Explicabilidade de modelos para Avaliação Automática de Redações

Gabriel Menezes Moreira[1], Raimundo Santos Moura[1]

Departamento de Computação – Universidade Federal do Piauí (UFPI)
Teresina – PI – Brasil

gabriel@ufpi.edu.br, rsm@ufpi.edu.br

Abstract. *In the context of Automated Essay Scoring (AES), one of the main learning points for the student is to receive feedback on their correction in order to understand where and what they have done wrong. In this sense, this work aims to research interpretability methods for AI models, with a focus on AES activity in the style of the National High School Exam (ENEM, Exame Nacional do Ensino Médio). The achieved results reveal that LIME (Local Interpretable Model-Agnostic Explanations) highlights crucial terms and trends that the model associates with in the AES task. These particularities played a fundamental role in identifying the strengths and weaknesses of the developed models.*

Resumo. *No contexto da Avaliação Automática de Redações (AAR), um dos principais pontos de aprendizagem do aluno é receber o feedback da sua correção para entender onde e o que errou. Nesse sentido, esse trabalho tem o objetivo de pesquisar métodos de interpretabilidade para modelos de AM, com o foco na atividade de AAR no estilo do Exame Nacional do Ensino Médio (ENEM). Os resultados alcançados revelam que o LIME (Local Interpretable Model-Agnostic Explanations) evidencia termos cruciais e tendências que o modelo vincula na tarefa de AAR. Essas particularidades desempenharam um papel fundamental na identificação dos pontos fortes e fracos dos modelos desenvolvidos.*

1. Introdução

A interpretabilidade de modelos de Aprendizado de Máquina - AM (do inglês: *Machine Learning* - ML) é um tópico de grande interesse na comunidade de Inteligência Artificial (IA). A importância desse tópico pode ser ilustrada pela crescente quantidade de pesquisas publicadas sobre ele nas últimas décadas.

Desde a década de 1990, os pesquisadores vêm explorando formas de tornar os modelos de AM mais compreensíveis, como o uso de árvores de decisão, regressão linear e modelos baseados em regras [Doshi-Velez & Kim, 2017]. No entanto, à medida que a complexidade dos modelos de AM aumentou, também aumentou a necessidade de métodos mais avançados de interpretabilidade.

Nos últimos anos, tem havido um crescente interesse por parte dos pesquisadores em explorar abordagens de interpretabilidade em modelos de AM. Foram desenvolvidas abordagens como o LIME (*Local Interpretable Model-Agnostic*

Explanations) [Ribeiro et al., 2016] e SHAP (*SHapley Additive exPlanations*) [Lundberg & Lee, 2017], que fornecem interpretabilidade em nível de instância para modelos de AM baseados em caixa preta. Esses métodos têm sido aplicados em diversos domínios, incluindo saúde, finanças, justiça criminal e Processamento de Linguagem Natural (PLN) [Doshi-Velez & Kim, 2017].

No contexto da Avaliação Automática de Redações (AAR), modelos de correção automática são utilizados e estudados [Marinho et al., 2022], porém, essas ferramentas não são muito exploradas na língua portuguesa. Além disso, os modelos de AM podem carregar e amplificar o viés humano na sua tomada de decisão.

Nesse sentido, será feito um estudo sobre interpretabilidade de modelos de AM a fim de detalhar os passos que foram tomados pelo modelo caixa preta para disponibilizar um melhor *feedback* para os alunos que terão suas redações corrigidas e avaliadas.

Portanto, o objetivo deste trabalho é avaliar a capacidade de modelos de AM em analisar e classificar a qualidade de redações de forma automática e objetiva. AAR é uma tarefa importante em áreas como a educação e a linguística computacional, pois pode ser utilizada para avaliar o nível de habilidade e conhecimento de um aluno em um determinado tema, ou para identificar características linguísticas e estilísticas presentes em diferentes gêneros textuais [Attali & Burstein, 2006].

O restante do artigo está organizado da seguinte maneira. A seção 2 apresenta brevemente conceitos da área de IA que são necessários para o entendimento do trabalho. A seção 3 descreve os principais trabalhos relacionados. A seção 4 detalha a metodologia seguida para os experimentos. A seção 5 exibe os resultados provenientes dos testes feitos. A seção 6 conclui o artigo e indica trabalhos futuros.

