Redin
REVISTA EDUCACIONAL INTERDISCIPLINAR
FACCAT
Núcleo de Educação On-line

# IA explicável para reduzir a assimetria de informação no consumo: uma análise comparativa de ferramentas e implicações educacionais

## Explainable AI to reduce information asymmetry in consumer relations: a comparative analysis of tools and educational implications

Maurício Zalamena Bavaresco[1]
Carine Geltrudes Webber[2]

## Resumo

Este artigo explora a necessidade de implementar a Inteligência Artificial Explicável (XAI) para mitigar a assimetria de informação nas relações de consumo. A opacidade dos algoritmos de IA frequentemente deixa os consumidores em desvantagem, incapazes de compreender as decisões automatizadas que impactam suas escolhas. Utilizando o Método Analítico Hierárquico (MAH), este estudo avalia ferramentas de XAI, como SHAP, LIME e InterpretML, identificando as mais eficazes para promover transparência e revelar a ética. Além de beneficiar os consumidores, a XAI também auxilia órgãos reguladores na garantia de práticas justas no mercado. Os resultados indicam que a adoção de XAII é essencial para desenvolver um mercado mais equilibrado e consciente, onde a explicabilidade das decisões algorítmicas fortalece a confiança e promove um consumo mais informado.

**Palavras-chave:** Inteligência artificial explicável; Assimetria de informação; Avaliação de ferramentas de IA; Educação em IA; Ética da IA.

## Abstract

This article explores the need to implement Explainable Artificial Intelligence (XAI) to mitigate information asymmetry in consumer relations. The opacity of AI algorithms often leaves consumers at a disadvantage, unable to understand the automated decisions that impact their choices. Using the Analytic Hierarchy Process (AHP), the study evaluates XAI tools, such as SHAP, LIME, and InterpretML, identifying the most effective ones to promote transparency and ethics. In addition to benefiting consumers, XAI also assists regulatory bodies in ensuring fair practices in the market. The results indicate that adopting XAI is essential to developing a more balanced and conscious market, where the explainability of algorithmic decisions strengthens trust and promotes more informed consumption.

**Keywords:** Explainable artificial intelligence; Information asymmetry; AI tools evaluation; AI education; AI ethics.

[1] Bacharel em Ciência da Computação pela Universidade de Caxias do Sul (UCS). E-mail: *mzbavaresco@ucs.br*
[2] Doutora em Ciência da Computação pela École Doctorale Mathématiques et Informatiques da Universite de Grenoble I Joseph Fourier (Grenoble/França). Professora Titular na Área de Conhecimento de Exatas e Engenharias da Universidade de Caxias do Sul (UCS), onde também integra os Programas de Pós-Graduação em Ensino de Ciências e Matemática e Computação Aplicada. E-mail: *cgwebber@ucs.br*

## 1. Introdução

A Inteligência Artificial (IA) tem se consolidado como uma área central no desenvolvimento tecnológico, com aplicações que vão desde a automação industrial até sistemas avançados de recomendação em plataformas digitais (LUGER, 2020). O ensino da IA já está bem estabelecido em cursos de graduação e pós-graduação em áreas das ciências exatas, como Ciências da Computação e Engenharias. No entanto, apesar do foco intenso nas técnicas de machine learning e deep learning, um aspecto crucial tem sido amplamente negligenciado na formação acadêmica: a Inteligência Artificial Explicável (XAI).

A XAI é a subárea da IA dedicada a tornar os processos decisórios das máquinas mais transparentes e compreensíveis para os seres humanos. Em um mundo cada vez mais dependente de algoritmos que operam como caixas pretas, a ausência de explicabilidade representa um risco significativo (WEBER, CARL e HINZ, 2024). O termo caixa preta refere-se a sistemas cujo funcionamento não é transparente ou facilmente compreensível aos seres humanos (ALVEY, ANDERSON e KELLER, 2024). Embora esses modelos sejam capazes de processar grandes volumes de dados e identificar padrões complexos, sua estrutura dificulta a interpretação dos cálculos e das camadas de processamento que levam a uma determinada saída.

