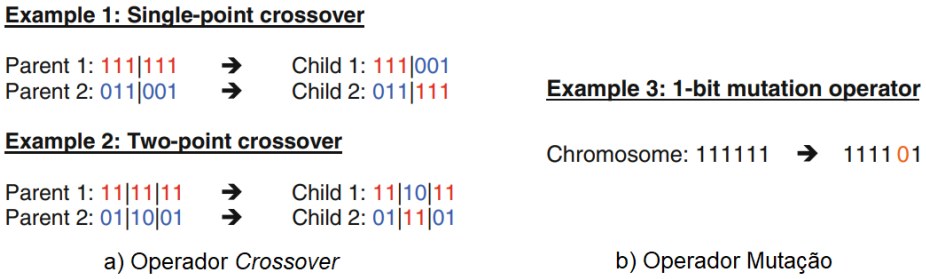
Na etapa de seleção, uma técnica deverá ser aplicada para escolher os indivíduos que irão assumir o papel de pais. Uma das possibilidades é a seleção por roleta viciada na qual a probabilidade de um indivíduo ser escolhido é proporcional a sua aptidão. Essa forma de seleção tende a privilegiar os indivíduos com função de avaliação mais alta sem

desprezar aqueles com avaliação mais baixa. Outra possibilidade é utilizar a seleção por torneio no qual a população pode ser dividida em subpopulações menores com membros de cada subpopulação competindo entre si, escolhendo apenas um indivíduo de cada subpopulação para reprodução (LINDEN, 2005; LIM, 2014).

Os indivíduos escolhidos na etapa de seleção são submetidos aos operadores genéticos de cruzamento e mutação. Para o cruzamento, dois indivíduos são selecionados e por meio de segmentos trocados de seus códigos são gerados os filhos que carregam informações parciais de ambos os pais. Na literatura dos AGs, existem alguns procedimentos de cruzamento, sendo os mais comuns o cruzamento de um ponto (*single-point crossover*) e o cruzamento de dois pontos (*two-point crossover*). No cruzamento de um ponto, um ponto de corte é escolhido aleatoriamente para dividir em partes o conjunto de genes de modo que as partes separadas sejam trocadas gerando os descendentes. No cruzamento de dois pontos, como o nome indica, dois pontos de corte são aleatoriamente selecionados e o conjunto de genes entre eles são trocados no par de indivíduos selecionados. Os dois tipos de procedimentos de cruzamento estão ilustrados na Figura 9(a) (LINDEN, 2005; LIM, 2014).

O cruzamento garante a hereditariedade e troca de informações dos pais para os filhos, porém, não induz novas variações genéticas, o que pode ser um problema especialmente quando a população se torna homogênea. Neste caso, é importante a utilização do operador de mutação, o qual permite que uma mudança de frequência genética normalmente lenta e pequena ocorra dentro a população. A mutação opera em um único indivíduo (Figura 9(b)).

Figura 9 – Operadores genéticos do AG.



Fonte: Extraído de Lim (2014).

Em geral nos AGs, a cada geração, todos os indivíduos da população atual são removidos e novos indivíduos são criados para a próxima geração usando a reprodução sobre a população atual. Ao fazer isso, podemos perder os melhores indivíduos de uma geração devido à natureza estocástica dos AGs. Diante disso, uma possível alteração na criação da nova população é considerar o uso do elitismo. Este consiste em preservar de uma geração para a outra os n cromossomos não redundantes melhor avaliados. O número

de indivíduos a preservar por elitismo pode ser 10%, 20% ou 50% de toda a população, dependendo do comportamento observado ao longo dos testes do algoritmo. O elitismo pode minimizar o custo de encontrar soluções ótimas em um número menor de iterações (RANI; SURI; GOYAL, 2019).

3 TRABALHOS RELACIONADOS

Wachter, Mittelstadt e Russell (2017) apresentam o SOC para geração de explicações contrafactuais aplicadas a modelos baseados em redes neurais para dois problemas. A técnica utilizada para geração dos contrafactuais foi a "*Adversarial Perturbations*". Resumidamente, a ideia é usar algoritmos capazes de computar contrafactuais para "confundir" classificadores existentes, gerando um ponto de dados sintético próximo a um existente, de modo que o novo ponto de dados sintético seja classificado de forma diferente do original.

Wachter, Mittelstadt e Russell (2017) demonstram interesse em discutir como as explicações contrafactuais podem contribuir diante das restrições impostas pela GDPR à tomada de decisão automatizada, especialmente em relação a importância das decisões serem justas. Segundo os autores, os contrafactuais podem fornecer evidências de que uma decisão algorítmica é afetada por um atributo protegido (por exemplo, raça, etnia ou gênero) e que, portanto, pode ser discriminatória.

O SOC foi testado em dois problemas. O primeiro foi um modelo de predição de desempenho acadêmico em uma faculdade de direito. Segundo os autores, a base de dados que deu origem ao modelo é comumente usada na literatura de justiça. O segundo problema consiste em modelo para classificação de risco de diabetes. Neste caso, a escolha se deu para avaliar o comportamento do método em um problema de maior complexidade. A Figura 10 ilustra a saída gerada para o método em um exemplo da base de predição de desempenho acadêmico (WACHTER; MITTELSTADT; RUSSELL, 2017).

Figura 10 – Exemplo de explicação gerada pelo SOC.

Person 1: If your LSAT was 34.0, you would have an average predicted score (0).
Person 2: If your LSAT was 32.4, you would have an average predicted score (0).
Person 3: If your LSAT was 33.5, and you were 'white', you would have an average predicted score (0).
Person 4: If your LSAT was 35.8, and you were 'white', you would have an average predicted score (0).
Person 5: If your LSAT was 34.9, you would have an average predicted score (0).

Fonte: Extraído de Wachter, Mittelstadt e Russell (2017).

Wachter, Mittelstadt e Russell (2017) concluíram que as explicações contrafactuais podem preencher a lacuna entre os interesses dos titulares e dos controladores dos dados,

atendem a necessidade do direito a explicação e, desta forma, aumenta potencialmente a aceitação pública de decisões automáticas.

Dhurandhar et al. (2018) propuseram um método de explicações contrastivas específico para redes neurais que explora o conceito de pertinentes positivos e negativos. Um pertinente positivo é um fator cuja presença é minimamente suficiente para justificar a classificação final. Já pertinente negativo é um fator cuja ausência é necessária para afirmar a classificação final.

Os autores apresentam resultados experimentais em três bases de dados: 1) uma base de dados de dígitos manuscritos MNIST; 2) uma base de dados de fraude de compras; 3) uma base de dados de imagens de ressonância magnética cerebral contendo padrões de atividade cerebral de pessoas autistas e sem autismo (DHURANDHAR et al., 2018).

Dhurandhar et al. (2018) afirmam que o método mostrou-se efetivo em diferentes domínios gerando explicações presumivelmente mais fáceis de compreender e mais precisas. Os autores destacam que a identificação de pertinentes negativos são particularmente úteis quando entradas próximas podem gerar classificações diferentes (por exemplo para distinguir um diagnóstico de gripe ou pneumonia). Se as entradas forem muito diferentes, provavelmente os pertinentes positivos são suficientes para caracterizar a entrada, pois possivelmente irão existir muitos pertinentes negativos sobrecarregando o usuário.

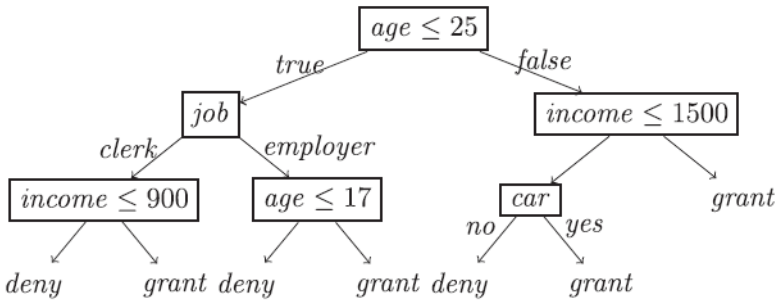
Guidotti et al. (2019) propõem um método de explicação chamado LORE, baseado em regras, e inclui a abordagem factual e contrafactual. O método é agnóstico e voltado para dados tabulares. O LORE aprende um classificador local interpretável (uma árvore de decisão) por meio de um conjunto de instância vizinhas à instância que se deseja explicar. As instâncias vizinhas são geradas por um AG ad-hoc e usadas como dados de treinamento para o aprendizado da árvore de decisão local. Por meio deste classificador, o método gera uma regra que explica as razões factuais da decisão e um conjunto de regras contrafactuais que indicam as mudanças na instância original que implicariam em uma saída diferente.

Guidotti et al. (2019) destacam as duas principais contribuições do trabalho. Primeiro, a elevada expressividade da explicação proposta supera os métodos existentes, pois fornecem não apenas evidências sobre porque uma instância recebeu um rótulo específico, mas também contrafactuais sugerindo o que deve ser diferente nas proximidades desta instância para reverter o resultado previsto. Segundo, a fronteira da decisão local, na vizinhança da instância a ser explicada, é explorada por meio de um algoritmo genético capaz de produzir dados de treinamento de alta qualidade para aprender a árvore de decisão local.

Para avaliação das regras factuais, Guidotti et al. (2019) realizam experimentos comparando seu método com o LIME e o Anchors (métodos apresentados na Seção 2.1.6,

página 23). Nestes os autores utilizaram as bases Compas*, German† e Adult‡ e desenvolveu modelos baseados em *Random Forest*, Redes Neurais e SVM. Já no caso do contrafactuais, os autores comparam o desempenho do método LORE com o SOC de Wachter, Mittelstadt e Russell (2017), descrito nesta seção. Neste caso, as bases Compas e German foram exploradas. A Figura 11 ilustra uma árvore gerada pelo método para a base Adult. Por meio da árvore as regras factuais e contrafactuais são extraídas.

Figura 11 – Árvore de decisão gerada pelo LORE que imita o comportamento local de um modelo caixa preta.



Fonte: Extraído de Guidotti et al. (2019).

Guidotti et al. (2019) afirmam que os resultados experimentais mostram que o LORE supera as abordagens existentes em termos de qualidade das explicações e da acurácia em imitar a caixa preta.

Rathi (2019) apresenta um método agnóstico de explicação contrafactual apoiado no SHAP. O autor também evidencia a compatibilidade das explicações contrafactuais com a GDPR. A ideia básica é a partir de uma determinada instância, obter por meio do SHAP os atributos que impactam negativamente, ou seja, aquelas que afastam essa instância da classe contrafactual para o qual se deseja migrar. O método tenta gerar instâncias contrafactuais, no qual cada possível contrafactual resulta da troca dos valores somente dos atributos de impacto negativo, utilizando como base um ponto vizinho a instância original.

O método foi testado em três bases de dados: IRIS, *Wine Quality*§ e *Mobile Price Classification*¶. Para cada uma destas bases de dados foram desenvolvidos modelos baseados em *K-Nearest Neighbors* (KNN), Redes Neurais, Random Forest e SVM. A Figura 12 exemplifica uma saída gerada para a base IRIS (RATHI, 2019).

Rathi (2019) afirma que os critérios de avaliação utilizados se diferenciam de outras

*Disponível em `https://github.com/propublica/compas-analysis`
†Disponível em `https://archive.ics.uci.edu/ml/datasets/statlog+(german+credit+data)`
‡Disponível em `https://archive.ics.uci.edu/ml/datasets/adult`
§Disponível em `https://www.kaggle.com/yasserh/wine-quality-dataset`
¶Disponível em `https://www.kaggle.com/iabhishekofficial/mobile-price-classification`

Figura 12 – Exemplo de saída gerada pelo método de Rathi (2019).

Original Datapoint	[4.4, 2.9, 1.4, 0.2]
Counterfactual points	[4.4, 2.9, 2.4, 0.2], [4.4, 2.9, 3.0, 0.2], [4.4, 2.9, 3.3, 0.2], [4.4, 2.9, 3.5, 0.2], [4.4, 2.9, 3.7, 0.2], [4.4, 2.9, 3.8, 0.2], [4.4, 2.9, 3.9, 0.2], [4.4, 2.9, 4.0, 0.2]
Why 0?	Algorithms Pro classification was primarily influenced by petal width (cm)
Why not 1?	Algorithms Anti classification was primarily influenced by petal length (cm)

Fonte: Extraído de Rathi (2019).

pesquisas na área. O autor mede a eficácia do método pelo número total de novos pontos de dados contrafactuais (pontos que não estão na base de dados) e o número médio de contrafactuais gerados por instância da base original. Segundo o autor, esta é uma avaliação mais adequada para a abordagem, pois leva em conta o quanto a distribuição de dados atual está longe da fronteira da decisão.

Com relação ao desempenho do método, Rathi (2019) considera os resultados satisfatórios, pois observou-se que a maioria dos pontos contrafactuais gerados não estão presentes na base de dados, ou seja, esses não teriam sido alcançados ao pesquisar no espaço vizinho a instância de interesse na base original. A abordagem alcançou os melhores resultados nas bases de dados mais densas (com mais atributos e instâncias), como a *Wine Quality* e *Mobile Price Classification*, pois nessas obteve menos pontos em comum considerando os contrafactuais gerados e os dados da base original.

Mothilal, Sharma e Tan (2020) desenvolveram um método que enfatiza a diversidade de exemplos e proximidade com a instância original. Os autores demonstram preocupação em gerar explicações viáveis ao usuário obtidas por meio da diversidade das explicações. A ideia é que exemplos contrafactuais diversos aumentam as chances de que pelo menos um ser aplicável para o usuário.

Para gerar explicações diversas, Mothilal, Sharma e Tan (2020) desenvolveram o DiverseCF[‖]. A ideia é a partir de um modelo de AM treinado (f) e uma instância original (x) gerar k exemplos contrafactuais $c_1, c_2, ..., c_k$ que conduzam a saída desejada y. O método utiliza uma função de perda que combina todos os contrafactuais gerados. A função é otimizada utilizando gradiente descendente. O método busca alcançar $f(c_i) = y$

[‖]Disponível em `https://github.com/microsoft/DiCE`

para cada exemplo contrafactual. Porém, o objetivo pode não ser alcançado em alguns casos.

Para avaliar o DiverseCF, Mothilal, Sharma e Tan (2020) realizaram experimentos com as bases de dados Adult, LendingClub**, German e Compas. Os autores compararam o desempenho do DiverseCF com o LIME em um modelo baseado em Redes Neurais. A Figura 13 exemplifica uma saída do DiverseCF para a base Compas.

Figura 13 – Exemplo de saída gerada pelo DiverseCF.

COMPAS	PriorsCount	CrimeDegree	Race	Age	Sex
Original input (outcome: Will Recidivate)	10.0	Felony	African-American	>45	Female
Counterfactuals (outcome: Won't Recidivate)	–	–	Caucasian	–	–
	0.0	–	–	–	Male
	0.0	–	Hispanic	–	–
	9.0	Misdemeanor	–	–	–

Fonte: Extraído de Mothilal, Sharma e Tan (2020).

No geral, os resultados mostraram que os exemplos gerados pelo DiverseCF podem aproximar-se da fronteira de decisão local pelo menos tão bem quanto o LIME. Uma limitação presente no método diz respeito ao fato deste necessitar conhecer o gradiente do modelo de aprendizado. Os autores ressaltam a importância de se construir métodos que possam ser utilizados com diferentes modelos caixa preta (MOTHILAL; SHARMA; TAN, 2020).

Embora os trabalhos aqui apresentados se assemelhem ao ECOSS na medida que abordam explicações contrafactuais, as propostas apresentam algumas diferenças. Os trabalhos possuem direcionamentos diferentes como, por exemplo, o fato de Mothilal, Sharma e Tan (2020) estarem focado na diversidade dos contrafactuais gerados, Guidotti et al. (2019) proporem uma saída baseada em regras e Dhurandhar et al. (2018) terem suas explicações centradas no conceito de pertinentes positivos e negativos.

O ECOSS é semelhante à proposta Rathi (2019) uma vez que ambos exploram recursos presentes no SHAP. No entanto, o papel do SHAP nos trabalhos é diferente. O ECOSS usa o SHAP para avaliar a proximidade da instância contrafactual da classe desejada. Rathi (2019) emprega-o para identificar atributos de impacto negativo e alterá-los. Porém, não há garantia de que o contrafactual mais próximo é gerado pela modificação dos atributos de impacto negativo. A melhor solução pode ser justamente aumentar a contribuição de um atributo que já é positivo para a classe a ser migrada. Além disso, no método proposto por Rathi (2019), cada candidato a contrafactual realiza as trocas nos valores dos atributos usando um ponto de dados na vizinhança, enquanto a estra-

**Disponível em `https://www.kaggle.com/datasets/wordsforthewise/lending-club`

tégia proposta neste trabalho permite combinar trocas com valores de múltiplos pontos ampliando a busca pela melhor solução.

Um dos grandes diferenciais do ECOSS está na busca por explicações orientadas ao usuário final alinhando-se assim às orientações apresentadas por Miller (2019), focadas nas diretrizes das ciências sociais para maior aceitação, confiança e compreensão dos usuários quanto às explicações. O método inclui recursos de parametrização que permitem aproximar as explicações do contexto e interesses do usuário, como por exemplo, a restrição das possíveis caraterísticas a constar na explicação. Mais detalhes estão disponíveis no Capítulo 4, página 44.

Finalmente, o caráter agnóstico do ECOSS também não está presente em todos os métodos citados. Sendo assim, não foi encontrado um método que inclua todos os recursos presentes na proposta aqui apresentada. No Capítulo 6 (página 63), será apresentada ainda uma comparação de desempenho do ECOSS com os métodos propostos por Wachter, Mittelstadt e Russell (2017) e Guidotti et al. (2019).

4 DESCRIÇÃO DO MÉTODO ECOSS PROPOSTO

Em linhas gerais, o método ECOSS apresentado neste trabalho para geração de exemplos contrafactuais consiste em: dado um modelo de classificação f e uma instância x informados pelo usuário, tal que $f(x) = y$, gerar k exemplos contrafactuais $c_1, c_2, ..., c_k$ tão próximos quanto possível de x no qual $f(c_i) = y'$, para $y' \neq y$. A saída apresentada ao usuário consiste nas diferenças entre x e c_i, ou seja, as mudanças necessárias para inverter a classe de x. A aplicação do método restringe-se a dados tabulares e modelos cuja saída é binária.

Para encontrar os exemplos contrafactuais, o método faz uso de um AG ad-hoc, cuja função objetivo (Seção 4.1, página 44) combina dois critérios: proximidade da classe desejada e similaridade em relação à instância original. A estratégia é que, a cada geração, o AG busque contrafactuais que se aproximem da classe desejada (y'), mantendo a similaridade em relação a x. O SHAP é utilizado para avaliar se o candidato a contrafactual está se aproximando de y'. Os detalhes da implementação do método estão apresentados na Seção 4.2, página 46.

O ECOSS inclui elementos cuja finalidade é alinhar a saída gerada pelo método das diretrizes destacadas por Miller (2019) em relação a como as pessoas melhor recebem explicações. Tais recursos estão apresentados na Seção 4.3, página 50.

4.1 Função objetivo do AG

A abordagem proposta busca, por meio de AG, otimizar a similaridade dos exemplos contrafactuais gerados, considerando a restrição destes pertencerem à classe desejada (classe contrafactual). Para tanto, foi usada a função objetivo descrita na Equação 4.1.

$$C(x) = arg \min_{c} \lambda_1 \; x\_dist(c, x) + \; \lambda_2 \; x\_nchg(c, x) + \; \lambda_3 \; y\_dist(f(c), y') + \; Pe \quad (4.1)$$

x é a instância original, c representa o candidato a contrafactual de x, $f(c)$ é a resposta do modelo caixa preta para a instância contrafactual, y' representa a classe desejada, $x\_dist(c, x)$ é a função de distância entre x e o candidato a contrafactual gerado, $x\_nchg(c, x)$ é o número de atributos modificados em x para gerar c e $y\_dist(f(c), y')$ calcula a distância de c para a classe desejada y'. Pe representa uma penalização dada ao exemplo que ainda pertence a classe da instância original, no qual $Pe = 1$ se c está

na classe y ou $Pe = 0$ se c está na classe y'. Sendo assim, x_dist e x_nchg mantém o candidato a contrafactual c próximo à instância original e y_dist e Pe visam conduzir c em direção a classe desejada.

Os parâmetros λ_1, λ_2 e λ_3 atribuem pesos às partes da funçao que medem a distância para a instância original, a quantidade de alterações necessárias na instância original e a proximidade do candidato a contrafactual para classe desejada, respectivamente. Esses parâmetros podem ser ajustados para cada problema.