2. Referencial Teórico

2.1. Explicabilidade de Modelos de IA

A área de explicabilidade de modelos de IA é uma área ativa de pesquisa [Lipton, 2018] que busca tornar os modelos mais compreensíveis para humanos. A explicabilidade é crucial em diversas aplicações, especialmente em aplicações sensíveis, como saúde, finanças e justiça, onde é necessário compreender a base da decisão do modelo para garantir a confiança e a transparência [Rudin, 2019].

Várias técnicas têm sido propostas para melhorar a explicabilidade dos modelos de IA, incluindo análise de importância de recursos [Breiman, 2001], visualização de decisão [Murdoch et al., 2019], análise de surtos [Lundberg & Lee, 2017], modelos de camadas simples [Molnar et al., 2021] e abordagens de interpretação [Doshi-Velez & Kim, 2017]. Além disso, há abordagens para a criação de modelos interpretáveis por natureza, como métodos de abordagem interpretáveis [Ribeiro et al., 2016].

2.2. Avaliação Automática de Redações

No contexto de AM, a AAR é uma subárea da IA que se concentra em avaliar a qualidade de uma redação de forma automatizada. Esta técnica é aplicada em muitos contextos, incluindo educação, recrutamento, tradução e outros [Marinho et al., 2022].

Nesse sentido, a AAR se baseia em modelos de AM, que são treinados com amostras de redações de alta qualidade e avaliadas por especialistas [Attali & Burstein, 2006]. Os modelos aprendem a identificar características específicas de redações de alta qualidade, como clareza, coesão, coerência, organização e uso adequado de gramática e vocabulário.

Existem várias abordagens para AAR, incluindo abordagens baseadas em regras, abordagens baseadas em modelos de AM supervisionado e abordagens baseadas em modelos de aprendizado profundo. Cada abordagem tem suas próprias vantagens e desvantagens, e a escolha de uma abordagem depende do contexto específico e da natureza do problema de avaliação de redações [Marinho et al., 2022].

2.3. Explicabilidade com LIME

LIME é uma abordagem inovadora no campo da interpretabilidade de modelos de AM. Desenvolvida para fornecer explicações acessíveis e compreensíveis sobre as previsões de modelos complexos, a técnica LIME oferece *insights* valiosos sobre como esses modelos tomam decisões em nível local, permitindo uma maior compreensão do processo de tomada de decisão dos modelos de AM.

Um dos fundamentos do LIME é o uso de modelos "explicativos", que são modelos mais simples e compreensíveis, como regressões lineares, que podem ser facilmente interpretados. Esses modelos explicativos são treinados para aproximar o comportamento do modelo complexo em uma região local em torno da instância de interesse. O processo de geração de explicações LIME envolve a perturbação dos dados de entrada e a observação de como o modelo complexo responde a essas perturbações. Com base nesses resultados, o modelo explicativo é treinado para refletir o comportamento do modelo complexo em torno da instância específica [Ribeiro et al., 2016].

3. Trabalhos relacionados

Existem diversos trabalhos que exploram técnicas de explicabilidade de modelos de AM na literatura. Nesta seção, destaca-se um dos principais trabalhos na abordagem LIME. Na parte final são apresentados dois trabalhos sobre AAR, incluindo o Corpus e a abordagem para pontuação automática.

O estudo seminal de [Ribeiro et al., 2016] introduziu o conceito de LIME e detalhou a metodologia por trás da abordagem. Ao demonstrar a aplicação do LIME em uma variedade de cenários, os autores forneceram evidências convincentes de sua eficácia em gerar explicações precisas e interpretáveis para modelos complexos.

No artigo [Marinho et al., 2022] sobre *Essay-Br*, é apresentado um *corpus* brasileiro destinado à tarefa de AAR. O estudo descreve a criação e as características

desse *corpus*, que é voltado para a língua portuguesa e tem como objetivo contribuir para a pesquisa e desenvolvimento de sistemas de AAR em língua portuguesa.

No trabalho [Marinho et al., 2022] sobre *Automated Essay Scoring*, é proposta uma abordagem para a pontuação automática de redações baseada nas competências do Exame Nacional do Ensino Médio (ENEM). Os autores detalham a metodologia empregada, incluindo a seleção de características relevantes das redações, a modelagem preditiva e a avaliação do desempenho da abordagem. A pesquisa visa fornecer uma alternativa eficaz e alinhada com os critérios de avaliação do ENEM para AAR.