A falta de interpretabilidade desses modelos levanta questões sobre confiabilidade e ética, especialmente em domínios críticos onde a explicabilidade das decisões é essencial. Além disso, quando desenvolvedores e engenheiros de software não são treinados para considerar a importância da transparência em seus modelos, a tendência é que continuem a produzir sistemas opacos, cujas decisões são difíceis, senão impossíveis, de entender (PHILLIPS et.al, 2021; LUNDBERG, 2017).

Esse cenário gera repercussões graves. A falta de explicabilidade na IA perpetua a criação de sistemas "caixas pretas" que, embora poderosos em termos de precisão e eficiência, falham em oferecer justificativas claras para suas decisões (MISHRA, 2023; ROSSETTI e ANGELUCI, 2021). Isso não apenas diminui a confiança dos usuários, mas também facilita a disseminação de informações falsas e a manipulação dos consumidores, que podem ser levados a decisões prejudiciais sem entender o porquê.

O resultado dessa lacuna na formação acadêmica é que, enquanto a IA avança rapidamente em termos de capacidade técnica, aspectos fundamentais para a segurança e a ética do uso dessas tecnologias ficam para trás. A ausência de XAI nos currículos acadêmicos implica que futuros profissionais, ao desenvolverem novas aplicações de IA, continuarão a repetir padrões problemáticos, exacerbando os riscos de desinformação e uso indevido de dados.

Outro fato importante, enfatizado neste artigo, é que a crescente utilização de modelos de IA nas relações de consumo tem ampliado significativamente a assimetria de informação entre empresas e consumidores (ROSSETTI e ANGELUCI, 2021). A assimetria de informação ocorre quando uma das partes em uma transação possui mais ou melhores informações do que a outra, criando um desequilíbrio que pode ser explorado de forma injusta (DOMINGUES, SILVA e SOUZA, 2021). No caso da IA, essa disparidade se manifesta através de sistemas opacos, conhecidos como *caixas-pretas*, onde as decisões automatizadas não são compreendidas pelos consumidores. Para mitigar essa desigualdade e promover um consumo mais consciente e ético, a área de explicabilidade da IA, emerge como uma solução crucial. A XAI visa reduzir essa assimetria, tornando os processos de tomada de decisão da IA mais transparentes e acessíveis.

A teoria de triagem (*screening*) de Joseph E. Stiglitz (1975) oferece uma lente adicional para compreender a importância da explicabilidade. Segundo Stiglitz, a triagem ocorre quando a parte menos informada, como consumidores ou reguladores, adota estratégias para extrair ou inferir informações ocultas da parte mais informada, que neste contexto são os sistemas de IA e as empresas que os controlam. A aplicação desse conceito à IA reflete a necessidade de ferramentas que permitam tanto aos consumidores quanto aos órgãos de regulação compreenderem e auditarem as decisões tomadas por algoritmos. A XAI desempenha um papel central nesse processo ao fornecer mecanismos para que as decisões algorítmicas sejam não apenas transparentes, mas também compreensíveis, permitindo que ambos os grupos tomem decisões mais informadas e monitoradas (VILONE e LONGO,2021).

Contudo, para que a explicabilidade seja efetivamente implementada e beneficie tanto os consumidores quanto os reguladores, é essencial que as habilidades necessárias para desenvolvê-la sejam ensinadas e promovidas nos currículos de ciência da computação e engenharia. Essas habilidades incluem a modelagem transparente, que possibilita a criação de algoritmos interpretáveis desde o início; o uso de ferramentas, como SHAP (TREVISAN, 2022; VELASCO, 2020) e LIME (VISANI, 2020), que tornam as decisões de IA mais compreensíveis; a integração da ética e responsabilidade no desenvolvimento de sistemas de IA; e a capacidade de comunicação técnica, fundamental para traduzir a complexidade dos algoritmos em termos compreensíveis para consumidores e reguladores (WEBER et al., 2023).