Para avaliar a similaridade de cada contrafactual em relação a x, o AG usa uma função de distância (euclidiana, por exemplo) e uma contagem do número de atributos modificados. Inicialmente considerou-se apenas a função de distância para medir a similaridade e por vezes o AG obtinha exemplos contrafactuais que, embora próximos da instância original, eram gerados as custas de pequenas mudanças em muitos atributos.

Assim percebeu-se que o conceito de "mínima mudança" na instância original precisava considerar o número de atributos a serem modificados. Para aplicação prática das explicações não é desejável que o usuário precise atuar sobre muitos atributos, ainda que para pequenas mudanças. Sendo assim, introduziu-se o número de trocas (x_nchg) como outra medida a ser minimizada. Vale ressaltar que o método permite que se atribua peso zero a λ_1 ou λ_2, mas não a ambos, trazendo uma perspectiva particular para a noção de similaridade a ser empregada. Por exemplo, atribuindo peso zero a λ_1, estará se privilegiando potencialmente explicações com um número menor de trocas e consequentemente mais simples (curtas), embora possam ocorrer mudanças maiores de valores nessas trocas.

Por outro lado, quanto maior o peso atribuído ao λ_1 mais se valoriza a função de distância. Essa opção de parametrização tende a privilegiar o uso de atributos de valores contínuos nas explicações. Em um atributo contínuo é possível realizar pequenas mudanças de valor. Já em uma variável binária, somente é possível alterações extremas de valor, o que gera um impacto grande na função objetivo e acaba por penalizar os candidatos a contrafactual que utilizam esse tipo de atributo. De fato, nos experimentos realizados percebeu-se que quanto mais se valoriza a função de distância, menor é a frequência dos atributos binários nas explicações.

Para calcular a distância entre os contrafactuais e a instância original é necessário que a base apresente apenas atributos numéricos. Optou-se por não automatizar tal transformação nos dados, pois é mais adequado que o usuário faça as devidas conversões de maneira que fique melhor representada a diferença entre os valores presentes nos atributos, o que é importante para os resultados a serem obtidos pelo ECOSS.

O cálculo do $y\_dist(f(c), y')$ consiste na diferença $|f(c) - y'|$, onde $f(c)$ corresponde ao *output predict value* (soma dos efeitos de todos os atributos) gerado pelo SHAP para o candidato a contrafactual c, e y' corresponde ao *base value* que representa a fronteira

entre as classes.

Antes do cálculo das distâncias entre a instância original e cada candidato a contrafactual, aplica-se um processo de normalização para todos os atributos. Esse ajuste é necessário para evitar que as diferentes escalas presentes nos atributos prejudiquem o cálculo das distâncias. Outra necessidade de normalização ocorre com as métricas que compõem a função objetivo (Equação 4.1), uma vez que seus os valores referem-se a métricas de naturezas e significados distintos. Em ambos os casos, o processo de normalização é dado pela função Min-Max apresentada pela Equação 4.2 (PATRO; SAHU, 2015), com os valores passando a variar no intervalo [0,1].

$$X' = \frac{X - min}{max - min} \tag{4.2}$$

onde X representa o valor a ser normalizado, max e min representam, respectivamente, os valores mais alto e mais baixo presentes no atributo e X' corresponde ao valor normalizado.

4.2 Implementação

O Algoritmo 2 exibe uma visão geral da implementação do método proposto. Este consiste em uma adaptação do AG proposto por Linden (2005), apresentado na Seção 2.3, página 35.

Os parâmetros do AG, os quais o usuário pode ajustar para melhor atender a cada problema, são mostrados na Tabela 2. Na mesma tabela apresenta-se os valores *default* dos parâmetros, os quais foram obtidos a partir dos experimentos realizados com o método (Capítulos 6, página 63 e Capítulo 7, página 67). Deixamos como trabalhos futuros a realização de outros experimentos e a utilização de algum método heurístico de otimização para aprimorar a sintonia dos valores *default* destes parâmetros. Sobre a estrutura do cromossomo, este é um vetor com os atributos de entrada de uma instância.

Conforme descrito em *"Input"*, o método de explicações contrafactuais necessita que o usuário forneça as seguintes entradas:

1. f: o modelo de classificação treinado;

2. x: a instância de interesse/original;

3. ds: o conjunto de dados fornecido pelo usuário. Por meio do conjunto de dados, pode-se identificar os possíveis valores que cada atributo pode assumir na explicação. Essa é uma forma de obter o domínio de cada atributo sem a necessidade de intervenções do usuário;

Algoritmo 2: ECOSS Pseudocódigo

Input: f, x, ds, k, $static\_list$
Output: $C(x_i)$, $i = 1, 2, \cdots$,k

```
1  G ← ShapExplanationModel(f);
2  t ← 0;
3  P(0) ← InitializePopulation(x, ds, static_list);
4  while not Number of Generations do
5  |  Fitness(P(t), x, G);
6  |  P(t + 1) ← Elitism(P(t));
7  |  while not Population Size do
8  |  |  parents ← Selection(P(t));
9  |  |  children ← Crossover(parents);
10 |  |  children ← Mutation(children, ds, static_list);
11 |  |  P(t + 1) ← P(t + 1) + children;
12 |  end
13 |  t ← t + 1;
14 end
15 C ← GetBest(P(t), k);
16 ShowChanges(C, x);
```

Tabela 2 – Parâmetros do AG

Parâmetro	Descrição	Valor default
num_gen	número de gerações	30
pop_size	tamanho da população	100
$\lambda_1, \lambda_2, \lambda_3$	pesos da função objetivo	1, 1, 1
per_elit	percentual de elitismo	0.1
$aval\_cros$	probabilidade de cruzamento	0.8
$aval\_mut$	probabilidade de mutação	0.1

Fonte: Elaborado pelo autor.

4. k: quantidade de exemplos contrafactuais, e consequentemente de explicações, que o usuário deseja receber. Assim, o usuário poderá avaliar o exemplo que mais lhe interessa;

5. $static\_list$: opcionalmente, de acordo com o contexto e objetivos, o usuário pode definir uma lista de atributos que não podem ser modificadas e, desta forma, não irão constar na explicação contrafactual. Na Seção 4.3 (página 50) são discutidas possíveis perspectivas de uso para a $static\_list$.

Em linhas gerais, conforme o Algoritmo 2, o AG consiste na seguinte implementação:

1. $ShapExplanationModel(f)$ (Linha 1): um modelo de explicação G é gerado pelo

SHAP a partir do modelo de classificação f. Neste ponto, o modelo de explicação G é treinado uma única vez para que, posteriormente na função $Fitness()$, G seja capaz de avaliar os candidatos a exemplos contrafactuais gerados ao longo das gerações em relação a proximidade da classe desejada;

2. $InitializePopulation(x, ds, static\_list)$ (Linha 3): a população do AG é inicializada com mutantes da instância original x. São gerados *pop_size* mutantes pela modificação de um atributo sorteado de forma aleatória. O AG gera o domínio de cada atributo a partir do próprio conjunto de dados fornecido pelo usuário. O operador de mutação escolhe aleatoriamente um dos valores do conjunto, diferente do valor atual, e modifica o atributo sorteado (atributos incluídas na *static_list* não são modificados). Estes serão os pais da primeira geração;

3. $Fitness(P(t), x, G)$ (Linha 5): basicamente, a função avalia a similaridade de cada indivíduo da população P em relação a instância original x e a proximidade destes para a classe desejada por meio do modelo de explicação G. A função objetivo está detalhada na Seção 4.1, página 44;

4. $Elitism(P(t))$ (Linha 6): a seguir são gerados *chil_elit* filhos por elitismo ($chil\_elit = pop\_size * per\_elit$). Os demais filhos ($pop\_size - chil\_elit$) são gerados por operações genéticas;

5. $Selection(P(t), op)$ (Linha 8): foi usado roleta viciada como estratégia de seleção de forma a privilegiar os indivíduos com função de avaliação mais alta sem desprezar aqueles com função de avaliação mais baixa;

6. $Crossover(parents)$ (Linha 9): se selecionado pela probabilidade de cruzamento, é executado o operador de *crossover* de um ponto;

7. $Mutation(parents, ds, static\_list)$ (Linha 10): considerando a taxa de mutação, mutantes são gerados pela mesma estratégia apresentada na população inicial;

8. $GetBest(P(t), k)$ (Linha 15): ao se atingir a condição de parada, a população final é ordenada considerando a função de avaliação. Em seguida, conforme em *"Output"*, os k contrafactuais mais similares a x que atingirem a classe desejada são retornados;

9. $ShowChanges(C, x)$ (Linha 16): uma vez obtidos os k exemplos contrafactuais $c_1, c_2, ..., c_k$, estes são comparados a instância original (x) para identificar em quais atributos se diferem. As diferenças entre c_i e x constitui um dos possíveis conjuntos de alterações necessárias para que a entrada escolhida mude para a classe desejada. A Tabela 3 exemplifica o formato da saída considerando $k = 3$.

Caso o usuário tenha selecionado alguma característica para se manter estática, estas são exibidas na parte inferior da Tabela 3.

Tabela 3 – *Template* da saída gerada pelo método

Atributos modificados				
	<atributo 1>	<atributo 2>	...	<atributo n>
Instância original classe: <saída original>	<valor original 1>	<valor original 2>	...	<valor original n>
Contrafactual(is) class: <classe contrafact.>	<novo valor 1>	–	...	–
	<novo valor 2>	<novo valor 3>	...	–
	–	<novo valor 4>	...	<novo valor n>
Atributo(s) estático(s): <atr. estático 1>, <atr. estático 2>, ..., <atr. estático n>				

Fonte: Elaborado pelo autor.

4.2.1 Utilização do método por meio de um pacote Python

Para ampliar o acesso e simplificar o processo de programação usando o ECOSS, o código foi estruturado para se tornar um pacote do Python. A implementação do método está disponível em `https://github.com/marcelobalbino/ECOSS`. O Algoritmo 3 ilustra como o usuário do método poderá incluí-lo em sua programação*

Algoritmo 3: Exemplo implementação usando ECOSS

```
1 from ecoss import ECOSS
2 ...
3 Código do Modelo de Classificação do Usuário
4 ...
5 #Instanciando o objeto
6 explainerECOSS ← ECOSS(f, x, ds)
7 #Retorna a lista dos contrafactuais
8 counterfactual_solution ← explainerECOSS.explain()
```

Inicialmente, o usuário deverá importar o pacote ECOSS (Linha 1) para utilizar seus recursos. Após o código do modelo de classificação, é necessário instanciar um objeto para acesso às funções do método (Linha 6). Os parâmetros obrigatórios para instanciar o objeto são: modelo de classificação (f), a instância para qual a explicação será gerada (x) e o conjunto de dados de entrada (ds). Os parâmetros opcionais, se não informados, assumem os valores padrão: exibição de três contrafactuais ($k = 3$), nenhuma restrição de atributo a constar na explicação (*static_list* vazia), se deve ou não imprimir os contrafactuais gerados ($print = True$) e os parâmetros padrão do AG (Tabela 2). Por

*O pacote será disponibilizado publicamente após a publicação do método.

meio do objeto instanciado deve-se chamar a função (*explain*()) que irá executar o método e retornar um objeto que consiste no conjunto de contrafactuais encontrado (Linha 8). Por meio deste objeto, se o usuário optar por não utilizar a saída padrão do método, este pode desenvolver sua própria implementação de saída ou manipular o conjunto de contrafactuais com o propósito que desejar.

Além disso, será desenvolvido um assistente (ajuda) para orientar na utilização do método, especialmente em relação às possibilidades e perspectivas oferecidas pelos parâmetros de entrada e ajuste dos parâmetros do AG.

4.3 Explicações orientadas ao usuário

Uma necessidade relatada na literatura que esta pesquisa busca superar refere-se as falhas de comunicação com os usuários dos atuais métodos de interpretabilidade. Para tal, propõe-se prover explicações alinhadas com as diretrizes destacadas por Miller (2019), segundo o qual explicações contrastivas, selecionadas e sociais possuem melhor capacidade de compreensão e relevância para o usuário final. Uma vez que o método proposto é baseado em explicações contrastivas, o desafio é incluir elementos que contribuam com os princípios da seletividade e socialização das explicações. Isso significa incluir recursos que tornem possível o usuário interagir com o método, permitir que seu contexto e intenções sejam considerados, tornando assim a explicação relevante e geradora de conhecimento.

Neste sentido, o método permite que o usuário determine uma quantidade k de explicações a receber. Assim, é possível selecionar a explicação que mais lhe interessa, seja para o entendimento de uma decisão do modelo, seja para uma tomada de decisão a partir da explicação. Com relação a esse recurso, é necessário fazer uma consideração, já que os AGs mono-objetivo geralmente são usados para encontrar uma única solução. Ao buscar mais de uma solução é comum encontrarmos soluções próximas, o que resultaria em explicações contrafactuais sem muita diversidade e possivelmente limitaria a escolha do usuário. Para contornar tal situação, o método exibe a melhor solução e as demais $k-1$ apresentadas são selecionadas, seguindo a ordenação, de forma a conter um conjunto diferente de atributos. Desta forma, temos algum grau de diversidade de soluções, gerando alternativas para o usuário. Na Seção 7.1.4 (página 73), detalhamos a estratégia para seleção das explicações com diversidade aplicada a uma instância da base de dados de TDAH. Consideramos que um exemplo prático facilita tal compreensão.

Outro recurso presente no método é a *static_list*, por meio da qual o usuário pode restrinjir o conjunto de possíveis de atributos a ser utilizado nas explicações. De acordo com as necessidades de cada usuário, a *static_list* pode trazer diferentes perspectivas de uso para o método. Dentre elas, pode-se destacar:

1. Apontar atributos que na prática não podem ser modificados, como o gênero de uma pessoa, por exemplo. Neste caso, o usuário está interessado com a aplicação prática das explicações;

2. Se o usuário está interessado em explicações que deem suporte a uma tomada de decisão de curto prazo, pode ser necessário retirar ainda atributos que o usuário acredita que não podem ser alteradas de forma imediata, como por exemplo elevar o grau de instrução ou mudar a classe social de uma pessoa;

3. Restringir possíveis atributos a constar na explicação aqueles que são de interesse e relevância para um determinado usuário, tornando a explicação próxima de um determinado contexto e ao encontro das capacidades de atuação do usuário. Por exemplo, suponha um problema que envolva atributos multidisciplinares. Para um médico, pode ser atrativo uma explicação que se restrinja a atributos da sua área, para que a explicação tenha sentido para o mesmo e inclua elementos sobre o qual este é capaz de agir;

4. Suponha que o usuário não utilize a *static_list*. Não havendo restrições, será possível visualizar os atributos que estão na fronteira das decisões, o que auxilia no seu entendimento. Além disso, outros aspectos podem ser evidenciados. Hipoteticamente, se em um modelo de aprovação de crédito bancário, a explicação mostra que a alteração do gênero, religião ou raça do indivíduo mudaria a previsão, isso pode indicar um viés discriminatório nas decisões.

Assim, tanto as opções de parametrização do AG quanto as entradas fornecidas ao método permitem que o usuário comunique aspectos relevantes ao seu contexto. Somando esses recursos à capacidade de comunicação inerente às explicações contrafactuais, consideramos que o método contempla os requisitos apontados por Miller (2019), o que segundo o autor gera explicações que aumentam a compreensão e a confiança do usuário sobre as decisões dos modelos de AM. Adicionalmente, confirma-se a Hipótese 1, apresentada na Seção 1.2 (página 12), de que é possível a implementação de um método agnóstico que atenda os requisitos apontados por Miller (2019).

5 MATERIAIS E MÉTODOS

Considerando a classificação de Vergara (2006), uma pesquisa pode ser vista em relação a seus fins e seus meios de investigação. Quanto aos fins, este trabalho se enquadra como pesquisa metodológica, uma vez que o objeto de estudo é o desenvolvimento de um instrumento (método) para atingir um determinado fim. Quanto aos meios, a pesquisa é bibliográfica, de laboratório e com estudo de caso.

A pesquisa é bibliográfica pois todo o estudo desenvolvido é baseado em material publicado em artigos e livros para levantamento dos conceitos e métodos relacionados a interpretabilidade, com destaque para o trabalho de Miller (2019) sobre o qual o ECOSS está pautado. Considera-se que esta é uma pesquisa de laboratório diante das simulações em computador utilizadas para desenvolvimento, avaliação e experimentação do método proposto. Por fim, ressaltam-se os estudos de caso nos quais o ECOSS foi aplicado que permitiram não só exemplificar seus recursos mas também atender demandas reais apresentadas.

As principais etapas que constituem este trabalho são:

- Levantamento bibliográfico: levantamento do referencial teórico relacionado a interpretabilidade, conceitos, os métodos atuais com seus recursos e limitações e os instrumentos necessários para o desenvolvimento do método proposto;
- Desenvolvimento do ECOSS: descrição do método proposto cujos elementos centrais são AG e a abordagem SHAP. A implementação foi realizada na linguagem Python utilizando a ferramenta Jupyter Notebook*. O método será disponibilizado como um pacote do Python;
- Avaliação do ECOSS: a Hipótese 2 deste trabalho considera que a utilização conjunta de AG e a abordagem SHAP é capaz de gerar contrafactuais tão próximos quanto possível da instância original. Para avaliar tal capacidade do método proposto realizou-se experimentos com o ECOSS comparando-o aos métodos LORE (Guidotti et al., 2019) e SOC (WACHTER; MITTELSTADT; RUSSELL, 2017). Os procedimentos metodológicos aplicados nestes experimentos estão descritos na Seção 5.1, página 53.
- Estudos de caso: aplicação do ECOSS em estudos de caso provenientes de demandas apresentadas por parceiros de diferentes departamentos da Universidade Federal de

*Disponível em `https://jupyter.org/`

Minas Gerais (UFMG). Tais estudos de caso referem-se a extração de conhecimento em conjuntos de dados relacionados ao desempenho acadêmico de crianças e adolescentes com TDAH e a investigação de sintomatologias depressivas também em crianças e adolescentes. Os procedimentos metodológicos aplicados a estes estudos estão detalhados na Seção 5.2, página 55.

Em todos experimentos foram computados o tempo de execução do método para cálculo do tempo médio por instância original. Os experimentos foram realizados em ambiente Windows, 8.0 GB de RAM, 2.80 GHz Intel Core i7.

5.1 Procedimentos metodológicos da avaliação do método

Guidotti et al. (2019) compararam seu método LORE[†] com o SOC de Wachter, Mittelstadt e Russell (2017). Nos experimentos realizados por Guidotti et al. (2019) foram utilizadas as bases de dados públicas Compas[‡] e German[§] cujas descrições encontram-se na Seção 5.1.1, página 54. Utilizamos os resultados publicados por Guidotti et al. (2019) para comparação com o ECOSS.

Guidotti et al. (2019) adotaram *the number of falsified conditions in counterfactual* (nf) como métricas para comparar seu método (LORE) ao SOC de Wachter, Mittelstadt e Russell (2017). O objetivo é encontrar o contrafactual com o mínimo de mudanças (menor valor de nf) gerando uma explicação mais simples, o que contribui com a compreensão e aplicação da explicação. Embora a proposta do LORE seja gerar explicações baseadas em regras e o ECOSS aborda explicações por exemplos, entendemos que a comparação é adequada considerando o número de atributos presentes nas regras ou nos casos contrafactuais. Logo, a comparação por meio do nf avalia o quanto os métodos são capazes de gerar contrafactuais próximos da instância original. Assim, o método com menor nf terá gerado uma explicação menor em relação a quantidade de atributos e, consequentemente mais simples.

A partir deste conceito de qualidade do contrafactual adotado por Guidotti et al. (2019), ao executar os experimentos com o ECOSS utilizou-se os parâmetros padrão do AG com exceção do λ_1 (peso da função de distância) o qual foi atribuído peso zero para privilegiar o número de atributos alterados no contrafactual, gerando assim explicações mais curtas e consequentemente mais simples. Este é um exemplo de como as opções de parametrização do ECOSS permitem o usuário adequar o método às necessidades específicas.

[†]Disponível em `https://github.com/riccotti/LORE`
[‡]Disponível em `https://github.com/propublica/compas-analysis`
[§]Disponível em `https://archive.ics.uci.edu/ml/datasets/statlog+(german+credit+data)`

Nos experimentos realizados com o ECOSS foi reproduzido o cenário utilizado por Guidotti et al. (2019). Desta forma, foram utilizados os mesmos conjuntos de dados, transformações e critérios de separação de dados adotados pelo autor, ou seja, 70% da base de dados foram utilizados para treinamento dos modelos caixa preta e 30% foram destinados como instâncias a serem explicadas. Para o modelo de classificação caixa preta, utilizou-se o algoritmo *Random Forest*. O conjunto de dados utilizado para explicação também foi usado para teste do modelo de classificação desenvolvido. Para avaliar a qualidade do modelo utilizou-se as métricas de *Precision*[¶], *Recall*[‖], *F-measure*[**].