O diferencial do nosso trabalho é analisar a eficiência de algoritmos de AM treinados em um *corpus* de redações do ENEM quanto a sua capacidade de correção e de interpretabilidade, por meio da ferramenta de interpretabilidade LIME.

4. Metodologia

4.1. Corpus

O *corpus* utilizado para treinar os modelos foi o *Essay-Br* [Marinho et al., 2022]. Essencialmente foram utilizados dois conjuntos de dados. O primeiro corresponde ao conjunto para treinamento dos modelos de AM, com as redações anotadas com a nota para cada competência e a nota geral associada. Já o segundo, corresponde às redações anotadas utilizadas nos testes de interpretabilidade LIME com os modelos treinados com o primeiro conjunto de dados.

4.2. Treinamento

Foi utilizado um modelo do tipo LSTM (Long Short-Term Memory) previamente criado no trabalho [Marinho et al., 2022]. Os testes conduzidos utilizaram como hiperparâmetros para o modelo LSTM os valores exibidos ao longo da seção. Nessa abordagem, antes da análise, as redações passaram por um processo preliminar no qual caracteres não alfanuméricos foram eliminados e todas as letras foram convertidas para letras minúsculas. Posteriormente, os textos foram transformados em vetores de dimensão 1.000.

A estrutura da RNN (Rede Neural Recorrente) foi construída com uma camada de *embeddings* como ponto de partida. Essa camada recebe os vetores de textos pré-processados e tokenizados e é responsável por gerar uma matriz composta pelos vetores de *embeddings*, cada um com 300 dimensões. A segunda camada da RNN consistiu em 10 células LSTM, seguidas por uma camada de saída com ativação softmax[1].

No treinamento das RNNs, foram executadas 15 épocas, com lotes de 10 redações processadas em cada iteração. Ao longo do processo de treinamento,

[1] A ativação softmax é uma função matemática usada em problemas de classificação para calcular probabilidades. Dada uma lista de números, a função softmax os transforma em valores entre 0 e 1, de forma que somem 1. [Nwankpa et al., 2018]

empregou-se o otimizador Adam [Kingma & Ba, 2015] com uma taxa de aprendizado de 0.001 e o erro médio quadrático foi adotado como a métrica de erro.

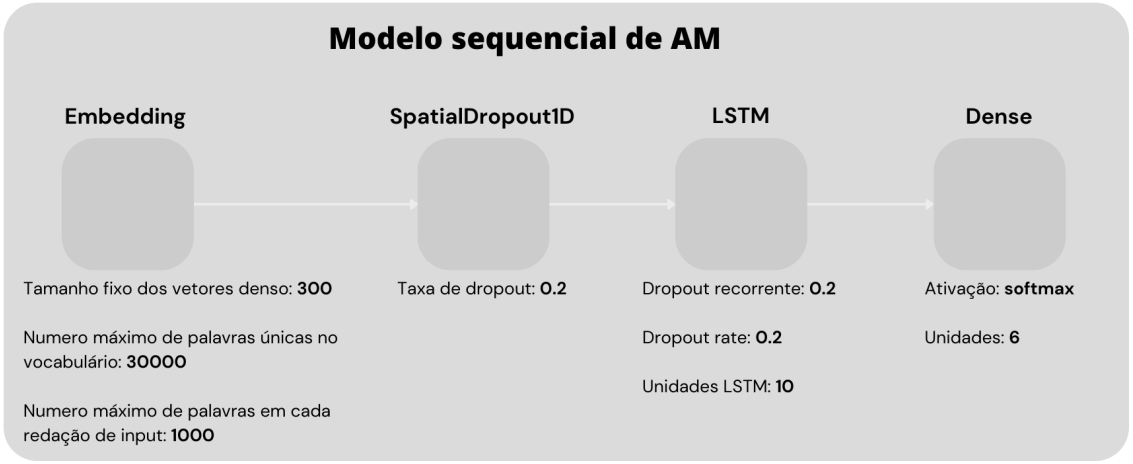

Figura 1. Arquitetura do modelo LSTM

4.3. Explicabilidade

A estratégia utilizada para realizar a explicabilidade de cada um dos modelos utilizando LIME seguiu o passo a passo de carregar as redações de treino para a competência em questão.

Após isso foi utilizado, o mesmo tokenizador da etapa de treinamento para converter os textos de entrada em sequências de números inteiros, realizando um mapeamento entre palavras ou *tokens* e suas representações numéricas correspondentes.

Logo em seguida, foi utilizado a função *pad_sequences* da biblioteca Keras para adicionar *padding* ou truncar essas sequências de modo que todas tenham o mesmo comprimento, no caso vetores de dimensão 1.000.

Com os dados de texto já pré-processados, essa informação é dada ao modelo LSTM treinado que retorna uma matriz de probabilidades previstas para cada classe, com uma linha para cada texto de entrada e uma coluna para cada classe. Cada classe corresponde à nota que uma competência do ENEM pode obter.

Essas probabilidades previstas podem ser usadas para determinar a classe mais provável para cada texto de entrada, bem como para gerar explicações para as previsões do modelo usando a biblioteca LIME.

5. Resultados

Nas próximas subseções discute-se os resultados obtidos e a interpretabilidade de cada modelo que foi gerado. Seguindo a metodologia de primeiro expor o mapa de frequência das notas esperadas e das notas obtidas pelo modelo. Logo em seguida, a explicabilidade para as redações com maior diferença entre nota esperada e nota obtida será comentada.

5.1 Competência 1

A competência 1, que versa sobre o domínio da escrita formal da língua portuguesa [INEP, 2019], possui uma tendência do modelo treinado a pontuar as redações com nota 120. Na Figura 2 podemos perceber uma redução no número de redações de nota 80, 160 e 200 mas não necessariamente tais redações foram todas para a classe de nota 120.

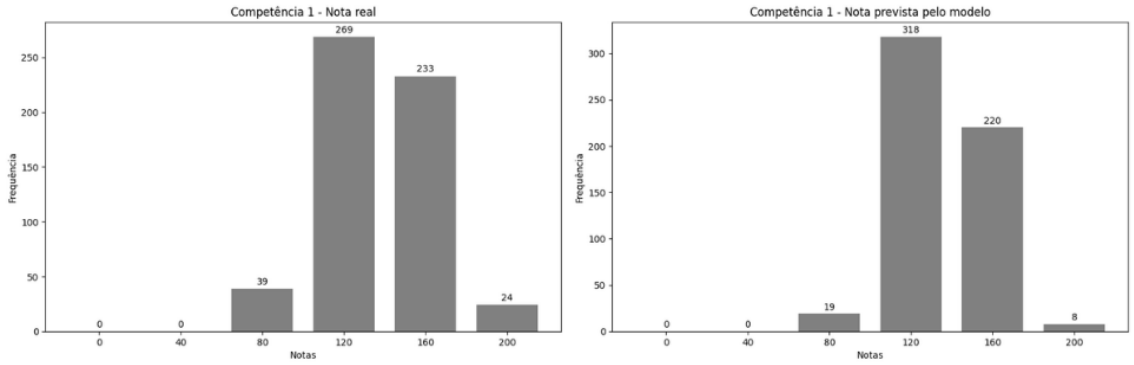

Figura 2. Notas reais X Notas previstas pelo modelo na competência 1

Com as notas reais e as notas previstas pelo modelo LSTM treinado, conseguimos identificar aquelas redações com a maior diferença entre as duas notas na competência 1. Dito isso, escolheu-se realizar a explicabilidade com uma redação com valor real 80 e valor previsto pelo modelo de 200.

Como método visual para o LIME, cada palavra da redação foi considerada como um token com peso associado. Cada token pode possuir peso positivo ou negativo, assim influenciando a probabilidade de cada classe de nota. Para cada classe de nota, o peso negativo é mostrado pelas palavras do lado esquerdo do gráfico (NOT 200) e o peso positivo é representado pelas palavras do lado direito do gráfico (200).