5.1.1 Bases de dados Compas e German

As bases de dados Compas e German foram utilizadas por Guidotti et al. (2019) nos experimentos para avaliação e comparação de método proposto LORE com o SOC (WACHTER; MITTELSTADT; RUSSELL, 2017). Tratam-se de bases públicas cujos dados já se encontram pré-processados. Contudo, especialmente na base Compas, os autores realizaram algumas transformações.

A base Compas contém dados relacionados ao histórico criminal, prisão e tempo de prisão, dados demográficos e três *Compas scores* (*"Risk of Recidivism", "Risk of Violence"* e *"Risk of Failure to Appear"*) de 7.214 réus do Condado de Broward de 2013 e 2014.

Quando a maioria dos réus é autuado na prisão, eles respondem a um questionário Compas para avaliar o risco do mesmo se tornar um reincidente. O condado de Broward usa principalmente o *Compas score* para determinar se deve liberar ou deter um réu antes de seu julgamento (LARSON et al., 2016). As pontuações do Compas para cada réu variaram de 1 a 10, sendo dez o risco mais alto. Guidotti et al. (2019) rotularam as pontuações de 1 a 6 como risco "Medium-Low" e 7 a 10 como "High". Considerando o *Risk of Recidivism*, a base possui 5219 réus classificados como "Medium-Low" e 1995 como "High". Originalmente, a base possui ao todo 32 atributos. Para os experimentos, os autores selecionaram determinados atributos e ainda realizaram algumas transformações na base de dados que resultaram em 11 atributos (descritos no Apêndice A, página 92), além do atributo classe.

A base German credit inclui dados de 1000 candidatos a empréstimos. Cada candidato é descrito por um conjunto de 20 atributos, além do atributo classe. Os atributos de entrada incluem dados pessoais, financeiros, de propriedades, profissionais, situação e histórico bancários e dados relacionados ao empréstimo em si. A listagem completa dos atributos está no Apêndice B, página 93.

[¶] $Precisão = \frac{VP}{VP+FP}$

[‖] $Sensibilidade = \frac{VP}{VP+FN}$

[**] $F\text{-}measure = \frac{2 \times Sensibilidade \times Precisão}{Sensibilidade + Precisão}$

A cada candidato é atribuída uma classificação de risco de crédito como “good”(0) ou “bad”(1). Na base de dados, 700 instâncias pertencem à classe de bons candidatos e 300 a classe dos candidatos ruins ao crédito.

5.2 Procedimentos metodológicos dos estudos de caso

O primeiro estudo foi desenvolvido junto a equipe médica do Departamento de Pediatria da Universidade Federal de Minas Gerais e refere-se a predição do desempenho acadêmico de crianças e adolescentes com TDAH em “Superior” ou “Inferior” em aritmética, leitura e escrita. Já o segundo estudo, originou de um levantamento realizado em parceria com o Programa de Pós-Graduação em Psicologia: Cognição e Comportamento da mesma instituição parceira. Neste caso, pretende-se classificar a sintomatologia de depressão em crianças e adolescente como “Alta” ou “Baixa”.

Em linhas gerais, ambos estudos seguiram os mesmos procedimentos metodológicos cujas principais etapas foram:

1. *Estudo e pré-processamento da base de dados*: os levantamentos realizados pelos parceiros deram origem as bases de dados dos referidos estudos. Tais dados foram analisados e tratados para melhor se adequar aos algoritmos de classificação. Paralelamente, fizeram-se necessários estudos específicos sobre cada um dos transtornos abordados (TDAH ou depressão). A descrição, pré-processamento e outros aspectos específicos de cada uma das bases de dados estão apresentados nas Seções 5.2.1 e 5.2.2;

2. *Desenvolvimento e avaliação dos modelos preditivos*: buscando obter uma melhor capacidade preditiva foram desenvolvidos modelos baseados em quatro algoritmos de AM: Árvore de Decisão, Redes Neurais, SVM e *Random Forest*. Os modelos foram implementados em Python usando a biblioteca Scikit-learn. Para avaliar a qualidade dos modelos foram utilizadas as métricas de *Precision*, *Recall* e *F-measure*. Todos os classificadores foram construídos e validados usando o processo de validação cruzada de k -fold, com $k = 10$. A intenção é trabalhar com o modelo de melhor desempenho preditivo possível, já que naturalmente não faz sentido explicar predições incorretas. Ressalta-se que no estudo de caso de TDAH foi criado um modelo de classificação para predição de desempenho em cada disciplina (aritmética, leitura e escrita);

3. *Aplicação do ECOSS*: uma vez identificado o melhor modelo de classificação, realizou-se alguns experimentos com o ECOSS utilizando o mesmo conjunto de instâncias o qual se aplicou os testes do modelo de classificação. Tais experimentos visaram avaliar a capacidade do método na geração de contrafactuais e extrair conhecimento

das bases de dados disponíveis. Nas execuções foram sempre mantidos os parâmetros padrão do método (Seção 4.2, página 46 - Tabela 2), dentre os quais ressalta-se a busca por três explicações contrafactuais para cada instância original. Mais especificamente, foram gerados os seguintes resultados:

- *Apresentação de explicações contrafactuais para instâncias selecionadas aleatoriamente*: a intenção foi exemplificar os recursos presentes no método. Em todos os estudos de caso, optou-se por apresentar explicações para instâncias da classe considerada indesejável, já que reverter um quadro desfavorável tende a ser um cenário mais comum do ponto de vista prático;

- *Cálculo de eficácia do método*: percentual de sucesso na obtenção das três explicações contrafactuais;

- *Ranking dos 10 atributos mais frequentes nas explicações contrafactuais*: possibilita observar os atributos com maior capacidade de reverter uma predição. Adicionalmente, foi realizada uma observação geral das explicações individuais no intuito de entender o comportamento dos principais atributos;

- *Cálculo da sensibilidade do modelo de classificação às mudanças nos atributos*: foi definido como "sensibilidade a mudanças"(SM), o fator inverso a quantidade média de trocas (QM) para geração do exemplo contrafactual. Mais especificamente, como apresentado na Equação 5.1, dado uma amostra de N instâncias X_i a serem explicadas, no qual serão gerados até k contrafactuais (C_{ij}) para cada instância (por exemplo, C_{13} representa o contrafactual 3 da instância original X_1), QM é expresso

$$QM = \frac{\sum_{i=1,j=1}^{N,k} QT_{i,j}}{\sum_{i=1}^{N} QC_i} \tag{5.1}$$

  onde QT_{ij} é quantidade de atributos alterados para gerar cada contrafactual C_{ij} e QC_i é a quantidade de contrafactuais obtidos para cada instância original X_i. Por exemplo, suponha uma amostra somente com duas instâncias originais X_1 e X_2 para as quais foram gerados para cada uma dois contrafactuais (C_{11}, C_{12}, C_{21}, C_{22}), a equação seria QM = $(QT_{11} + QT_{12} + QT_{21} + QT_{22})/4$, já que quatro contrafactuais foram gerados no total.
  A sensibilidade a mudanças (SM) é inversamente proporcional a QM. Em outras palavras, quanto menor a quantidade média de trocas necessária para gerar os contrafactuais, mais sensível é o modelo às mudanças nos atributos. No geral, esta alta sensibilidade indica a presença de um ou mais atributos

com uma importância muito destacada, cuja mudança é capaz de reverter a predição da instância sobre o qual está sendo aplicado.

É importante destacar que, por se tratar de um método de explicação local, na prática espera-se que o usuário do método direcione sua aplicação a instâncias de interesse previamente definidas.

5.2.1 Contextualização e descrição da base de dados do estudo de TDAH

Presente no *Diagnostic and Statistical Manual of Mental Disorders* (DSM), o Transtorno de Déficit de Atenção/Hiperatividade (TDAH) é definido por níveis prejudiciais de desatenção, desorganização e/ou hiperatividade e impulsividade, sintomas esses excessivos quando comparado com outras pessoas de mesma idade e grau de desenvolvimento. Levantamentos sugerem que o TDAH ocorre na maioria das culturas em cerca de 5% das crianças e 2,5% dos adultos, sendo mais frequente no sexo masculino (ASSOCIATION et al., 2014).

Os prejuízos causados pelo TDAH afetam a vida do indivíduo em vários aspectos. Nesse contexto, a escola tem papel fundamental no desenvolvimento cognitivo e socioemocional do ser humano. No entanto, devido ao seu funcionamento peculiar, os alunos com TDAH tendem a apresentar variados problemas acadêmicos, como dificuldades de aprendizagem, comportamentos considerados impróprios ao ambiente escolar, dificuldades de relacionamento com os colegas, tendo perdas pedagógicas e sociais consideráveis. Sendo assim, é necessário que as escolas adotem estratégias adequadas para esse público, de forma a contribuir efetivamente para o desenvolvimento dos discentes com TDAH. Porém, de uma forma geral, as instituições de ensino têm encontrado dificuldades para lidar com estudantes com o transtorno (JÚNIOR; LOOS, 2011). Diante disso, como resultado das dificuldades inerentes ao TDAH e ainda a pouca adequação das escolas para lidar com discentes com o transtorno é natural encontrar um quadro geral de baixo desempenho dos alunos.

Por outro lado, acredita-se que o TDAH não é um fator definitivo de baixo desempenho. Apesar das dificuldades impostas pelo transtorno, existem indivíduos com TDAH que não apresentam déficits acadêmicos (FRAZIER et al., 2007). Sendo assim, é possível que outras características possam potencializar ou minimizar os prejuízos causados pelo transtorno.

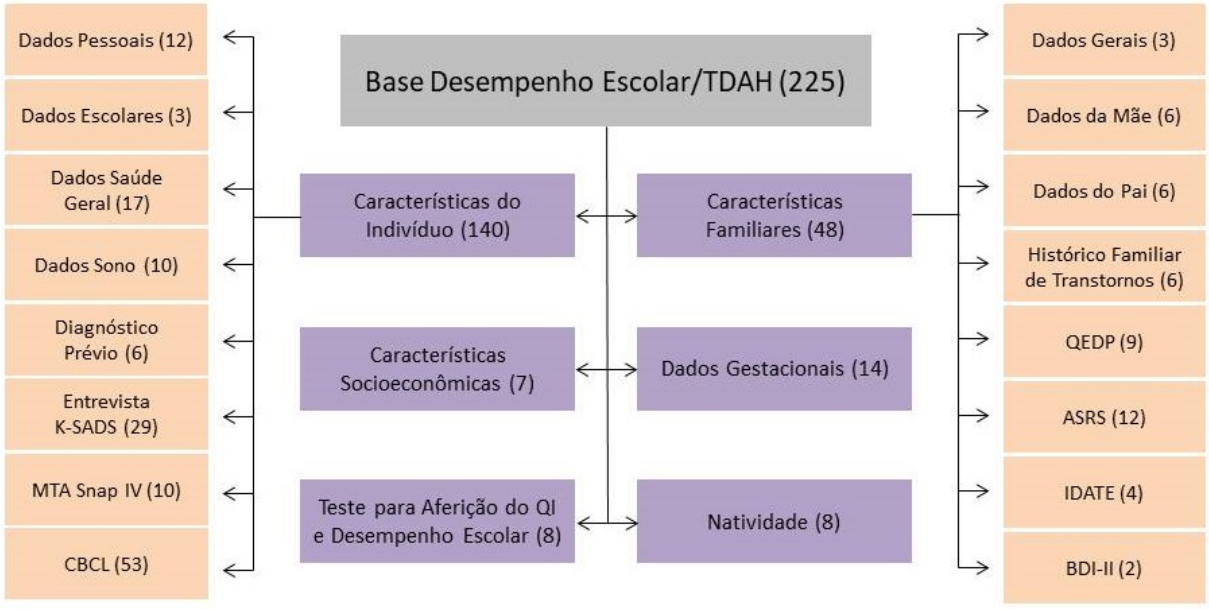
Em vista do cenário descrito, é preciso levantar as variáveis que interferem no desempenho dos alunos com TDAH. Identificar casos de sucesso e compreender quais fatores levaram a esse resultado positivo pode ajudar a guiar ações que beneficiem outros estudantes. Estudar os casos de baixo desempenho, entendendo quais variáveis influenciaram

nesse quadro, pode dar subsídios para que pais, educadores e demais profissionais (psicólogos, psiquiatras e neurologistas) direcionem suas ações na busca por melhores resultados para os discentes com o transtorno.

O Departamento de Pediatria da Universidade Federal de Minas Gerais realiza um trabalho de apoio a crianças e adolescentes com diagnóstico prévio de TDAH. Uma vez confirmado o diagnóstico, os pacientes passam a ser acompanhados no hospital vinculado à instituição.

Por meio de questionários manuais e entrevistas, a equipe médica responsável por este acompanhamento realizou um levantamento com 266 crianças/adolescentes de idade entre 6 e 18 anos dos quais 196 foram diagnosticadas com TDAH e 70 possuem diagnóstico negativo para o transtorno. Esse levantamento gerou uma base de dados com 225 atributos incluindo dados pessoais, familiares, gestacionais de saúde, médicos, sócioeconômicos, cuidados parentais, escolaridade, entre outros, além das pontuações relativas às provas de aritmética, escrita e leitura de cada indivíduo presente na base de dados. A Figura 14 apresenta uma visão geral da base com os tipos de atributos e suas quantidades. A lista completa de atributos encontra-se no Apêndice C, página 96.

Figura 14 – Visão geral da base de dados de TDAH.



Fonte: Adaptado de Jandre et al. (2021).

A existência de uma base de dados de alunos com TDAH é uma fonte valiosa de pesquisa e precisa ser objeto de estudo na busca de conhecimento sobre esse cenário que se encontra com questões em aberto.

5.2.1.1 Pré-processamento da base TDAH

Para melhorar a qualidade dos modelos obtidos foi realizado o pré-processamento da base de dados. Os detalhes do pré-processamento estão descritos em Jandre et al. (2021). Em linhas gerais, foram realizados os seguintes procedimentos: exclusão de atributos duplicados, codificação numérica de atributos nominais, tratamento dos dados ausentes e redução de dimensionalidade. Ressalta-se ainda a transformação das notas obtidas no Teste de Desempenho Escolar, que é subdividido em escrita, aritmética e leitura, nas classes Superior e Inferior. Utilizou-se do Manual do Teste de Desempenho Escolar (STEIN, 1994) para comparar as notas dos alunos com a nota média do mesmo estado onde o teste foi aplicado.

Sobre a divisão da base para treinamento/validação e teste, realizou-se a separação aleatória de 15% das instâncias de cada classe para realizar a etapa dos teste. Os dados restantes foram balanceados por meio do algoritmo *SpreadSubsample* presente na ferramenta WEKA[††] (*Waikato Environment for Knowledge Analysis*). Para realizar o balanceamento foi necessário separar a base em três arquivos, um por disciplina, uma vez que o número de instâncias de cada classe é diferente em cada disciplina. A Tabela 4 apresenta o número de instâncias reservadas para treinamento/validação e teste em cada disciplina. O número de instâncias para teste em leitura é menor, pois havia algumas instâncias com desempenho ausente. Após todo processo as bases de aritmética, escrita e leitura ficaram com 29, 31 e 22 atributos de entrada respectivamente.

Tabela 4 – Número de instâncias de treinamento/validação e testes da base TDAH

Disciplina	Treinamento/Validação	Testes	
		Superior	Inferior
Aritmética	118	11	29
Escrita	100	09	31
Leitura	78	08	27

Fonte: Dados da Pesquisa.

[††]WEKA é um software de código aberto emitido sob a Licença Pública Geral GNU que contém uma coleção de algoritmos de AM (SINGHAL; JENA, 2013). Disponível em `http://www.cs.waikato.ac.nz/ml/weka`.

5.2.2 Contextualização e descrição da base de dados do estudo de depressão

A depressão é um termo usado para se referir aos Transtornos Depressivos (TD), sendo entendido como uma patologia que altera e compromete o corpo e a mente, afetando principalmente o humor. O indivíduo com TD pode apresentar tristeza persistente, falta de interesse ou prazer em atividades que anteriormente eram gratificantes, perda de confiança e autoestima, sentimento injustificado de culpa, ideias de morte e suicídio, perturbações do sono e apetite, fadiga, baixa concentração e sintomas de ansiedade. Os seus efeitos podem ser duradouros ou recorrentes e podem afetar a capacidade de uma pessoa no cotidiano familiar, social, acadêmico, ocupacional e outras áreas importantes de funcionamento (APA et al., 2013; WHO, 2017, 2018).

Apenas de 2005 a 2015, houve um aumento de 18% na quantidade de pessoas com depressão em todo mundo, resultando em mais de 300 milhões de pessoas (WHO, 2017). Além disso, estima-se que uma em cada seis pessoas (cerca de 16,67%) sofrerá com a depressão em algum momento de sua vida, o que significa mais de um bilhão pessoas afetadas pelo transtorno no planeta (APA, 2017).

Estudos indicam que os TD têm sido a principal causa de doenças e incapacidades na adolescência (WHO, 2017). Além disso, metade das pessoas que desenvolvem transtornos mentais tem os primeiros sintomas até os 14 anos (YOON; TAHA; BAKKEN, 2014). Diante disso, é fundamental a identificação e tratamento dos indivíduos com depressão ainda na infância/adolescência para evitar que as perdas ocasionadas pelo transtorno o acompanhem ao longo da vida.

Contudo, a definição da depressão na juventude não é tratada especificamente no Manual Diagnóstico e Estatístico de Transtornos Mentais (APA et al., 2013). Não se diferencia os critérios diagnósticos dos transtornos depressivos para crianças, adolescentes ou adultos. Porém, vários autores afirmam que as particularidades da infância devem ser consideradas na avaliação e diagnóstico da depressão em crianças (VERSIANI; REIS; FIGUEIRA, 2000; LACERDA-PINHEIRO et al., 2014; QUEVEDO; NARDI; SILVA, 2018; BERNARAS; JAUREGUIZAR; GARAIGORDOBIL, 2019).

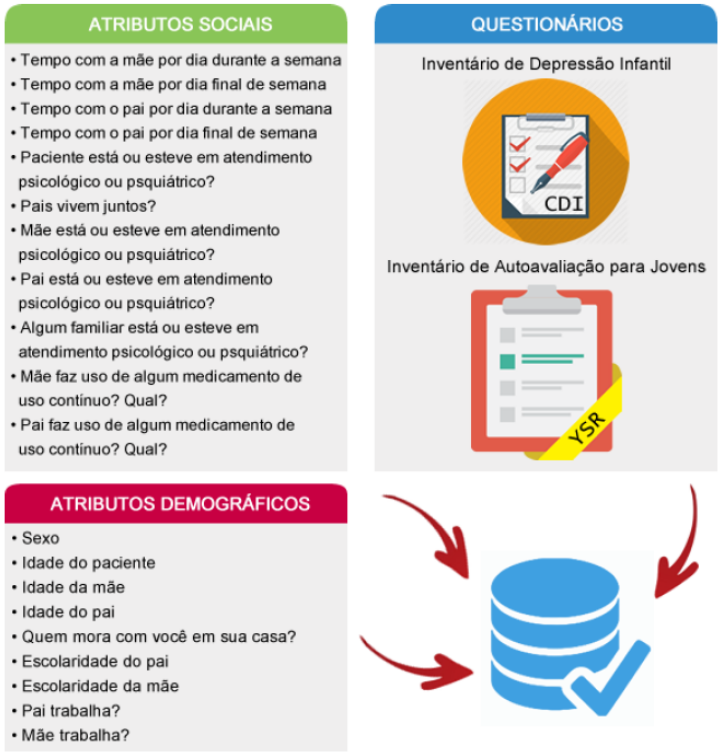
No entanto, um dos obstáculos para tratamento da depressão é justamente sua avaliação e diagnóstico, levando a falta de tratamento ou uma condução inadequada do mesmo (PAVLOVA; UHER, 2020). Tal cenário, evidencia a importância de instrumentos que possam dar suporte ao correto diagnóstico da depressão.

Um levantamento realizado em parceria com o Programa de Pós-Graduação em Psicologia: Cognição e Comportamento da Universidade Federal de Minas Gerais deu origem a uma base de dados que contém informações de crianças e adolescentes entre 10 e

16 anos, sendo 158 do sexo masculino e 219 do sexo feminino, totalizando 377 instâncias com diferentes sintomatologias depressivas. Esse levantamento foi realizado por meio de questionários manuais e entrevistas realizadas pela equipe médica responsável pelo acompanhamento dos pacientes.