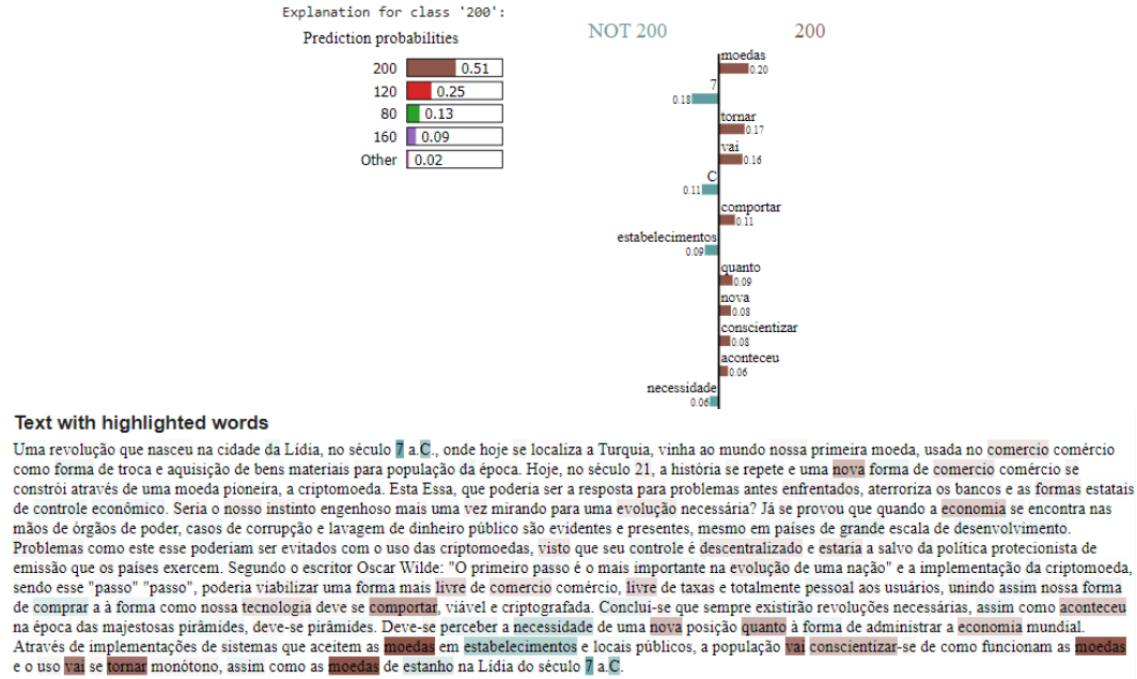

Figura 3. Explicabilidade da nota real na competência 1

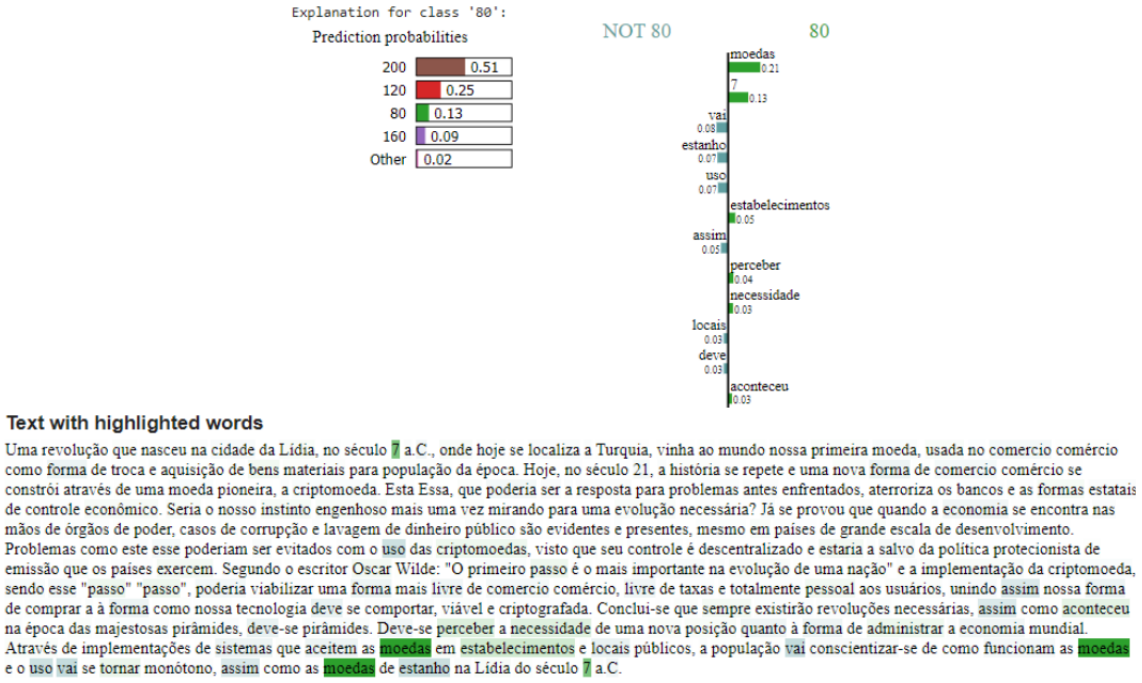

Figura 4. Explicabilidade da nota prevista na competência 1

No contexto da classe de nota 200, o LIME conferiu pesos positivos em termos como ‘comercio’ sem a acentuação correta de ‘comércio’. De tal forma, termos que caracterizam um mau domínio da língua, não obtiveram peso algum, como em ‘Esta Essa’, ‘deve-se pirâmides’, ‘este esse’ deveriam contribuir com peso negativo para a classe de nota 200, uma vez que, esses termos interferem na fluidez do texto e fogem à gramática normativa.

Curiosamente, essas palavras também não receberam pontuação positiva na classe de nota 80. Tal padrão de atribuição de pesos revelou a previsão do modelo, onde a probabilidade de uma redação receber a nota 200 na competência 1 atingiu 51%, enquanto a probabilidade de obter a nota 80 ficou em apenas 13%.

6. Conclusão e Trabalhos Futuros

A explicabilidade de modelos de AM é uma questão crítica na avaliação automática de redações, uma vez que a compreensão dos fatores que influenciam as decisões do modelo é essencial para garantir a justiça, a transparência e a responsabilidade do processo de avaliação.

Acerca dos resultados obtidos, pode-se observar que o método LIME destaca palavras-chave e padrões que o modelo associa a diferentes competências de redação. Tais características contribuíram para a identificação das falhas e acertos dos modelos treinados. De tal forma, esses *insights* destacam a importância de um treinamento mais contextualizado e aprimorado para melhorar a acurácia do modelo em relação às competências específicas da redação do ENEM.

Pontos de melhoria no trabalho proposto podem partir tanto do lado do refinamento do modelo de AM ao analisar novos modelos, quanto da integração dos textos motivadores e tema da redação no treinamento.

7. Referências Bibliográficas

Attali, Y., & Burstein, J. (2006). Automated essay scoring with e-rater® V. 2. Disponível em: https://ejournals.bc.edu/index.php/jtla/article/view/1650

Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32. Disponível em: https://link.springer.com/content/pdf/10.1023/A:1010933404324.pdf

Doshi-Velez, F., & Kim, B. (2017). Towards a Rigorous Science of Interpretable Machine Learning. Disponível em: https://arxiv.org/pdf/1702.08608.pdf

Kingma, D. P. and Ba, J. (2015). Adam: A method for stochastic optimization. Disponível em: https://arxiv.org/pdf/1412.6980.pdf%5D

Lipton, Z. C. (2018). The mythos of model interpretability. Disponível em: https://arxiv.org/pdf/1606.03490.pdf

Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. Disponível em: https://proceedings.neurips.cc/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf

Marinho, J. C., Anchiêta, R. T., Moura, R. S. (2022) Essay-BR: a Brazilian Corpus to Automatic Essay Scoring Task. Disponível em: https://sol.sbc.org.br/journals/index.php/jidm/article/view/2340/1978

Marinho, Jeziel C.; Cordeiro, Fábio; Anchiêta, Rafael T.; Moura, Raimundo S. (2022) Automated Essay Scoring: An approach based on ENEM competencies. Disponível em: https://sol.sbc.org.br/index.php/eniac/article/view/22769

MEC. (2019). Cartilha do INEP. Acesso em 04-08-2023. Disponível em: http://portal.mec.gov.br/ultimas-noticias/418-enem-946573306/81381-conheca-as-cinco-competencias-cobradas-na-redacao-do-enem

Molnar, C., Casalicchio, G., & Bischl, B. (2021). Interpretable machine learning–a brief history, state-of-the-art and challenges. Disponível em: https://arxiv.org/pdf/2010.09337.pdf

Murdoch, W. J., Singh, C., Kumbier, K., Abbasi-Asl, R., & Yu, B. (2019). Interpretable machine learning: definitions, methods, and applications. Disponível em: https://arxiv.org/pdf/1901.04592.pdf?fbclid=IwAR2frcHrhLc4iaH5-TmKKq263NVvAKHtG4uQoiVNDeLAG3QFzdje-yzZjiQ

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why Should I Trust You?": Explaining the Predictions of Any Classifier. Disponível em: https://arxiv.org/pdf/1602.04938.pdf?source=post_page----------

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Disponível em: https://arxiv.org/pdf/1811.10154.pdf?fbclid=IwAR01WIlfIiC1cgM99nhwIjAT0tHWY