A Figura 15 apresenta uma visão geral da base de dados. Em linhas gerais, são 75 atributos que incluem características pessoais, demográficas, sociais e as pontuações obtidas pelos inventários *Children's Depression Inventory* (CDI) e *Young Self Report* (YSR). Outras questões consideradas importantes pela comunidade de saúde mental também foram incluídas, principalmente fatores como ansiedade, problemas sociais, falta de atenção, agressividade, problemas de conduta (APA et al., 2013). A lista completa dos atributos da base encontra-se no Apêndice D, página 109.

Figura 15 – Combinação de atributos que compõem a base de dados de Depressão.



Fonte: Extraído de Santana et al. (2019)

Tendo em vista o alto índice de incidência do TD e as suas peculiaridades em crianças e adolescentes, é essencial explorar a referida base de dados, tornando-o uma fonte de conhecimento que possa subsidiar o diagnóstico e ações para auxiliar indivíduos nesta faixa etária.

Os sistemas de AM tem alcançado resultados satisfatórios no suporte ao diagnóstico de doenças, como é o caso do problema em questão. No entanto, em se tratando

da área de saúde, a interpretabilidade do modelo é essencial, pois o especialista precisa entender e confiar nos resultados encontrados para efetivamente utilizá-los (KHADEMI; NEDIALKOV, 2015; RAVÌ et al., 2017). Sendo assim, o desenvolvimento de um modelo de classificação somado a interpretabilidade promovida por meio da aplicação do ECOSS pode trazer benefícios significativos para o cenário estudado.

5.2.3 Pré-processamento da base de depressão

Foi realizado o pré-processamento dos dados visando uma melhor adequação aos algoritmos selecionados e obtenção de modelos mais consistentes. Os detalhes do pré-processamento estão descritos em Balbino. et al. (2022). Resumidamente foram realizados os seguintes procedimentos: tratamento de inconsistências de dados, discretização dos valores de alguns atributos, codificação numérica de atributos nominais, tratamento de dados ausentes e redução de dimensionalidade. Após todo processo a base ficou com 45 atributos de entrada.

Inicialmente a base de dados não incluía um atributo para classificar os indivíduos pela sua sintomatologia. A partir da recomendação de Kovacs (2003) e com o apoio do especialista obteve-se 63 classificados como sintomatologia “High” e 314 como “Low”.

Foi realizada a separação aleatória de 15% das instâncias de cada classe para realizar a etapa de teste. Em seguida, os 85% restantes dos dados foram balanceados usando o algoritmo *SpreadSubsample* presente na ferramenta WEKA (SINGHAL; JENA, 2013). A tabela 5 mostra o número de instâncias reservadas por classe para treinamento/validação e teste.

Tabela 5 – Número de instâncias dos conjuntos de treinamento/validação e teste da base Depressãos

Classe	Instâncias Pré-processadas	Criação do Modelo	Teste
HIGH	63	53	10
LOW	314	53	50
Total	**377**	**106**	**60**

Fonte: Dados da Pesquisa.

6 AVALIAÇÃO DO MÉTODO PROPOSTO

Neste capítulo, é realizada uma experimentação do ECOSS visando avaliá-lo no que se refere a capacidade de gerar contrafactuais tão próximos quanto possível da instância original. Para tal, realizou-se experimentos para comparar o desempenho do ECOSS em relação aos resultados publicados por Guidotti et al. (2019), que por sua vez comparou seu método (LORE) com o SOC desenvolvido por Wachter, Mittelstadt e Russell (2017). Na Seção 6.1 (página 63), apresenta-se uma exemplificação de aplicação do método no intuito de esclarecer seu funcionamento e alguns recursos. Além disso, mediu-se o tempo médio de execução do ECOSS por instância orginal. Na Seção 6.2 (página 65), se descreve a referida avaliação e comparação do método. Nos experimentos utilizou-se as bases de dados Compas e German e o algoritmo *Random Forest*. Os detalhes dos procedimentos metodológicos e as descrições das bases de dados destes experimentos estão apresentados na Seção 5.1, página 53.

Os resultados obtidos com os testes do modelo de classificação nas bases Compas e German encontram-se nas Tabelas 6 e 7, respectivamente. Em ambos os casos, percebe-se uma diferença significativa de desempenho preditivo para as classes. Presume-se que o balanceamento das bases poderia equilibrar tal desempenho. No entanto, para manter-se consistente com os experimentos de Guidotti et al. (2019), optou-se por preservar as bases como utilizadas pelo autor.

Tabela 6 – Desempenho do modelo de classificação na etapa de teste na base Compas

Classe	Métricas (%)		
	Precisão	Sensibilidade	*F-measure*
High	72.0	52.0	60.0
Medium-Low	83.0	92.0	88.0
Acurácia	81.0		

6.1 Aplicação do método

Para exemplificar a aplicação do ECOSS utilizou-se a base German e, dentre as instâncias destinadas a explicação, uma instância da classe negativa (candidato com risco de crédito "bad"(1)) foi escolhida aleatoriamente. Além disso, optou-se pelo retorno de três casos contrafactuais ($k = 3$) e não se restringiu qualquer atributo a constar na

Tabela 7 – Desempenho do modelo de classificação na etapa de teste na base German

Classe	Métricas (%)		
	Precisão	Sensibilidade	*F-measure*
0 (Good)	89.0	95.0	92.0
1 (Bad)	79.0	62.0	70.0
Acurácia	87.0		

explicação (*static_list* vazia). Com relação aos parâmetros do AG, manteve-se os valores *default* apresentados na Tabela 2.

Para a instância escolhida, conforme apresentado na Tabela 8, o ECOSS indicou que o cliente seria classificado como "good" para o risco de crédito se: 1) tivesse 30 anos ao invés de 21 anos; ou 2) aumentar o valor em poupança/títulos (savings); ou 3) reduzir o valor do crédito solicitado (credit_amount).

Tabela 8 – Resultado da aplicação do ECOSS na base German considerando todos os atributos

Modified features			
	age	savings	credit_amount
Original instance: class: 1 (Bad)	21	1 (... <100 DM*)	15653
Counterfactuals class: 0 (Good)	30 — —	— 2 (100 <= ... <500 DM) —	— — 9157
No static feature			

*DM = Deutsche Mark = Marco Alemão

Fonte: Dados da Pesquisa.

A explicação contrafactual pode ser utilizada para compreensão de um comportamento local do modelo e eventualmente é seguida de alguma tomada de decisão. Neste caso, a mudança da idade do cliente apresentada na Tabela 8 como explicação para alterar a classificação do risco de crédito para "good" não seria viável de forma imediata. Hipoteticamente, poderíamos repetir a execução do método, porém incluindo o atributo "*age*" na *static_list* para que este não seja considerado na busca pelos exemplos contrafactuais. Assim, pode-se exemplificar um dos recursos presentes no método. A nova explicação obtida está apresentada na Tabela 9.

Diante da natureza estocástica dos AGs poderia-se obter explicações diferentes, porém os atributos relativos ao valor em poupança/títulos (*savings*) e valor do crédito solicitado (*credit_amount*) permaneceram, sendo este último com um valor ligeiramente

Tabela 9 – Resultado da aplicação do ECOSS na base German incluindo o atributo "age" na *static_list*

	Modified features		
	savings	credit_amount	duration_in_month
Original instance: class: 1 (Bad)	1 (... <100 DM)	15653	60
Counterfactuals class: 0 (Good)	2 (100 <= ... <500 DM)	_	_
	_	9277	_
	_	_	33
Static feature: age			

Fonte: Dados da Pesquisa.

superior ao encontrado anteriormente. A nova execução incluiu uma explicação que indica que a redução do tempo de pagamento do empréstimo (*duration_in_month*) mudaria a classificação do risco do candidato ao crédito para "good".

Em relação ao tempo de execução do ECOSS para geração de contrafactuais, o tempo médio por instância para a base German foi de 27.34 segundos e para a base Compas foi 87.18 segundos.

6.2 Resultados da avaliação do método

Por definição espera-se que instâncias contrafactuais sejam geradas por mudanças mínimas na instância original. Guidotti et al. (2019) consideraram que essa mudança mínima implica em alterar a menor quantidade possível de atributos na instância original ao gerar o contrafactual. Sendo assim, os autores adotaram *the number of falsified conditions in counterfactual* (nf) como métricas para comparar seu método (LORE) ao SOC de Wachter, Mittelstadt e Russell (2017). Neste caso, na busca por contrafactuais para uma instância original qualquer, quanto menor o nf mais simples (curta) será a explicação, o que contribui com a compreensão e aplicação da explicação por parte do usuário.

Considerando o critério adotado por Guidotti et al. (2019), o ECOSS foi ajustado de forma a privilegiar contrafactuais gerados a partir de trocas na menor quantidade possível de atributos em relação a instância original. Os detalhes dos procedimentos adotados e bases de dados utilizados nos experimentos foram descritos na Seção 5.1, página 53.

A Tabela 10 mostra o desempenho dos métodos LORE e SOC publicados por Guidotti et al. (2019) e os resultados encontrados nos experimentos com o ECOSS propostos neste trabalho.

Tabela 10 – Comparação de performance dos métodos contrafactuais.

Base de dados	***Método***	***nf***
German	LORE	1.52 ±1.18
	SOC	14.80 ±1.59
	ECOSS (método proposto)	1.47 ±0.67
Compas	LORE	1.84 ±0.78
	SOC	6.24 ±1.45
	ECOSS (método proposto)	1.17 ±0.38

Fonte: Dados da Pesquisa.

Os resultados mostraram que o ECOSS obteve um desempenho melhor que os outros métodos em ambos conjuntos de dados. O ECOSS e o LORE obtiveram resultados próximos quando comparados entre si, mas foram consideravelmente melhores que o SOC.

Enfim, diante dos cenários apresentados e dos critérios estabelecidos, o método alcançou um desempenho bastante satisfatório considerando a busca por contrafactuais tão próximos quanto possível da instância original. O ponto chave para o bom desempenho é o uso do SHAP como meio de identificar se os candidatos a contrafactuais produzidos estão se aproximando da classe desejada combinado a capacidade de explorar o espaços de busca do AG. Sendo assim, os resultados evidenciam que é verdadeira a Hipótese 2 (Seção 1.2, página 12) que declara que a utilização conjunta de AG e a abordagem SHAP é capaz de gerar contrafactuais com mudanças mínimas em relação a instância original.

7 ESTUDOS DE CASO

A aplicação de explicações contrafactuais é ainda mais consonante com problemas no qual uma das classes se refere a algo indesejável/negativo. Este é o caso dos estudos de caso apresentados a seguir. Na Seção 7.1, o modelo visa predizer o desempenho acadêmico de crianças e adolescentes com TDAH em "Superior" ou "Inferior" em aritmética, leitura e escrita. Já na Seção 7.2, o modelo pretende classificar a sintomatologia de depressão em crianças e adolescente como "Alta" ou "Baixa". Em ambos estudos, as explicações contrafactuais permitirão não só perceber os atributos que estão na fronteira das predições, como também direcionar ações para reverter os quadros considerados indesejáveis para cada cenário. Os procedimentos metodológicos e as descrições das bases aplicadas nestes estudos de caso estão apresentados na Seção 5.2, página 55.

7.1 Resultados do estudo de caso desempenho escolar de crianças e adolescentes com TDAH

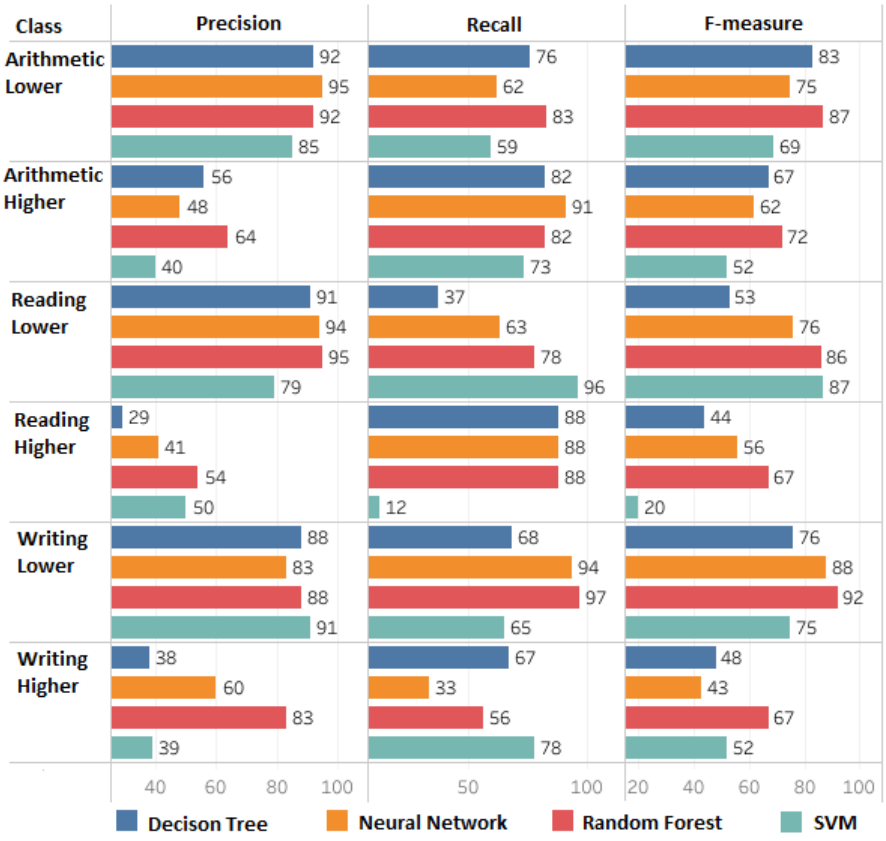
O Departamento de Pediatria da Universidade Federal de Minas Gerais realiza um trabalho de apoio a crianças e adolescentes com diagnóstico prévio de TDAH. A equipe médica que integra este trabalho realizou um levantamento com 266 pessoas de idade entre 6 e 18 anos dos quais 196 foram diagnosticadas com TDAH e 70 possuem diagnóstico negativo para o transtorno, dando origem a base de dados já apresentada na Seção 5.2.1, página 57.

Na busca por obter a melhor capacidade preditiva em relação ao desempenho acadêmico dos estudantes em aritmética, leitura e escrita foram desenvolvidos modelos basedos em quatro métodos de AM: Árvore de Decisão, Redes Neurais, SVM e *Random Forest*. A proposta foi aplicar o método apenas nos modelos de melhor desempenho em cada disciplina.

A Figura 16 apresenta os resultados da fase de teste dos modelos preditivos. É possível perceber que, de uma forma geral, o modelo baseado em *Random Forest* obteve os melhores resultados. Em relação à aritmética, o modelo com *Random Forest* obteve *F-measure* de 87% para classe inferior e 72% para classe superior. Para leitura, o modelo obteve 86% para classe inferior e 67% para classe superior. Em relação ao comportamento do modelo para escrita, o mesmo alcançou o *F-measure* de 92% para classe inferior e 67% para classe superior. Pontualmente os outros modelos apresentaram resultado igual ou superior para *Precision* ou *Recall* em uma determinada classe/disciplina. Sendo assim,

optou-se por aplicar o ECOSS nos modelos com *Random Forest* para todos os cenários.

Figura 16 – Avaliação do desempenho dos modelos no contexto de TDAH.



Fonte: Dados da Pesquisa.

7.1.1 Resultados obtidos pela aplicação do ECOSS no modelo para aritmética

Nestes experimentos, utilizando o modelo de predição de desempenho acadêmico em aritmética com *Random Forest*, aplicou-se o ECOSS para gerar explicações para as predições das 40 instâncias selecionadas como conjunto de teste/explicação. O tempo médio de execução por instância foi de 17.46 segundos.

Os atributos mencionados na descrição dos experimentos estão listados na Tabela 11, bem como a transformação numérica de seus valores. A lista completa de atributos da base encontra-se no Apêndice A.

Considerando os parâmetros padrões, o ECOSS buscou apresentar três explicações contrafactuais para cada uma das 40 instâncias. As Tabelas 12 e 13 exemplificam explicações contrafactuais encontradas para 2 instâncias preditas como desempenho Inferior e escolhidas aleatoriamente.

No primeiro caso (Tabela 12), são apresentadas três explicações como formas que

Tabela 11 – Atributos destacados nas explicações

Atributo	Domínio
ADHD (Se o paciente tem TDAH)	Sim(1) ou Não(0)
Friendly_father	Sim(1) ou Não(0)
Mother_age	22...67
Mother_hyperactivity	0...35
Mother_schooling/ Father_schooling	1 a 4 anos(0), 5 a 8 anos(1), Ensino Médio Incompleto(2), Ensino Médio Completo(3), Graduação Incompleta(4), ou Graduação Completa(5)
Raven_Z	-3.58...2.41
School_type	Pública(0) ou Privada(1)
School_year	Educação Infantil(0), Primeiro Ano(1), Segundo Ano(2), Terceiro Ano(3), Quarto Ano(4), Quinto Ano(5), Sexto Ano(6), Sétimo Ano(7), Oitavo Ano(8), Nono Ano(9) ou Ensino Médio(10)
Sex	Masculino(0) ou Feminino(1)
Social_class	Pobre(0), Vulnerável(1), Classe Média(2) ou Classe Alta(3)

Fonte: Elaborado pelo autor.

tornariam uma instância predita como desempenho Inferior em Superior: 1) se a pontuação no teste *Raven_Z* * fosse 0.9672 ao invés de -3.577; ou 2) a escolaridade da mãe fosse graduação completa ao invés de ensino médio completo; ou 3) se o pai fosse amigável.

Tabela 12 – Resultado da aplicação do ECOSS - Instância 1 (Aritmética)

	Modified features		
	Raven_Z	Mother_Schooling	Friendly_father
Original instance: class: Lower	-3.577	3 (Ens. médio completo)	0 (Não)
Counterfactuals class: Higher	0.9672 — —	— 5 (Graduação completa) —	— — 1 (Sim)
No static feature			

Fonte: Dados da Pesquisa.

No segundo caso (Tabela 13), a instância da classe Inferior reverteria sua predição nas seguintes situações: 1) se a pontuação no teste *Raven_Z* fosse 1.423 ao invés de -0.3725; ou 2) se a pontuação no teste *Raven_Z* fosse 1.039 ao invés de -0.3725 e o estudante estivesse em uma escola privada; ou 3) se a pontuação no teste *Raven_Z* fosse 1.039 ao invés de -0.3725 e o estudante fosse de classe social alta.

*O *Raven_Z* se refere a média do desvio padrão do valor obtido no Teste das Matrizes Progressivas de Raven.

Tabela 13 – Resultado da aplicação do ECOSS - Instância 2 (Aritmética)

	Modified features		
	Raven_Z	School	Social_class
Original instance: class: Lower	-0.3725	0 (Pública)	0 (Pobre)
Counterfactuals class: Higher	1.423	_	_
	1.039	1 (Privada)	_
	1.039	_	3 (Classe alta)
No static feature			

Fonte: Dados da Pesquisa.

Sobre a eficácia do método na geração dos contrafactuais para as 40 instâncias, o mesmo obteve:

- 3 explicações para 38 instâncias;
- 2 explicações para 1 instância;
- somente 1 explicação para 1 instância.

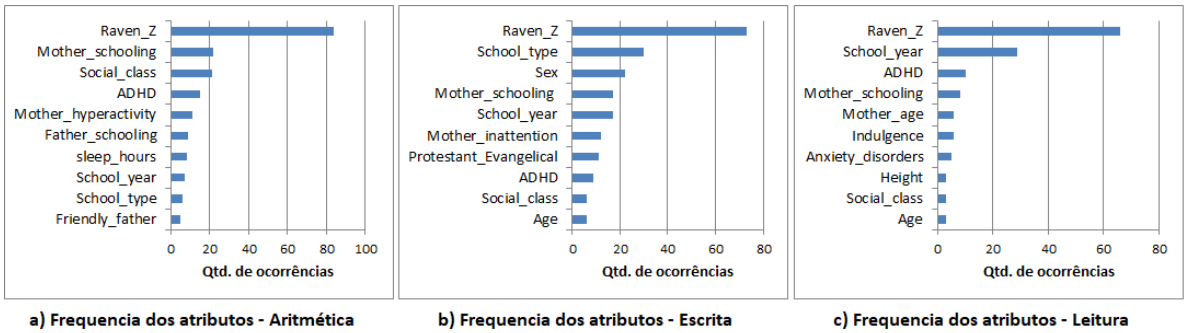
Por esta perspectiva, considera-se que o método obteve um bom desempenho na geração de explicações com o modelo em aritmética, já que em 95% dos casos as três explicações foram obtidas e não houve casos onde ao menos uma explicação não foi apresentada.

A Figura 17(a) apresenta os atributos mais frequentes nas explicações geradas para aritmética. Somando-se os resultados presentes na figura a uma apreciação geral das explicações individuais, é possível expandir a visão sobre o modelo de aritmética.

Percebe-se que o *Raven_Z*[†] se sobressai de forma expressiva, no qual maiores valores no teste Raven_Z tendem a conduzir a um desempenho superior em aritmética. Em seguida, destaque-se influências positivas e significativas das mães de alta escolaridade e famílias de classes sociais mais altas. As explicações confirmam que o fato de o aluno ter TDAH é um complicador para o desempenho escolar em aritmética. Outro atributo de destaque é a hiperatividade da mãe, que apresentou um comportamento particular. No geral, quando avaliados individualmente os atributos apresentam um comportamento linear. No caso da hiperatividade da mãe, percebe-se que valores baixos ou altos para este atributo contribuem para predição do desempenho como Inferior em aritmética. Já valores intermediários favorecem a predição como desempenho Superior na mesma disciplina. Além disso, pais de escolaridade mais alta também contribuem com o desempenho escolar. Outros atributos ainda surgiram, porém em menor frequência.

[†]*Raven_Z* é um teste de inteligência (pontuação de QI) amplamente usado (QIU; HATTON; HOU, 2020).

Figura 17 – Ranking dos atributos mais frequentes nas explicações da base TDAH.



Fonte: Dados da Pesquisa.

Já com relação ao número de trocas necessário para encontrar os três contrafactuais, o ECOSS precisou em média de 1.82 ± 0.81 mudanças nos atributos da instância original, o que evidencia que a classificação do modelo é razoavelmente sensível a alterações de valores, especialmente dos atributos considerados mais importantes.

Enfim, observando as explicações geradas e os demais indicadores calculados, percebe-se uma influência muito destacada na pontuação do *Raven_Z* para a predição em aritmética. Para muitas instância, um valor maior de *Raven_Z* sozinho seria capaz de reverter um desempenho inferior na disciplina, independente do estudante possuir ou não TDAH. Vale salientar ainda, a importância da escolaridade da mãe e da classe social, o que indica que o suporte em torno do aluno é um fator importante no seu desempenho escolar em aritmética.

7.1.2 Resultados obtidos pela aplicação do ECOSS no modelo para escrita

Utilizando o modelo de predição de desempenho acadêmico em escrita com *Random Forest*, novamente aplicou-se o ECOSS para gerar explicações para as predições das 40 instâncias do conjunto de teste/explicação. O tempo médio de execução por instância foi de 24.33 segundos.

Em se tratando da eficácia na obtenção das explicações contrafactuais, o método obteve:

- 3 explicações para 39 instâncias.
- 2 explicação para 1 instância.

O método alcançou sucesso na geração das três explicações em 97.5% dos casos e não houve casos sem apresentar explicação.

A Figura 17(b) apresenta os atributos que foram mais frequentemente usados na

geração dos contrafactuais para as predições em escrita. Novamente, o *Raven_Z* aparece como a característica mais relevante, o qual valores mais altos no teste contribuem para previsão de desempenho superior. O segundo atributo mais frequente foi o tipo de escola, de modo que estudar em escola privada apresenta impacto positivo na performance do aluno. Surge com significativa importância o atributo sexo, no qual segundo as explicações individuais observadas, ser do sexo feminino favorece o desempenho em escrita. Tal qual em aritmética, a escolaridade da mãe aparece com destaque e comportamento semelhante para escrita.

Outro atributo destacado foi o ano escolar, sobre o qual percebe-se que as séries intermediárias, mais especificamente do terceiro ao sétimo ano, tem impacto negativo na performance em escrita. Por outro lado, os primeiros anos (educação infantil, primeiro e segundo anos) e os últimos anos escolares (oitavo, novo anos e ensino médio) tem impacto positivo na performance do discente. Uma possibilidade levantada junto ao especialista para tal comportamento é que nos primeiros anos escolares os conteúdos são mais simples e os recursos existentes de auxílio a aprendizagem estão mais consolidados. A medida que se avança nos anos escolares, a maior complexidade dos conteúdos tendem a explicitar as dificuldades dos estudantes. Porém, aqueles que conseguem superar os anos intermediários tendem a apresentar bom desempenho nos últimos anos. Contudo, pretende-se aprofundar na investigação deste atributo com o auxílio do especialista e sustentação na literatura.

Sobre a sensibilidade do modelo em relação a mudanças nos valores dos atributos, o ECOSS precisou em média 1.78 ± 0.66 mudanças na instância original para gerar os contrafactuais. Novamente, percebe-se que poucas trocas de valores possibilitam reverter as predições.

Em linhas gerais, novamente o *Raven_Z* teve uma importância significativa nas explicações embora com uma frequência menor quando comparado a aritmética. Destaca-se a frequencia relativamente alta do tipo de escola e sexo na geração dos contrafactuais. Outros atributos importantes, mas com menor destaque quando comparados aos anteriormente citados, são a escolaridade da mãe e o ano escolar.

7.1.3 Resultados obtidos pela aplicação do ECOSS no modelo para leitura

Para explicação das decisões do modelo de classificação com *Random Forest* para leitura, novamente aplicou-se o ECOSS para as 35 instâncias destinadas ao conjunto de teste/explicação. O tempo médio de execução por instância foi de 15.67 segundos.

Em se tratando da eficácia na obtenção das explicações contrafactuais, o método obteve:

- 3 explicações para 34 instâncias.

- 2 explicações para 1 instância.

O método alcançou sucesso na geração das três explicações em 97.1 % dos casos e não houve casos sem apresentar explicação.

A Figura 17(c) apresenta os atributos que foram mais frequentes nas explicações para leitura. A pontuação no teste *Raven_Z* foi mais uma vez o atributo mais relevante para o modelo com comportamento semelhante as demais disciplinas, já que altas pontuações no teste indicam melhor desempenho em leitura. O ano escolar é o segundo atributo em importância, sendo que o discente melhora seu desempenho em leitura com o decorrer dos anos. Na sequência, destaca-se o atributo que indica se o estudante possui ou não TDAH, no qual o modelo considera que possuir TDAH é um complicador para o desempenho em leitura. Assim como em aritmética e escrita, a escolaridade da mãe também está frequentemente presente nas explicações e com o mesmo comportamento, no qual mães de alta escolaridade contribuem para o desempenho acadêmico em leitura. Por fim, observa-se a relevância do atributo idade da mãe, o qual, segundo o modelo, ser filho de mães de mais idade impacta positivamente no desempenho do aluno.

Sobre a sensibilidade do modelo em relação a mudanças nos valores dos atributos, o ECOSS precisou em média 1.44 ± 0.57 mudanças na instância original para gerar os contrafactuais. Dos três modelos estudados, o de leitura mostrou-se o mais sensível a troca nos valores dos atributos. Em 60% dos contrafactuais gerados para esse modelo, a troca de um único atributo seria suficiente para mudar a classe da instância.

Em termos gerais, as explicações geradas mostraram que o teste *Raven_Z* e o ano escolar são atributos muito significativos para o desempenho em leitura. Ter ou não TDAH, a escolaridade e idade da mãe também merecem destaque.

7.1.4 Diversidade das explicações

O ECOSS utiliza um AG mono-objetivo para identificar os melhores contrafactuais. No AG mono-objetivo, normalmente as soluções seguintes a melhor são próximas a esta, o que normalmente não é um problema já que nos cenários no qual este tipo de algoritmo é aplicado se está interessado apenas na melhor solução. No entanto, pretende-se que o ECOSS permita ao usuário escolher uma quantidade k de explicações a receber e para que estas sejam úteis necessita-se de alguma diversidade.

Diante do cenário exposto, foi necessário desenvolver um meio que permita ao método, mesmo baseado em um AG mono-objetivo, gerar explicações com diversidade. Resumidamente, a estratégia adotado pelo método foi: exibe-se a melhor solução e as demais $k - 1$ apresentadas são selecionadas, seguindo a ordenação, de forma a conter um conjunto diferente de atributos. Assim, explicações que presume-se não agregar na

diversidade de explicações são descartadas.

A Figura 18 ilustra a estratégia adotada na geração de três explicações ($k = 3$) para uma instância da classe Inferior escolhida aleatoriamente na base de Aritmética. Neste caso, para compreensão da lógica empregada no método, foram apresentadas também as explicações descartadas. Em linhas gerais a estratégia é:

- Exibe-se o melhor contrafactual (C1);
- Qualquer contrafactual C_j gerado a partir dos mesmos atributos de um contrafactual C_i (onde $i < j$) é descartado (C2). No exemplo em questão C1 e C2 utilizam o *Raven_Z*;
- Considere um contrafactual C_i apto (selecionado) e um contrafactual C_j a ser analisado (onde $i < j$) no qual os atributos modificados em C_i para gerar o contrafactual são um subconjunto dos atributos modificados em C_j. Neste caso tem-se duas possibilidades:

  1) casos como em C3 são descartados, pois a solução está contida em uma solução já selecionada (C1). C3 gera um contrafactual com a mesma alteração que C1 acrescido de uma alteração na idade da mãe (*Mother_Age*), que no caso seria desnecessária para gerar um contrafactual;

  2) C4 é selecionado para o usuário, pois embora também contenha *Raven_Z* na solução (como C1), o valor em C4 está mais próximo da instância original que em C1. Assim, o usuário pode optar por alterar apenas um atributo mas com uma distância maior de valores (C1) ou dois atributos mas um deles com uma distância menor de valores (C4);
- Qualquer contrafactual gerado a partir de um conjunto de atributos modificados diferente dos anteriores é considerado útil para o usuário (C5). C5 é gerado por alterações na escolaridade da mãe (*Mother_schooling*) e classe social (*Social_class*), diferente de todos os contrafactuais anteriores.

Consideramos que ao diversificar as explicações, eleva-se a possibilidade que uma delas seja relevante para o usuário, indo ao encontro das diretrizes apontadas por Miller (2019) para explicações que atendam as necessidades dos usuários.

7.2 Resultados do estudo de caso sintomatologia de depressão

Um levantamento realizado em parceria com o Programa de Pós-Graduação em Psicologia: Cognição e Comportamento da Universidade Federal de Minas Gerais deu origem a uma base de dados que contém informações de 377 crianças e adolescentes com diferentes sintomatologias depressivas. Esta foi previamente apresentada na Seção 5.2.2, página 60.

Figura 18 – Exemplo de diversidade das explicações.

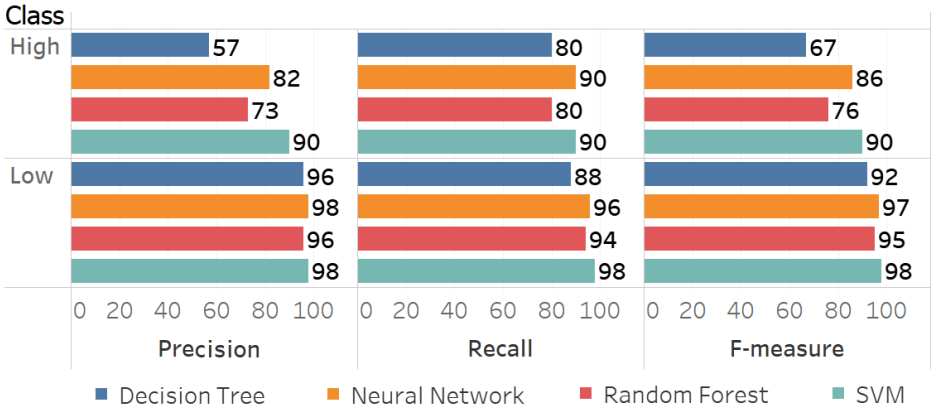
	Modified features				
	Raven_Z	Mother_Age	Mother_schooling	Social_class	
Original instance: class: Lower	-0.3725	37	3 (Ensino Médio)	0 (Pobre)	
Counterfactuals class: 0 (Good)	0.9672	_	_	_	C1
	~~1.089~~	_	_	_	C2
	~~0.9672~~	~~34~~	_	_	C3
	0.3115	_	5 (Graduação)	_	C4
	_	_	5 (Graduação)	3 (Alta)	C5
No static feature					

Fonte: Elaborado pelo autor.

Para obter uma melhor capacidade preditiva foram desenvolvidos modelos baseados em quatro algoritmos de AM: Árvore de Decisão, Redes Neurais, SVM, e *Random Forest*. Como no estudo anterior, propõe-se gerar explicações a partir do modelo de melhor desempenho na etapa de teste.

A Figura 19 apresenta os resultados dos modelos preditivos na fase de teste. Nota-se que para sintomatologia "Low" todos os modelos tiveram um desempenho expressivo, com um desempenho levemente superior para o modelo com SVM com *F-measure* de 98%. Já na predição da sintomatologia "High", o modelo com SVM obteve destacadamente o melhor desempenho com *F-measure* de 90% e por isso aplicou o ECOSS para esse modelo.

Figura 19 – Avaliação da performance dos modelos para a base de Depressão.



Fonte: Dados da Pesquisa.

7.2.1 Aplicação do ECOSS no modelo de Depressão

Nestes experimentos, utilizando o modelo de predição de sintomatologia de depressão com SVM, aplicou-se o ECOSS para gerar explicações para as predições das 60 instâncias selecionadas como conjunto de teste.

O tempo médio de execução por instância foi de 10 minutos e 07 segundos. O tempo de execução para o modelo de depressão foi bastante superior às outras bases. A explicação é que, neste caso, o método fez uso do *Kernel SHAP*, a versão agnóstica do SHAP. Constata-se que as versão específicas do SHAP, como por exemplo *SHAP Tree Explainer*, tem um desempenho bastante superior em termos de tempo de resposta. Percebe-se que o tempo de resposta do SHAP é o principal elemento de influência no tempo de resposta do ECOSS.

Os atributos mencionados na descrição dos experimentos estão listados na Tabela 14, bem como a transformação numérica de seus valores. Vale ressaltar que os valores referentes aos atributos foram normalizados, o que pode dificultar a leitura dos dados. A lista completa de atributos encontra-se no Apêndice B.

Considerando os parâmetros padrões, o ECOSS buscou apresentar três explicações contrafactuais para cada uma das 60 instâncias. As Tabelas 15 e 16 exemplificam explicações contrafactuais encontradas para duas instâncias preditas como sintomatologia "Alta". Propositalmente, optou-se por apresentar as explicações nos quais alguma situação inesperada ocorreu. A intenção é mostrar que as explicações também podem ser utilizadas para identificar algum possível erro na predição ou algo considerado diferente do esperado.

No primeiro caso (Tabela 15), apresenta-se três explicações como formas que tornariam uma instância de sintomatologia "Alta" em "Baixa". Existe em comum às três explicações a necessidade de alterar o CDI20 de 0.5 (Eu me sinto sozinho(a) muitas vezes) para 0 (Eu não me sinto sozinho(a)). Além disso, estas incluem uma segunda alteração que consistiria em: 1) alterar o CDI1 de 0.5 (Eu fico triste muitas vezes) para 0 (Eu fico triste de vez em quando); ou 2) mudar a escolaridade da mãe para graduação completa; ou 3) mudar a situação dos pais de juntos (1) para separados (0). As explicações 1 e 2 são relevantes pois evidenciam que os sentimentos de solidão (CDI20) e tristeza (CDI1) e a escolaridade da mãe são significativos na predição da sintomatologia de depressão. No entanto, a terceira explicação apresentou uma situação que, a princípio, contradiz o senso comum uma vez apresenta que ser filho de pais que estão juntos contribui com a depressão.

As explicações apresentadas no segundo caso também consistem em 3 formas de tornar um indivíduo com sintomatologia para depressão "Alta" em "Baixa". Como pode-se

Tabela 14 – Atributos destacados nos modelos de explicação

Atributo	Domínio
CDI1	Eu fico triste de vez em quando (0) Eu fico triste muitas vezes (0.5) Eu estou sempre triste (1)
CDI3	Eu faço bem a maioria das coisas (0) Eu faço errado a maioria das coisas (0.5) Eu faço tudo errado (1)
CDI8	Normalmente, eu não me sinto culpado(a) pelas coisas ruins que acontecem (0) Muitas coisas ruins que acontecem são por minha culpa (0.5) Tudo de mal que acontece é por minha culpa (1)
CDI11	Eu me sinto preocupado(a) de vez em quando (0) Eu me sinto preocupado(a) frequentemente (0.5) Eu me sinto sempre preocupado(a) (1)
CDI14	Eu tenho boa aparência (0) Minha aparência tem alguns aspectos negativos (0.5) Eu sou feio(a) (1)
CDI15	Fazer os deveres de casa não é um grande problema para mim (0) Com frequência eu tenho que ser pressionado(a) para fazer os deveres de casa (0.5) Eu tenho que me obrigar a fazer os deveres de casa (1)
CDI20	Eu não me sinto sozinho(a) (0) Eu me sinto sozinho(a) muitas vezes (0.5) Eu sempre me sinto sozinho(a) (1)
CDI25	Eu tenho certeza que sou amado(a) por alguém (0) Eu não tenho certeza se alguém me ama (0.5) Ninguém gosta de mim realmente (1)
CDI26	Eu costumo fazer o que me mandam (0) Eu não faço o que me mandam com frequência (0.5) Eu nunca faço o que me mandam (1)
Genre	Masculino (0), Feminino (1)
Mother_Complete_Graduation	Não (0), Sim (1)
Parents	Separados/Divorciados (0), Juntos (1)
Time_with_Mother (hours per week)	[0 a 3 horas] (0), [4 a 7 horas] (1), [8 a 11 horas] (2) [12 a 15 horas] (3), [16 a 19 horas] (4), [20 ou mais] (5)

Fonte: Elaborado pelo autor.

perceber na Tabela 16, em linhas gerais, as explicações combinam mudanças em atributos que demonstram como positivo pertencer ao gênero masculino, uma postura de obediência (CDI26), ser filho(a) de mulheres de alta escolaridade (graduação completa) e não possuir níveis altos de preocupação (CDI11). Contudo, surpreende o comportamento do atributo relacionado ao tempo com a mãe por semana (*Time_with_mother*). Segundo a explicação, ter menos tempo com a mãe contribui com a baixa sintomatologia para

Tabela 15 – Resultado da aplicação do ECOSS - Instância 1

	Modified features			
	CDI20	CDI1	Mother_Complete_Graduation	Parents
Original instance: class: High	0.5	0.5	0 (Não)	1 (Juntos)
Counterfactuals class: Low	0	0	_	_
	0	_	1 (Sim)	_
	0	_	_	0 (Separados)
No static feature				

Fonte: Dados da Pesquisa.

depressão.

Tabela 16 – Resultado da aplicação do ECOSS - Instância 2

	Modified features				
	Gender	Time_with_Mother	CDI26	Mother_Complete_Graduation	CDI11
Original instance: class: High	1 (Feminino)	1 (4 a 7 horas)	1	0 (Não)	1
Counterfactuals class: Low	0 (Masculino)	0 (0 a 3 horas)	0	1 (Sim)	_
	0 (Masculino)	_	0	1 (Sim)	0
	0 (Masculino)	0 (0 a 3 horas)	_	1 (Sim)	0
No static feature					

Fonte: Dados da Pesquisa.

Nas situações descritas anteriormente, deparou-se com um comportamento do modelo contrário ao que se espera. Neste caso, é preciso aprofundar na referida questão, por vezes com o auxílio de um especialista, pois: 1) pode-se identificar um erro de predição; ou 2) pode-se estar diante de uma descoberta. A se confirmar a identificação de um erro de predição, exemplifica-se um papel fundamental da interpretabilidade, em um primeiro momento de interesse dos desenvolvedores. Em caso de uma descoberta, se evidencia outra face de destaque da interpretabilidade de utilidade especial para pesquisadores e outros envolvidos no contexto do problema.

Em relação à eficácia do método para geração dos contrafactuais para as 60 instâncias de teste, o mesmo obteve:

- 3 explicações para 58 instâncias;
- somente 1 explicação para 1 instância;
- não obteve explicações para 1 instância.

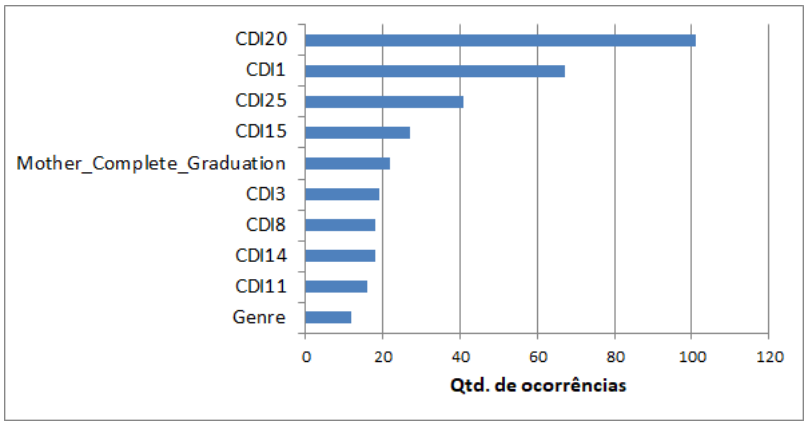
Novamente o método obteve um bom desempenho na geração de explicações, já que em 96,6% dos casos as três explicações foram obtidas. Contudo, houve um caso onde

nenhuma explicação foi apresentada.

Em um segundo momento, considerando que possivelmente para essa instância não se encontrou solução em virtude da necessidade de uma quantidade maior de mudanças na instância original para encontrar o exemplo contrafactual, uma nova execução foi realizada aumentando o número de indivíduos (de 100 para 150) e a quantidade de gerações (de 30 para 40) do AG. De fato, com essas mudanças nos parâmetros do método, encontrou-se os três casos contrafactuais desejados.

A Figura 20 apresenta os atributos mais frequentes nas explicações geradas para as decisões do modelo. Considerando os resultados apresentados na figura e observações realizadas nas explicações individuais, algumas considerações gerais são possíveis.

Figura 20 – Ranking dos atributos mais frequentes nas explicações da base de Depressão.



Fonte: Dados da Pesquisa.

É notável que o CDI20 é o atributo de influência mais significativa na previsão da sintomatologia de um indivíduo para depressão. Logo, para o modelo, o sentimento de solidão é o que mais evidencia a depressão.

Outros CDIs aparecem na sequência, como o CDI11 que aponta níveis excessivos de tristeza como um fator de influência destacada para a predição como sintomatologia "Alta". O CDI25 está relacionado ao sentimento exacerbado de rejeição do indivíduo propenso à depressão. O CDI15 destaca a falta de motivação com atividades escolares como outro fator de influência destacada para a sintomatologia "Alta".

As explicações apontam ainda uma relação entre a escolaridade da mãe e a predição de sintomatologia das crianças e adolescentes. Nota-se um impacto positivo para mães com graduação completa. Entende-se que esse atributo está relacionado a um suporte maior que estes indivíduos possam receber das mães. É importante salientar que a influência da escolaridade do pai na situação de depressão das crianças/adolescentes não foi observada nos experimentos realizados.

Sobre a influência da escolaridade da mãe, Park, Fuhrer e Quesnel-Vallée (2013) obtiveram resultados similares. Os autores concluíram que filhos de mulheres que não terminaram o ensino médio têm duas vezes mais chances de sofrer um episódio grave de depressão no início da vida adulta do que crianças cujas mães alcançaram diploma universitário/ensino médio. Curiosamente, neste trabalho, os autores também observaram que o nível de escolaridade do pai não teve impacto na depressão dos filhos. Segundo os autores, uma mãe com melhor nível de escolaridade pode ter mais confiança para lidar com as dificuldades decorrentes da criação dos filhos. Essa maior confiança e senso de autodomínio podem servir como um modelo para seus filhos.

Ainda sobre os atributos mais frequentes nas explicações, outros CDIs surgiram, porém em menor frequência. Finalmente, a se destacar o atributo gênero, no qual pertencer ao gênero masculino tem impacto positivo para sintomatologia "Baixa".

Sobre a sensibilidade do modelo em relação a mudanças nos valores dos atributos, o ECOSS precisou em média 2.35 ± 1.04 mudanças na instância original para gerar os contrafactuais. Comparado aos outros experimentos realizados, este é o modelo menos sensível a mudanças nos valores da instância original. Em outras palavras, dos modelos estudados, este é o modelo que em média necessita de mais alterações na instância original para gerar contrafactuais. Apenas 15% dos contrafactuais foram gerados com mudança de valor em um único atributo. Ainda que o CDI20 e CDI1 sejam muito frequentes nas explicações, na maioria dos casos foi necessário combinar alterações nestes atributos com outras mudanças na instância original.

Acreditamos que os resultados alcançados nesta pesquisa podem auxiliar familiares, educadores e profissionais de saúde a identificar e encaminhar o tratamento de crianças e adolescentes com depressão, sendo esta uma contribuição importante diante da gravidade e a quantidade de pessoas acometidas pelo transtorno.

8 CONSIDERAÇÕES FINAIS

Pesquisas em AM têm apontado a necessidade de avanços no campo da interpretabilidade de maneira que os usuários possam compreender melhor as decisões dos modelos e consequentemente tenham maior confiança para utilizá-los. Segundo Miller (2019), isso implica em considerar elementos que vão além das questões computacionais para entender como as pessoas melhor recebem explicações. Sendo assim, baseados em estudos provenientes especialmente das ciências sociais, o autor aponta três aspectos a serem considerados: explicações devem ser contrafactuais, selecionadas e sociais.

Consideramos como hipótese para o trabalho que é possível incluir em um método agnóstico recursos que atendam as três diretrizes apontadas por Miller (2019) para uma explicação orientada ao usuário (Hipótese 1). Adicionalmente, gerar explicações contrafactuais implica em encontrar as mudanças mínimas na instância original que permitam inverter a classe da mesma. Para tal, apontamos como hipótese que a ação conjunta de AG e abordagem SHAP é capaz de gerar contrafactuais tão próximos quanto possível da instância original (Hipótese 2). A Hipótese 1 foi confirmada por meio dos recursos desenvolvidos para o método descritos na Seção 4.3, página 50. Já a Hipótese 2 se mostrou verdadeira por meio dos resultados dos experimentos realizados com o método proposto apresentados na Seção 6.2, página 65.

Diante disso, este trabalho propõe o ECOSS, um método que inclui recursos para que as explicações locais geradas estejam alinhadas às diretrizes destacados por Miller (2019). Acredita-se que ao atender tais diretrizes, o método proporciona uma contribuição importante para o campo da interpretabilidade de modelos de AM, pois apresenta uma abordagem voltada ao usuário final, especialmente não especialistas em AM. Outra vantagem é o fato do método ser agnóstico e, portanto, aplicável a qualquer modelo de classificação.

Este trabalho envolveu ainda dois estudos de caso, desenvolvidos com uma instituição parceira, que incluíram a construção de modelos de classificação e a interpretabilidade dos mesmos, sendo este último o foco maior da pesquisa. No estudo de caso relativo a predição do desempenho acadêmico de crianças e adolescentes com TDAH foi desenvolvido um modelo baseado em *Random Forest*, cuja aplicação do ECOSS culminou em um instrumento de auxílio aos pais, educadores e outros profissionais envolvidos na busca por melhores resultados acadêmicos em aritmética, escrita e leitura para os discentes com o transtorno. Os estudos realizados permitiram ainda aprofundar no entendimento do

comportamento do modelo e dos principais atributos utilizados nas decisões.

O segundo estudo consistiu na predição de sintomatologias depressivas em crianças e adolescentes. Os resultados alcançados por meio do modelo de classificação desenvolvido utilizando SVM somado a interpretabilidade gerada pela aplicação do ECOSS, podem auxiliar na identificação de depressão, permitem entender as características mais relevantes nesta predição e ainda levantar possíveis desvios no comportamento do modelo. Logo, o trabalho constitui um importante aparato tanto para os desenvolvedores do modelo quanto para os familiares, profissionais de saúde e demais pessoas próximas as crianças e adolescentes que potencialmente podem ser acometidos pelo transtorno.

Nos experimentos realizados com o ECOSS que visaram compará-lo a outros métodos, estes foram avaliados quanto a capacidade de gerar contrafactuais tão próximos quanto possível da instância original. Nestes experimentos, o ECOSS apresentou um desempenho geral superior a outros métodos presentes na literatura. Entende-se que tal performance foi atingida especialmente pela união da capacidade de explorar o espaço de busca presentes nos AGs com a possibilidade de avaliar a proximidade dos indivíduos para a classe desejada proporcionada pelo SHAP.

Esta capacidade de busca de soluções proveniente do uso combinado do SHAP e do AG é um diferencial do método proposto. Assim, em bases de dados com maior número de atributos de entrada ou cenários que exijam uma maior quantidade de mudanças na instância original para encontrar o contrafactual, acredita-se que o ECOSS apresentará um desempenho ainda mais destacado quando comparado a outros métodos. Logo, como trabalho futuro, sugere-se realizar outros experimentos que confirmem tal hipótese.

Uma limitação do ECOSS está relacionada a possibilidade de as mudanças nos atributos indicados nas explicações gerarem uma instância inconsistente para o mundo real. Neste sentido, outra possibilidade de trabalho futuro é permitir que o usuário crie uma lista de restrições de consistência ou criar uma forma de identificar automaticamente tais inconsistências. Esta última seria ainda mais interessante para retirar tal incumbência do usuário.

Por fim, indica-se a realização de experimentos com a participação dos usuários finais visando uma avaliação das explicações geradas e dos recursos disponíveis no método.

9 CRONOGRAMA

Segundo o Regulamento do Programa de Pós-Graduação em Informática, os requisitos para obtenção do título de Doutor são:

1. 34 (trinta e quatro) créditos em disciplinas;
2. Publicação e apresentação de artigo em conferência que seja um produto ou subproduto da tese;
3. Publicação em periódico no estrato superior;
4. Estágio em docência;
5. Aprovação em exame de qualificação;
6. Aprovação na defesa da tese.

Sobre os requisitos obrigatórios do programa citados anteriormente, já foram cumpridos os itens 1, 2 e 4. Em relação aos artigos decorrentes da tese, os seguintes artigos foram publicados em conferências (Tabela 17):

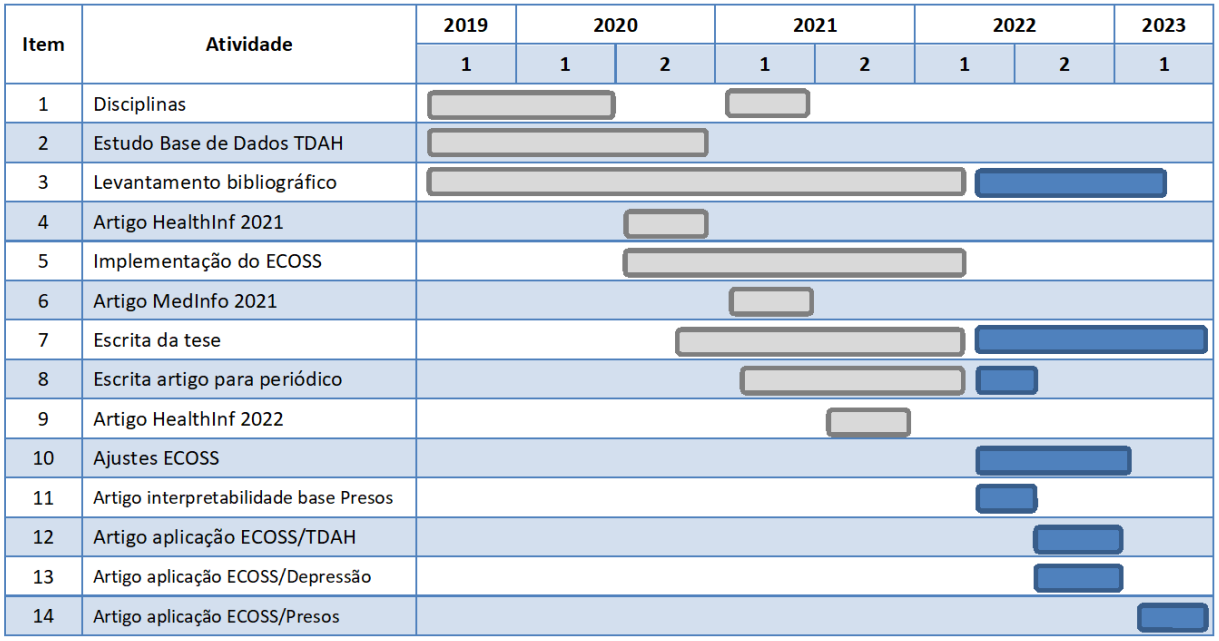
A Figura 21 apresenta o cronograma com as atividades já executadas e aquelas que se pretende cumprir até a defesa da tese. Dentre os itens apresentados, ressalta-se que o artigo para periódico citado (item 8) refere-se a um artigo de apresentação do método ECOSS proposto neste trabalho. Uma vez que o método for publicado, pretende-se escrever três artigos com a aplicação do mesmo nas bases de TDAH (item 12), Depressão

Tabela 17 – Artigos publicados

Título	Ano	Evento	Qualis
Analysis of School Performance of Children and Adolescents with Attention-Deficit/Hyperactivity Disorder: A Dimensionality Reduction Approach	2021	HEALTHINF 2021, 14th International Joint Conference on Biomedical Engineering Systems and Technologies	A3
Predictions of Academic Performance of Children and Adolescents with ADHD using the SHAP Approach	2021	MedInfo 2021, 18th World Congress on Medical and Health Informatics	A3
Predicting Depression in Children and Adolescents Using the SHAP Approach	2022	HEALTHINF 2022, 15th International Joint Conference on Biomedical Engineering Systems and Technologies	A3

(item 13) e Presos (item 14) com as quais já foram produzidos artigos relacionados a interpretabilidade usando a abordagem SHAP (itens 4, 6, 9 e 11). O artigo com a base de Presos (item 11) está em andamento.

Figura 21 – Cronograma de atividades.



Fonte: Elaborado pelo autor.

Sobre os ajustes mencionados para o ECOSS (item 10), pretendemos testar o método em cenários diversos, especialmente em bases de dados com um número maior de atributos, gerando gráficos e outras formas de visualização que permitam avaliar o comportamento do método. Consideramos ainda aprimorar o formato da saída exibida ao usuário e buscar alternativas para reduzir o tempo de execução do método.

REFERÊNCIAS

ABDUL, A. et al. Trends and trajectories for explainable, accountable and intelligible systems: An hci research agenda. In: Proceedings of the 2018 CHI conference on human factors in computing systems. New York, NY, USA: Association for Computing Machinery, 2018. p. 1–18. Disponível em: <https://doi.org/10.1145/3287560.3287574>.

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.

ANDREWS, R.; DIEDERICH, J.; TICKLE, A. B. Survey and critique of techniques for extracting rules from trained artificial neural networks. Knowledge-based systems, Elsevier, v. 8, n. 6, p. 373–389, 1995.

APA. Depression. American Psychiatric Association, 2017. `https://www.psychiatry.org/psychiatrists/practice/quality-improvement/quality-measures-for-mips-quality-category`.

APA, A. P. A. et al. Diagnostic and statistical manual of mental disorders (DSM-5®). Washington, D.C: American Psychiatric Pub, 2013.

ASSOCIATION, A. P. et al. DSM-5: Manual diagnóstico e estatístico de transtornos mentais. Artmed Editora, 2014. Disponível em: <http://www.niip.com.br/wp-content/uploads/2018/06/Manual-Diagnosico-e-Estatistico-de-Transtornos-Mentais-DSM-5-1-pdf>.

BALBINO., M. et al. Predicting depression in children and adolescents using the shap approach. In: INSTICC. Proceedings of the 15th International Joint Conference on Biomedical Engineering Systems and Technologies - HEALTHINF,. SciTePress, 2022. p. 514–521. ISBN 978-989-758-552-4. Disponível em: <https://doi.org/10.5220/0010842500003123>.

BERNARAS, E.; JAUREGUIZAR, J.; GARAIGORDOBIL, M. Child and adolescent depression: a review of theories, evaluation instruments, prevention programs and treatments. Frontiers in Psychology, Frontiers, v. 10, p. 543, 2019.

BIRAN, O.; COTTON, C. Explanation and justification in machine learning: A survey. In: IJCAI-17 workshop on explainable AI (XAI). [s.n.], 2017. v. 8, n. 1, p. 8–13. Disponível em: <http://www.cs.columbia.edu/~orb/papers/xai_survey_paper_2017.pdf>.

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.

CHEN, Z.-Y. et al. Evolutionary feature and instance selection for traffic sign recognition. Computers in Industry, Elsevier, v. 74, p. 201–211, 2015.

DERRAC, J.; GARCÍA, S.; HERRERA, F. A survey on evolutionary instance selection and generation. In: Modeling, Analysis, and Applications in Metaheuristic Computing: Advancements and Trends. IGI Global, 2012. p. 233–266. Disponível em: <https://doi.org/10.4018/jamc.2010102604>.

DHURANDHAR, A. et al. Explanations based on the missing: Towards contrastive explanations with pertinent negatives. arXiv preprint arXiv:1802.07623, 2018.

DU, M.; LIU, N.; HU, X. Techniques for interpretable machine learning. Communications of the ACM, ACM New York, NY, USA, v. 63, n. 1, p. 68–77, 2019.

El Shawi, R. et al. Interpretability in healthcare a comparative study of local machine learning interpretability techniques. In: 2019 IEEE 32nd International Symposium on Computer-Based Medical Systems (CBMS). IEEE, 2019. p. 275–280. Disponível em: <https://ieeexplore.ieee.org/document/8787506>.

FRAZIER, T. W. et al. Adhd and achievement: Meta-analysis of the child, adolescent, and adult literatures and a concomitant study with college students. Journal of learning disabilities, Sage Publications Sage CA: Los Angeles, CA, v. 40, n. 1, p. 49–65, 2007.

GARCÍA, S.; HERRERA, F. Evolutionary undersampling for classification with imbalanced datasets: Proposals and taxonomy. Evolutionary computation, MIT Press One Rogers Street, Cambridge, MA 02142-1209, USA journals-info . . . , v. 17, n. 3, p. 275–306, 2009.

GILPIN, L. H. et al. Explaining explanations: An overview of interpretability of machine learning. In: IEEE. 2018 IEEE 5th International Conference on data science and advanced analytics (DSAA). 2018. p. 80–89. Disponível em: <https://doi.org/10.48550/arXiv.1806.00069>.

Guidotti, R. et al. Factual and counterfactual explanations for black box decision making. IEEE Intelligent Systems, v. 34, n. 6, p. 14–23, 2019.

GUIDOTTI, R. et al. 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.

HANSON, N. R. Patterns of discovery: An inquiry into the conceptual foundations of science. CUP Archive, 1965. Disponível em: <https://books.google.com.br/books?id=XD44AAAAIAAJ>.

HAUPT, R. L.; HAUPT, S. E. Practical genetic algorithms. Wiley Online Library, 2004.

HERLOCKER, J. L.; KONSTAN, J. A.; RIEDL, J. Explaining collaborative filtering recommendations. In: Proceedings of the 2000 ACM conference on Computer supported cooperative work. New York, NY, USA: Association for Computing Machinery, 2000. p. 241–250. Disponível em: <https://doi.org/10.1145/358916.358995>.

HOFMANN, H. Statlog (german credit data) data set. UCI Repository of Machine Learning Databases, v. 53, 1994. Disponível em: <https://archive.ics.uci.edu/ml/datasets/statlog+(german+credit+data)>.

HONEGGER, M. Shedding light on black box machine learning algorithms: Development of an axiomatic framework to assess the quality of methods that explain individual predictions. arXiv preprint arXiv:1808.05054, 2018.

JANDRE, C. et al. Analysis of school performance of children and adolescents with attention-deficit/hyperactivity disorder: A dimensionality reduction approach. In: HEALTHINF. [s.n.], 2021. p. 155–165. Disponível em: <https://www.scitepress.org/Papers/2021/102404/102404.pdf>.

JANDRE, C. R. da S. O uso de personas e aprendizado de máquina na análise do perfil de crianças e adolescentes com transtorno de déficit de atenção/hiperatividade. 2021. Disponível em: <https://search.ebscohost.com/login.aspx?direct=truedb=cat06909aAN=sib.540516lang=pt-brsite=eds-live>.

JÚNIOR, É. d. B. R.; LOOS, H. Escola e desenvolvimento psicossocial segundo percepções de jovens com tdah. Paidéia (Ribeirão Preto), SciELO Brasil, v. 21, p. 373–382, 2011.

KARIM, A. et al. Machine learning interpretability: A science rather than a tool. arXiv preprint arXiv:1807.06722, 2018.

KHADEMI, M.; NEDIALKOV, N. S. Probabilistic graphical models and deep belief networks for prognosis of breast cancer. In: IEEE. Machine Learning and Applications (ICMLA), 2015 IEEE 14th International Conference on. Miami, FL, 2015. p. 727–732.

KIM, K.-j. Artificial neural networks with evolutionary instance selection for financial forecasting. Expert Systems with Applications, Elsevier, v. 30, n. 3, p. 519–526, 2006.

KMENT, B. Counterfactuals and Explanation. Mind, v. 115, n. 458, p. 261–310, 04 2006. ISSN 0026-4423. Disponível em: <https://doi.org/10.1093/mind/fzl261>.

KOVACS, M. Children's Depression Inventory (CDI): Technical Manual Update. Multi-Health Systems, Incorporated, 2003. Disponível em: <https://books.google.com.br/books?id=fZN5tAEACAAJ>.

LACERDA-PINHEIRO, S. F. et al. Are there depression and anxiety genetic markers and mutations? a systematic review. Journal of affective disorders, Elsevier, v. 168, p. 387–398, 2014.

LARSON, J. et al. How we analyzed the compas recidivism algorithm. 2016. Disponível em: <https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm>.

LIM, T. Y. Structured population genetic algorithms: a literature survey. Artificial Intelligence Review, Springer, v. 41, n. 3, p. 385–399, 2014.

LINDEN, R. Algoritmos genéticos: Teoria e implementação. Editora Ciência Moderna, 3ª Edição, 2005.

LINDEN, R. Algoritmos genéticos (2a ediçao). [S.l.]: Brasport, 2008.

LIPTON, Z. C. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. Queue, ACM New York, NY, USA, v. 16, n. 3, p. 31–57, 2018.

LUNDBERG, S. M.; LEE, S.-I. A unified approach to interpreting model predictions. In: GUYON, I. et al. (Ed.). Advances in Neural Information Processing Systems. Curran Associates, Inc., 2017. v. 30, p. 4765–4774. Disponível em: <https://proceedings.neurips.cc/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf>.

MANGALATHU, S.; HWANG, S.-H.; JEON, J.-S. Failure mode and effects analysis of rc members based on machine-learning-based shapley additive explanations (shap) approach. Engineering Structures, Elsevier, v. 219, p. 110927, 2020.

MARTENS, D. et al. Performance of classification models from a user perspective. Decision Support Systems, Elsevier, v. 51, n. 4, p. 782–793, 2011.

MILLER, T. Explanation in artificial intelligence: Insights from the social sciences. Artificial Intelligence, Elsevier, v. 267, p. 1–38, 2019.

MITTELSTADT, B.; RUSSELL, C.; WACHTER, S. Explaining explanations in ai. In: Proceedings of the Conference on Fairness, Accountability, and Transparency. New York, NY, USA: Association for Computing Machinery, 2019. (FAT* '19), p. 279–288. ISBN 9781450361255. Disponível em: <https://doi.org/10.1145/3287560.3287574>.

MOKHTARI, K. E.; HIGDON, B. P.; BAsAR, A. Interpreting financial time series with shap values. In: Proceedings of the 29th Annual International Conference on Computer Science and Software Engineering. USA: IBM Corp., 2019. (CASCON '19), p. 166–172.

MOLNAR, C. Interpretable machine learning. Lulu.com, 2020. Disponível em: <https://christophm.github.io/interpretable-ml-book/>.

MONTAVON, G. et al. Explaining nonlinear classification decisions with deep taylor decomposition. Pattern Recognition, Elsevier, v. 65, p. 211–222, 2017.

MOTHILAL, R. K.; SHARMA, A.; TAN, C. Explaining machine learning classifiers through diverse counterfactual explanations. In: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. New York, NY, USA: Association for Computing Machinery, 2020. (FAT* '20), p. 607–617. ISBN 9781450369367. Disponível em: <https://doi.org/10.1145/3351095.3372850>.

PARK, A. L.; FUHRER, R.; QUESNEL-VALLÉE, A. Parents' education and the risk of major depression in early adulthood. Social Psychiatry and Psychiatric Epidemiology, Springer Science and Business Media LLC, v. 48, n. 11, p. 1829–1839, maio 2013. Disponível em: <https://doi.org/10.1007/s00127-013-0697-8>.

PATRO, S.; SAHU, K. K. Normalization: A preprocessing stage. arXiv preprint arXiv:1503.06462, 2015.

PAVLOVA, B.; UHER, R. Assessment of psychopathology: Is asking questions good enough? JAMA psychiatry, American Medical Association, v. 77, n. 6, p. 557–558, 2020.

POULIN, B. et al. Visual explanation of evidence with additive classifiers. In: MENLO PARK, CA; CAMBRIDGE, MA; LONDON; AAAI PRESS; MIT PRESS; 1999. Proceedings Of The National Conference On Artificial Intelligence. 2006. v. 21, n. 2, p. 1822. Disponível em: <https://www.aaai.org/Papers/IAAI/2006/IAAI06-018.pdf>.

QIU, C.; HATTON, R.; HOU, M. Variations in raven's progressive matrices scores among chinese children and adolescents. Personality and Individual Differences, Elsevier, v. 164, p. 110064, 2020.

QUEVEDO, J.; NARDI, A. E.; SILVA, A. G. da. Depressão-: Teoria e Clínica. Brasil: Artmed Editora, 2018.

RANI, S.; SURI, B.; GOYAL, R. On the effectiveness of using elitist genetic algorithm in mutation testing. Symmetry, Multidisciplinary Digital Publishing Institute, v. 11, n. 9, p. 1145, 2019.

RATHI, S. Generating counterfactual and contrastive explanations using SHAP. arXiv preprint arXiv:1906.09293, 2019.

RAVÌ, D. et al. Deep learning for health informatics. IEEE journal of biomedical and health informatics, IEEE, v. 21, n. 1, p. 4–21, 2017.

RAYMER, M. L. et al. Dimensionality reduction using genetic algorithms. IEEE transactions on evolutionary computation, IEEE, v. 4, n. 2, p. 164–171, 2000.

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. New York, NY, USA: Association for Computing Machinery, 2016. (KDD '16), p. 1135–1144. ISBN 9781450342322. Disponível em: <https://doi.org/10.1145/2939672.2939778>.

RIBEIRO, M. T.; SINGH, S.; GUESTRIN, C. Anchors: High-precision model-agnostic explanations. In: Proceedings of the AAAI conference on artificial intelligence. [s.n.], 2018. v. 32, n. 1. Disponível em: <https://ojs.aaai.org/index.php/AAAI/article/view/11491>.

ROCHA, M. M. d.; ARAúJO, L. G. d. S.; SILVARES, E. F. d. M. Um estudo comparativo entre duas traduções brasileiras do inventário de auto-avaliação para jovens (ysr). Psicologia: teoria e prática, scielopepsic, v. 10, p. 14 – 24, 06 2008. ISSN 1516-3687. Disponível em: <http://pepsic.bvsalud.org/scielo.php?script=sci_arttext&pid=S1516-36872008000100002&nrm=iso>.

RUDIN, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence, Nature Publishing Group, v. 1, n. 5, p. 206–215, 2019.

RÜPING, S. Learning interpretable models. 2006. Tese (Doutorado).

SANTANA, R. et al. Genetic algorithms for feature selection in the children and adolescents depression context. In: IEEE. 2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA). 2019. p. 1470–1475. Disponível em: <https://ieeexplore.ieee.org/abstract/document/8999209>.

SILVA, W. et al. Towards complementary explanations using deep neural networks. In: Understanding and Interpreting Machine Learning in Medical Image Computing Applications. Springer, 2018. p. 133–140. Disponível em: <https://link.springer.com/chapter/10.1007/978-3-030-02628-8_15>.

SINGH, S.; RIBEIRO, M. T.; GUESTRIN, C. Programs as black-box explanations. arXiv preprint arXiv:1611.07579, 2016.

SINGHAL, S.; JENA, M. A study on weka tool for data preprocessing, classification and clustering. International Journal of Innovative technology and exploring engineering (IJItee), v. 2, n. 6, p. 250–253, 2013.

STEIN, L. M. TDE: teste de desempenho escolar: manual para aplicação e interpretação. [S.l.]: Casa do Psicólogo, São Paulo, 1994. 1–17 p.

STEPIN, I. et al. A survey of contrastive and counterfactual explanation generation methods for explainable artificial intelligence. IEEE Access, IEEE, v. 9, p. 11974–12001, 2021.

STRUMBELJ, E.; KONONENKO, I. An efficient explanation of individual classifications using game theory. The Journal of Machine Learning Research, JMLR. org, v. 11, p. 1–18, 2010.

SWARTOUT, W. R. Xplain: A system for creating and explaining expert consulting programs. Artificial intelligence, Elsevier, v. 21, n. 3, p. 285–325, 1983.

Tjoa, E.; Guan, C. A survey on explainable artificial intelligence (xai): Toward medical xai. IEEE Transactions on Neural Networks and Learning Systems, p. 1–21, 2020.

TSAI, C.-F.; EBERLE, W.; CHU, C.-Y. Genetic algorithms in feature and instance selection. Knowledge-Based Systems, Elsevier, v. 39, p. 240–247, 2013.

VERGARA, S. C. Projetos e relatórios de pesquisa. São Paulo: Atlas, 2006.

VERSIANI, M.; REIS, R.; FIGUEIRA, I. Diagnóstico do transtorno depressivo na infância e adolescência. J. bras. psiquiatr, v. 49, n. 10/12, p. 367–82, 2000.

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.

WHO. Depression. World Health Organization, 2018. http://www.who.int/mediacentre/factsheets/fs369/en/.

WHO, W. H. O. Depression and other common mental disorders: global health estimates. 2017. http://www.who.int/iris/handle/10665/254610.

YOON, S.; TAHA, B.; BAKKEN, S. Using a data mining approach to discover behavior correlates of chronic disease: A case study of depression. In: Studies in Health Technology and Informatics. NIH Public Access, 2014. v. 201, p. 71–78. ISBN 9781614994145. ISSN 18798365. Disponível em: <http://www.ncbi.nlm.nih.gov/pubmed/24943527 http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4580372>.

ZENG, W.; DAVOODI, A.; TOPALOGLU, R. O. Explainable DRC hotspot prediction with random forest and SHAP tree explainer. In: IEEE. 2020 Design, Automation & Test in Europe Conference & Exhibition (DATE). 2020. p. 1151–1156. Disponível em: <https://ieeexplore.ieee.org/abstract/document/9116488>.

APÊNDICE A – DESCRIÇÃO DA BASE DE DADOS COMPAS

A Tabela 18 apresenta a lista dos atributos da base de dados Compas.

Tabela 18 – Atributos da base Compas

Atributo	Descrição	Domínio
age	Idade	18..96
age_cat	Faixa etária	Less than 25(0), 25 - 45(1), Greater than 45(2)
sex	Gênero	Male, Female
race	Raça	African-American, Asian, Caucasian, Hispanic, Native American, Other
priors_count	Contagem de antecedentes	0..38
days_b_screening_arrest	Quantidade de dias entre o crime e a triagem no Compas	-414..1057
c_charge_degree	Grau da acusação	F, M, O
is_recid	É reincidente	0,1
is_violent_recid	Reincidência violenta	0,1
two_year_recid	Reincidência em 2 anos	0,1
length_of_stay	Tempo de prisão (dias)	

Fonte: Elaborado pelo autor.

APÊNDICE B – DESCRIÇÃO DA BASE DE DADOS GERMAN

A Tabela 19 apresenta a lista dos atributos da base de dados German.

Tabela 19 – Atributos da base German

Atributo	Descrição	Domínio
age	Idade	19..75
personal_status_sex	Estado Civil/Gênero	masc.:divorciado/separado(1), fem.: divorciada/separada/casada (2), masc.: solteiro (3), masc.: casado/viúvo (4), fem.: solteira (5)
present_res_since	Duração residência atual	1..4
people_under_maintenance	Número de pessoas responsáveis por fornecer manutenção	1..2
installment_as_income_perc	Taxa de parcelamento (em percentual do rendimento disponível)	1..4
other_debtors	Outros devedores/fiadores	Nenhum (1), co-requerente (2), fiador (3)
property	Propriedade	Imobiliária (1) Se não (1), construção de acordo de poupança da sociedade / Seguro de vida (2) Se não (1) e (2), Carro ou outro (não no atributo "savings") (3) Sem propriedade (4)
housing	Habitação	Aluguel (0), Própria (1), Gratuita (2)

telephone	Telefone	Nenhum (0), Registrado no cadastro do cliente (1)
job	Emprego	Desempregado/não qualificado - não residente (0), Não qualificado (1), Funcionário qualificado/oficial (2), Gestor/Autônomo/Funcionário altamente qualificado/Policial (3)
present_emp_since	Duração do emprego atual	Desempregado (1), Tempo < 1 ano (2), 1 <= Tempo < 4 anos (3), 4 <= Tempo < 7 anos (4), Tempo >= 7 anos (5)
foreign_worker	trabalhador estrangeiro	Sim (1), Não (2)
account_check_status	Situação da conta corrente existente	... < 0 DM (1), 0 <= ... < 200 DM (2), ... >= 200 DM (3), sem conta corrente (4)
credit_history	Histórico de crédito	Não possui/Todos os créditos pagos devidamente (0), Todos os créditos neste banco foram devidamente pagos (1), Créditos existentes pagos devidamente até agora (2), Atraso no pagamento no passado (3), Conta crítica/Créditos existentes em outros bancos (4)

savings	Situação da Conta poupança/títulos	... < 100 DM (1), 100 <= ... < 500 DM (2), 500 <= ... < 1000 DM (3), .. >= 1000 DM (4), Desconhecido/sem conta poupança(5)
credits_this_2	Número de créditos existentes	1..4
duration_in_month	Duração do crédito em meses	4..72
credit_amount	Valor Total do crédito	250 DM .. 18.424 DM
purpose	Motivo do crédito	Carro novo (0), Carro usado (1), Móveis/Equipamentos (2), Rádio / Televisão (3), Eletrodomésticos (4), Reparos (5), Educação (6), Férias (7), Reconversão profissional (8), Negócios (9), Outros (10)
other_installment_plans	Outrosplanos deparcelamento	Banco (0), Lojas (1), Nenhum (2)

Fonte: Adaptado de Hofmann (1994).

APÊNDICE C – DESCRIÇÃO DA BASE DE DADOS TDAH

As tabelas abaixo apresentam a lista dos atributos da base TDAH. Estes podem ser divididos em: características do indivíduo (Tabela 20), características familiares (Tabela 21), características socioeconômicas (Tabela 22), dados gestacionais (Tabela 23), natividade (Tabela 24) e Testes para aferição do QI e Desempenho Escolar (Tabela 25).

Tabela 20 – Atributos referentes a características do indivíduo.

Dados Pessoais		
Atributo	Descrição	Possíveis valores
Gênero	Gênero do paciente	Masculino ou Feminino
Idade	Tempo de vida do paciente contado em anos	6...17
Peso	Com quantos quilogramas o paciente se encontrava no momento do preenchimento	19.00...113.00
Altura	Com quantos metros e centímetros o paciente se encontrava no momento do preenchimento	0,76...1,72
Protestante_ Evangélico	Se o paciente é protestante ou evangélico	Sim ou Não
Religião_Outro	Se a religião do paciente é diferente das já listadas	Sim ou Não
Espírita	Se o paciente é espírita	Sim ou Não
Católico	Se o paciente é católico	Sim ou Não
Ateu_Sem_ Religião	Se o paciente é ateu ou sem religião	Sim ou Não
Cidade_Outras_ Regiões	Se a cidade em que o paciente reside é diferente de Belo Horizonte ou da região metropolitana de Belo Horizonte	Sim ou Não
Belo_Horizonte	Se a cidade em que o paciente reside é Belo Horizonte	Sim ou Não

Região_ Metropolitana_ BH	Se a cidade em que o paciente reside faz parte da metropolitana de Belo Horizonte	Sim ou Não
Dados Escolares		
Atributo	Descrição	Possíveis valores
Ano_Escolar	Série do paciente na escola	Educação infantil, Primeiro ano, Segundo ano, Terceiro ano, Quarto ano, Quinto ano, Sexto ano, Sétimo ano, Oitavo ano, Nono ano ou Ensino médio
Tipo_Escola	Tipo de escola frequentada pelo paciente	Pública ou Particular
Dados Saúde Geral		
Atributo	Descrição	Possíveis valores
Problema_Visão	Se o paciente já teve/ainda têm problema de visão	Sim ou Não
Problema_Audição	Se o paciente já teve/ainda têm problema de audição	Sim ou Não
Ganhou_Pouco_Peso	Se o paciente já teve/ainda têm um peso considerado baixo	Sim ou Não
Ganhou_Muito_Peso	Se o paciente já teve/ainda têm um peso considerado alto	Sim ou Não
Alergia_Alimentar	Se o paciente já teve/ainda têm alergia alimentar	Sim ou Não
Convulsões_Epilepsia	Se o paciente já teve/ainda têm convulsões ou epilepsia	Sim ou Não
Asma_Bronquite	Se o paciente já teve/ainda tem asma ou bronquite	Sim ou Não
Internação	Se o paciente já foi internado	Sim ou Não
Acidente	Se o paciente já sofreu algum acidente (quedas, acidente de carro...)	Sim ou Não
Osso_Quebrado	Se o paciente já teve algum osso quebrado	Sim ou Não

Fonoaudiologia	Se o paciente já fez (faz) acompanhamento Fonoaudiológico	Sim ou Não
Psicologia	Se o paciente já fez (faz) acompanhamento Psicológico	Sim ou Não
Terapia_Ocupacional	Se o paciente já fez (faz) Terapia Ocupacional	Sim ou Não
Psicopedagogia	Se o paciente já fez (faz) acompanhamento Psicopedagógico	Sim ou Não
Neurologista	Se o paciente já fez (faz) acompanhamento Neurológico	Sim ou Não
Psiquiatra	Se o paciente já fez (faz) acompanhamento Psiquiátrico	Sim ou Não
Psicostimulante	Se o paciente faz uso de Concerta, Ritalina ou Venvanse	Sim ou Não
Antipsicótico	Se o paciente faz uso de Neuleptil, Risperidona ou Aristab	Sim ou Não
	Dados Sono	
Atributo	Descrição	Possíveis valores
Sono_em_Horas	Quantas horas o paciente dorme por noite	5,00...13,00
Cochila_Dia	Se o paciente cochila durante o dia	Sim ou Não
Cama_Própria	Se o paciente dorme em cama individual	Sim ou Não
Quarto_Próprio	Se o paciente tem seu quarto individual	Sim ou Não
Sono_Muito_Dia	Se o paciente tem muito sono durante o dia	Sim ou Não
Pesadelos	Se o paciente tem pesadelos enquanto dorme	Sim ou Não
Conversa_Dormindo	Se o paciente conversa enquanto dorme	Sim ou Não
Sonambulismo	Se o paciente tem episódios de sonambulismo	Sim ou Não

Medo_Dormir_Sozinho	Se o paciente tem medo de dormir sem companhia	Sim ou Não
Ritual	Se o paciente executa alguma prática ou costume antes de dormir	Sim ou Não
Entrevista K-SADS		
Atributo	Descrição	Possíveis valores
TDAH	Se o paciente tem TDAH	Sim ou Não
TDAH_SOE	Se o paciente tem TDAH sem especificação	Sim ou Não
TDAH_Hiperativo	Se o paciente tem TDAH do subtipo predominantemente hiperativo-impulsivo	Sim ou Não
TDAH_Desatento	Se o paciente tem TDAH do subtipo predominantemente desatento	Sim ou Não
TDAH_Combinado	Se o paciente tem TDAH do subtipo combinado	Sim ou Não
Transtornos_Eliminação	Se o paciente tem Enurese ou Encoprese	Sim ou Não
Transtornos_Comportamento	Se o paciente tem Transtorno da Conduta ou Transtorno de Oposição Desafiante	Sim ou Não
Transtornos_Ansiedade	Se o paciente tem Transtorno de Pânico, Transtorno de Ansiedade Social (Fobia Social), Agorafobia, Transtorno de Ansiedade Generalizada ou Transtorno Obsessivo-Compulsivo	Sim ou Não
Transtornos_Humor	Se o paciente tem Depressão ou Mania	Sim ou Não
TEA	Se o paciente apresenta o Transtorno do Espectro Autista	Sim ou Não

Fonte: Extraído de Jandre (2021).

Tabela 21 – Atributos referentes a características familiares.

Dados Gerais		
Atributo	Descrição	Possíveis valores
Mora_Mãe	Se o paciente mora somente com a mãe	Sim ou Não
Mora_Pais	Se o paciente mora com o pai e a mãe	Sim ou Não
Mora_Mãe_Parceiro	Se o paciente mora com a mãe e um cônjuge, que não é o pai do paciente	Sim ou Não
Número_Irmãos	Quantos irmãos o paciente tem	0...9
Pais_Biológicos	Se a mãe e o pai do paciente são biológicos	Sim ou Não
Dados da Mãe		
Atributo	Descrição	Possíveis valores
Número_Parceiros_Mãe	Com quantos parceiros românticos a mãe do paciente já morou, além do pai do paciente, depois do nascimento do paciente	0 ou 1
Idade_Mãe	Tempo de vida da mãe do paciente contado em anos	22...67
Escolaridade_Mãe	Qual o grau de escolaridade da mãe do paciente	Ensino fundamental, Ensino médio ou Ensino Superior
Mãe_Solteira	Se a mãe do paciente está solteira	Sim ou Não
Mãe_Divorciada	Se a mãe do paciente está divorciada	Sim ou Não
Mãe_Casada	Se a mãe do paciente está casada	Sim ou Não
Mãe_Amigada	Se a mãe do paciente está amigada	Sim ou Não

Dados do Pai		
Atributo	Descrição	Possíveis valores
Idade_Pai	Tempo de vida do pai do paciente contado em anos	26...81
Escolaridade_Pai	Qual o grau de escolaridade do pai do paciente	Ensino fundamental, Ensino médio ou Ensino Superior
Pai_Solteiro	Se o pai do paciente está solteiro	Sim ou Não
Pai_Divorciado	Se o pai do paciente está divorciado	Sim ou Não
Pai_Casado	Se o pai do paciente está casado	Sim ou Não
Pai_Amigado	Se o pai do paciente está amigado	Sim ou Não
Histórico Familiar de Transtornos		
Atributo	Descrição	Possíveis valores
HF_Primeiro	Se o pai e/ou a mãe do paciente não possui algum transtorno	Sim ou Não
HF_Mãe	Se a mãe do paciente possui algum transtorno	Sim ou Não
HF1_Transtornos_Ansiedade	Se no histórico familiar de primeiro grau consta algum Transtorno de Ansiedade	Sim ou Não
HF1_Transtornos_Humor	Se no histórico familiar de primeiro grau consta algum Transtorno do Humor	Sim ou Não
Um_Avós_Paternos	Se um dos avós paternos do paciente possui algum transtorno	Sim ou Não
Um_Avós_Maternos	Se um dos avós maternos do paciente possui algum transtorno	Sim ou Não
HF_Segundo	Se um irmão e/ou um dos avós do paciente não possui algum transtorno	Sim ou Não

HF2_Depressão	Se no histórico familiar de segundo grau consta algum Transtorno Depressivo	Sim ou Não
QEDP - Questionário de Estilos e Dimensões Parentais		
Atributo	Descrição	Possíveis valores
Indulgência	Pontuação que indica o quanto o responsável do paciente é permissível	5...24
Acolhimento_Emocional	Pontuação que indica o quanto o responsável do paciente o apoia e dá afeto	12...25
Diálogo	Pontuação que indica o quanto o responsável do paciente conversa para estabelecer as regras	9...25
Autonomia	Pontuação que indica o quanto o responsável do paciente incentiva a participação democrática	5...25
Coerção_Física	Pontuação que indica o quanto o responsável do paciente bate nele	2...18
Hostil_Verbal	Pontuação que indica o quanto o responsável do paciente grita ou fala alto com ele	4...20
Punição	Pontuação que indica o quanto o responsável do paciente o coloca de castigo com pouca ou nenhuma explicação	1...17

ASRS - Adult Self-Report Scale		
Atributo	Descrição	Possíveis valores
Desatenção_Mãe	Pontuação da mãe do paciente na soma dos itens de "Desatenção"	1...34
Hiperatividade_Mãe	Pontuação da mãe do paciente na soma dos itens de "Hiperatividade-Impulsividade"	0...35
Mãe_Antes_12_Anos	Se a mãe do paciente apresentou os sintomas presentes no questionário antes dos 12 anos	Sim ou Não
Mãe_Prejuízo_Funcional	Se a manifestação dos sintomas presentes no questionário ocasionam perturbações que influenciam negativamente a vida da mãe do paciente	Sim ou Não
Desatenção_Pai	Pontuação do pai do paciente na soma dos itens de "Desatenção"	5...23
Hiperatividade_Pai	Pontuação do pai do paciente na soma dos itens de "Hiperatividade-Impulsividade"	1...24
Pai_Prejuízo_Funcional	Se a manifestação dos sintomas presentes no questionário ocasionam perturbações que influenciam negativamente a vida do pai do paciente	Sim ou Não

IDATE - State-Trait Anxiety Inventory		
Atributo	Descrição	Possíveis valores
Mãe_Estado	Pontuação da mãe do paciente na análise do grau de ansiedade numa reação transitória diretamente relacionada a uma situação de adversidade que se apresenta em dado momento	Ansiedade baixa, Ansiedade média ou Ansiedade alta
Mãe_Traço	Pontuação da mãe do paciente na análise do aspecto mais estável relacionado à propensão do indivíduo lidar com maior ou menor ansiedade ao longo de sua vida	Ansiedade baixa, Ansiedade média ou Ansiedade alta
Pai_Estado	Pontuação do pai do paciente na análise do grau de ansiedade numa reação transitória diretamente relacionada a uma situação de adversidade que se apresenta em dado momento	Ansiedade baixa, Ansiedade média ou Ansiedade alta
Pai_Traço	Pontuação do pai do paciente na análise do aspecto mais estável relacionado à propensão do indivíduo lidar com maior ou menor ansiedade ao longo de sua vida	Ansiedade baixa, Ansiedade média ou Ansiedade alta
BDI-II - Beck Depression Inventory		
Atributo	Descrição	Possíveis valores
BDI_Mãe	Pontuação total da detecção de sintomas depressivos na mãe do paciente	Depressão leve, Depressão moderada ou Depressão severa
BDI_Pai	Pontuação total da detecção de sintomas depressivos no pai do paciente	Depressão leve, Depressão moderada ou Depressão severa

Fonte: Extraído de Jandre (2021).

Tabela 22 – Atributos referentes a características socioeconômicas.

Atributo	Descrição	Possíveis valores
N_Habitantes	Número de moradores no domicílio	2...7
Instrução_Chefe	Grau de escolaridade do chefe da família	Analfabeto/Fundamental I incompleto, Fundamental I completo/ Fundamental II incompleto, Fundamental II completo/ Médio incompleto, Médio completo/Superior incompleto ou Superior completo
Chefe_Pai	Se o pai é o chefe da família	Sim ou Não
Chefe_Mãe	Se a mãe é a chefe da família	Sim ou Não
Chefe_Outro	Se outra pessoa, que não é nem a mãe nem o pai do paciente, é a chefe da família	Sim ou Não
SAE	Classe socioeconômica de acordo com a Secretaria de Assuntos Estratégicos (SAE) segundo a renda	Pobre, Vulnerável, Classe média ou Classe alta

Fonte: Extraído de Jandre (2021).

Tabela 23 – Atributos gestacionais.

Atributo	Descrição	Possíveis valores
Gravidez_Planejada	Se a gravidez foi planejada pelos pais do paciente	Sim ou Não
Gravidez_Aceita	Se a gravidez foi bem aceita pelos pais do paciente	Sim ou Não
Anticoncepcional	Se a mãe do paciente engravidou mesmo fazendo uso de anticoncepcional	Sim ou Não
Preferência_Sexo	Se os pais do paciente tinham preferência pelo sexo do bebê	Sim ou Não

Dificuldade_Engravidar	Se os pais do paciente tiveram dificuldades para a gravidez acontecer	Sim ou Não
Pré_Natal	Se houve acompanhamento médico pelo pré-natal	Sim ou Não
Gravidez_ Problema_Saúde	Se a mãe do paciente teve algum problema de saúde durante a gravidez	Sim ou Não
Gravidez_ Internação	Se a mãe do paciente ficou internada alguma vez durante a gravidez	Sim ou Não
Gravidez_ Sangramentos	Se houve algum sangramento durante a gravidez	Sim ou Não
Gravidez_Ameaça_Aborto	Se houve ameaça de aborto espontâneo durante a gravidez	Sim ou Não
Gravidez_Álcool	Se a mãe do paciente fez uso de álcool durante a gravidez	Sim ou Não
Cigarros_Por_Dia	Quantos cigarros a mãe do paciente fumou por dia durante a gravidez	0...30
Gravidez_Medicação	Se a mãe do paciente fez uso de alguma medicação durante a gravidez	Sim ou Não

Fonte: Extraído de Jandre (2021).

Tabela 24 – Atributos de Natividade.

Atributo	Descrição	Possíveis valores
Parto	Tipo de parto	Normal ou Cesária
Idade_Gestacional	Com qual tempo gestacional o paciente nasceu	Prematuro, Termo ou Pós-termo
Peso_Nascimento	Qual foi o peso de nascimento do paciente, considerando as quilogramas	Baixo, Adequado ou Elevado
Comprimento_Nascimento	Qual foi o comprimento de nascimento do paciente, considerando os centímetros	Pequeno, Adequado ou Grande
Apgar_5min	Somatório da pontuação obtida no teste Apgar durante a avaliação do paciente ao nascer	Ótimas condições, Dificuldade leve ou moderada, ou Dificuldade grave
Chorou_Nascimento	Se o paciente chorou ao nascer	Sim ou Não
Problema_Nascimento	Se houve algum problema no nascimento do paciente	Sim ou Não
Alta_Com_Mãe_Nacimento	Se o paciente teve alta do hospital com a mãe	Sim ou Não

Fonte: Extraído de Jandre (2021).

Tabela 25 – Atributos do Testes para aferição do QI e Desempenho Escolar.

Atributo	Descrição	Possíveis valores
Teste_Raven	Resultado obtido no Teste das Matrizes Progressivas de Raven, por meio da média do desvio padrão	Incerto, Inferior, Médio ou Superior
Desempenho_Arit	Nível do paciente no Teste de Desempenho Escolar ao comparar sua nota no subteste de "Aritmética" com a nota média do estado de Minas Gerais no mesmo subteste	Inferior, Médio ou Superior

Desempenho_Esc	Nível do paciente no Teste de Desempenho Escolar ao comparar sua nota no subteste de "Escrita" com a nota média do estado de Minas Gerais no mesmo subteste	Inferior, Médio ou Superior
Desempenho_Leit	Nível do paciente no Teste de Desempenho Escolar ao comparar sua nota no subteste de "Leitura" com a nota média do estado de Minas Gerais no mesmo subteste	Inferior, Médio ou Superior

Fonte: Extraído de Jandre (2021).

APÊNDICE D – DESCRIÇÃO DA BASE DE DADOS DEPRESSÃO

As tabelas abaixo apresentam a lista dos atributos da base de Depressão. Estes podem ser divididos em: atributos demográficos (Tabela 26), sociais (Tabela 27), pontuações obtidas pelos inventários CDI (Tabela 28) e YSR (Tabela 29) e outras questões consideradas importantes pela comunidade de saúde mental (Tabela 30).

Tabela 26 – Atributos demográficos.

Descrição dos atributos	**Tipo**	**Possíveis Valores**
Sexo	Categórico	Masculino, feminino
Idade	Numérico	Em anos
Idade da mãe	Numérico	Em anos
Escolaridade da mãe	Categórico	[Não estudou], [Ensino Fundamental Incompleto], [Ensino Fundamental Completo], [Ensino Médio Incompleto], [Ensino Médio Completo], [Superior Incompleto], [Superior Completo], [Não Sabe]
Mãe trabalha?	Categórico	Sim, não
Idade do pai	Numérico	Em anos
Escolaridade do pai	Categórico	[Não estudou], [Ensino Fundamental Incompleto], [Ensino Fundamental Completo], [Ensino Médio Incompleto], [Ensino Médio Completo], [Superior Incompleto], [Superior Completo], [Não Sabe]
Pai trabalha?	Categórico	Sim, não

Tabela 27 – Atributos sociais.

Descrição	Tipo	Possíveis Valores
Tempo em horas que você passa com a sua mãe por dia durante a semana?	Categórico	[0 a 3], [4 a 7], [8 a 11], [12 a 15], [16 a 19], [20 ou mais]
Tempo em horas que você passa com a sua mãe por dia no final de semana?	Categórico	[0 a 3], [4 a 7], [8 a 11], [12 a 15], [16 a 19], [20 ou mais]
Tempo em horas que você passa com seu pai por dia durante a semana?	Categórico	[0 a 3], [4 a 7], [8 a 11], [12 a 15], [16 a 19], [20 ou mais]
Tempo em horas que você passa com seu pai por dia no final de semana?	Categórico	[0 a 3], [4 a 7], [8 a 11], [12 a 15], [16 a 19], [20 ou mais]
Você está ou esteve em atendimento psicológico ou psiquiátrico?	Categórico	Sim, não
Pais vivem juntos ou separados	Categórico	Juntos, Separ./divorciados
A mãe está ou esteve em atendimento psicológico ou psiquiátrico?	Categórico	Sim, não
Alguém da família da mãe está ou esteve em atendimento psicológico ou psiquiátrico?	Categórico	Sim, não
A mãe toma algum tipo de medicação de uso contínuo?	Categórico	Sim, não
Qual remédio a mãe toma?	Texto	Livre
Tempo em horas que a mãe passa com seu(s) filho(s) por dia durante a semana	Categórico	[0 a 3], [4 a 7], [8 a 11], [12 a 15], [16 a 19], [20 ou mais]
Tempo em horas que a mãe passa com seu(s) filho(s) por dia durante o final de semana	Categórico	[0 a 3], [4 a 7], [8 a 11], [12 a 15], [16 a 19], [20 ou mais]
O pai está ou esteve em atendimento psicológico ou psiquiátrico?	Categórico	Sim, não
Alguém da família do pai está ou esteve em atendimento psicológico ou psiquiátrico?	Categórico	Sim, não
O pai toma algum tipo de medicação de uso contínuo?	Categórico	Sim, não
Qual remédio o pai toma?	Texto	Livre
Tempo em horas que o pai passa com seu(s) filho(s) por dia durante a semana	Categórico	[0 a 3], [4 a 7], [8 a 11], [12 a 15], [16 a 19], [20 ou mais]
Tempo em horas que o pai passa com seu(s) filho(s) por dia durante o final de semana	Categórico	[0 a 3], [4 a 7], [8 a 11], [12 a 15], [16 a 19], [20 ou mais]

Tabela 28 – Atributos que representam as questões do CDI.

Questões	Tipo	Possíveis Valores
Questão 1	Categórico	[0 - Eu fico triste de vez em quando], [1 - Eu fico triste muitas vezes], [2 - Eu estou sempre triste]
Questão 2	Categórico	[0 - Para mim tudo se resolverá bem], [1 - Eu não tenho certeza se as coisas darão certo para mim], [2 - Nada vai dar certo para mim]
Questão 3	Categórico	[0 - Eu faço bem a maioria das coisas], [1 - Eu faço errado a maioria das coisas], [2 - Eu faço tudo errado]
Questão 4	Categórico	[0 - Eu me divirto com muitas coisas], [1 - Eu me divirto com algumas coisas], [2 - Nada é divertido para mim]
Questão 5	Categórico	[0 - Eu sou mau (má) de vez em quando], [1 - Eu sou mau (má) com frequência], [2 - Eu sou sempre mau (má)]
Questão 6	Categórico	[0 - De vez em quando eu penso que coisas ruins vão me acontecer], [1 - Eu temo que coisas ruins me aconteçam], [2 - Eu tenho certeza que coisas terríveis me acontecerão]
Questão 7	Categórico	[0 - Eu gosto de mim mesmo], [1 - Eu não gosto muito de mim], [2 - Eu me odeio]
Questão 8	Categórico	[0 - Normalmente, eu não me sinto culpado pelas coisas ruins que acontecem], [1 - Muitas coisas ruins que acontecem são por minha culpa], [2 - Tudo de mal que acontece é por minha culpa]
Questão 9	Categórico	[0 - Eu não penso em me matar], [1 - Eu penso em me matar], [2 - Eu quero me matar]
Questão 10	Categórico	[0 - Eu sinto vontade de chorar de vez em quando], [1 - Eu sinto vontade de chorar frequentemente], [2 - Eu sinto vontade de chorar diariamente]
Questão 11	Categórico	[0 - Eu me sinto preocupado de vez em quando], [1 - Eu me sinto preocupado frequentemente], [2 - Eu me sinto sempre culpado]
Questão 12	Categórico	[0 - Eu gosto de estar com pessoas], [1 - Frequentemente, eu não gosto de estar com pessoas], [2 - Eu não gosto de estar com pessoas]
Questão 13	Categórico	[0 - Eu tomo decisões facilmente], [1 - É difícil para mim tomar decisões], [2 - Eu não consigo tomar decisões]
Questão 14	Categórico	[0 - Eu tenho boa aparência], [1 - Minha aparência tem alguns aspectos negativos], [2 - Eu sou feio (feia)]
Questão 15	Categórico	[0 - Fazer os deveres de casa não é um grande problema para mim], [1 - Com frequência eu tenho que ser pressionado para fazer os deveres de casa], [2 - Eu tenho que me obrigar a fazer os deveres de casa]
Questão 16	Categórico	[0 - Eu durmo bem à noite], [1 - Eu tenho dificuldade para dormir algumas noites], [2 - Eu tenho sempre dificuldades para dormir à noite]
Questão 17	Categórico	[0 - Eu me canso de vez em quando], [1 - Eu me canso frequentemente], [2 - Eu estou sempre cansado (cansada)]
Questão 18	Categórico	[0 - Eu como bem], [1 - Alguns dias eu não tenho vontade de comer], [2 - Quase sempre eu não tenho vontade de comer]
Questão 19	Categórico	[0 - Eu não temo sentir dor nem adoecer], [1 - Eu temo sentir dor e ficar doente], [2 - Eu estou sempre temeroso de sentir dor e ficar doente]
Questão 20	Categórico	[0 - Eu não me sinto sozinho (sozinha)], [1 - Eu me sinto sozinho(a) muitas vezes], [2 - Eu sempre me sinto sozinho (sozinha)]
Questão 21	Categórico	[0 - Eu me divirto na escola frequentemente], [1 - Eu me divirto na escola de vez em quando], [2 - Eu nunca me divirto na escola]
Questão 22	Categórico	[0 - Eu tenho muitos amigos], [1 - Eu tenho muitos amigos e gostaria de ter mais], [2 - Eu não tenho amigos]
Questão 23	Categórico	[0 - Meus trabalhos escolares são bons], [1 - Meus trabalhos escolares não são tão bons como eram antes], [2 - Eu tenho me saído mal em matérias em que costumava ser bom (boa)]
Questão 24	Categórico	[0 - Sou tão bom quanto outras crianças], [1 - Se eu quiser, posso ser tão bom quanto outras crianças], [2 - Não posso ser tão bom quanto outras crianças]
Questão 25	Categórico	[0 - Eu tenho certeza que sou amado (a) por alguém], [1 - Eu não tenho certeza se alguém me ama], [2 - Ninguém gosta de mim realmente]
Questão 26	Categórico	[0 - Eu sempre faço o que me mandam], [1 - Eu não faço o que me mandam com frequência], [2 - Eu nunca faço o que me mandam]
Questão 27	Categórico	[0 - Eu não me envolvo em brigas], [1 - Eu me envolvo em brigas com frequência], [2 - Eu estou sempre me envolvendo em brigas]
Escore Total	Numérico	

Fonte: Extraído de Kovacs (2003).

Tabela 29 – Atributos referente ao questionário YSR.

Descrição dos atributos	Tipo
Total de Ansiedade e depressão (Internalizante)	Numérico
Total de Retraimento e depressão (Internalizante)	Numérico
Total de Queixas somáticas (Internalizante)	Numérico
Total de Problemas sociais	Numérico
Total de Problemas de Pensamento	Numérico
Total de Problemas de Atenção	Numérico
Total de Comportamento de quebrar regras (Externalizante)	Numérico
Total de Comportamento agressivo (Externalizante)	Numérico
Total de Problemas internalizantes	Numérico
Total de Problemas externalizantes	Numérico
Total de Outros problemas	Numérico

Fonte: Extraído de Rocha, Araújo e Silvares (2008).

Tabela 30 – Outras questões consideradas importantes pela comunidade de saúde mental.

Descrição dos atributos	Tipo
Total da Escala orientada pelo DSM de Problemas Afetivos	Numérico
Total da Escala orientada pelo DSM de Problemas de Ansiedade	Numérico
Total da Escala orientada pelo DSM de Problemas somáticos	Numérico
Total da Escala orientada pelo DSM de Problemas de Hiperatividade/Déficit de Atenção	Numérico
Total da Escala orientada pelo DSM de Problemas de Comportamento opositor-desafiador	Numérico
Total da Escala orientada pelo DSM de Problemas de Conduta	Numérico
Total da Escala de Qualidades Positivas de Problemas Obsessivo-Compulsivo	Numérico
Total da Escala de Qualidades Positivas de Problemas de Estresse pós-traumático	Numérico
Total da Escala de Qualidades Positivas de Qualidades Positivas	Numérico

Fonte: Extraído de APA et al. (2013).