Assim, a formação dos futuros profissionais de tecnologia deve não apenas focar na eficácia dos algoritmos, mas também na necessidade de torná-los transparentes e compreensíveis, garantindo que o avanço tecnológico vá de encontro a um consumo mais equilibrado e consciente, ao mesmo tempo que facilita o trabalho dos órgãos reguladores em garantir práticas justas e éticas.

A partir desta contextualização, propõe-se neste artigo explorar e avaliar os caminhos disponíveis para construir modelos de IA explicáveis. Utilizando o Método Analítico Hierárquico (MAH), este estudo avalia ferramentas de XAI, como SHAP, LIME e InterpretML, identificando a sua eficácia para promover transparência e ética (SANTOS, 2017). Os resultados destacam a importância de integrar XAI no currículo educacional para formar profissionais capazes de desenvolver sistemas de IA que reforcem a confiança dos consumidores e promovam práticas de consumo mais informadas e conscientes. Ao capacitar desenvolvedores com ferramentas e conhecimentos que promovam a transparência e a compreensão dos processos automatizados, podemos garantir que as futuras gerações de sistemas de IA sejam não apenas tecnicamente avançadas, mas também alinhadas com princípios éticos e de consumo consciente.

## 2. Inteligência Artificial Explicável e suas características

A Inteligência Artificial Explicável é uma abordagem que visa tornar os sistemas de IA mais transparentes e compreensíveis para os usuários. Em um contexto onde as decisões dos modelos de IA são frequentemente vistas como "caixas-pretas", a explicabilidade proporciona mecanismos para que essas decisões sejam interpretadas e auditadas, o que é essencial tanto para a confiança do usuário quanto para a conformidade regulatória (RIBEIRO e SAMEER, 2016; SETZU, 2021).

Segundo Philips et al. (2021), os quatro princípios da explicabilidade são: explicação, acurácia, significado e limites. A explicação refere-se à capacidade de uma IA fornecer uma explicação clara e compreensível para suas decisões e processos. A acurácia elaciona-se à exatidão da explicação fornecida, garantindo que a explicação seja consistente com a saída do modelo. O signficado enfatiza que a explicação deve ser significativa e relevante para os usuários, ajudando-os a entender o comportamento do modelo. Por fim, o limite destaca a importância de definir claramente as capacidades da IA e as situações em que as explicações podem ser menos confiáveis. Esses pilares são fundamentais para garantir que as explicações oferecidas pelas IA sejam úteis, precisas e compreensíveis para todos os tipos de usuários.

Com base nessas premissas, a explicabilidade pode ser desdobrada em várias categorias, cada uma oferecendo abordagens distintas para tornar os modelos de IA mais transparentes e compreensíveis, atendendo às necessidades de diferentes públicos e contextos de aplicação.

A explicabilidade numérica refere-se ao uso de métricas quantitativas para explicar as decisões dos modelos de IA. Esse tipo de explicabilidade é frequentemente usado em métodos agnósticos de modelo, que podem ser aplicados independentemente do algoritmo subjacente. Ferramentas como SHAP (SHapley Additive exPlanations) e LIME (Local Interpretable Model-agnostic Explanations) exemplificam esse tipo de abordagem, fornecendo uma análise quantitativa de como cada característica de entrada contribui para a saída do modelo.

A explicabilidade baseada em regras utiliza regras extraídas de modelos de IA para fornecer uma interpretação das decisões tomadas. Essa abordagem é particularmente útil em cenários onde é necessário entender a lógica subjacente a uma previsão, sendo amplamente empregada em ambientes regulatórios. Modelos específicos, como redes neurais, podem gerar explicações de três maneiras: métodos decomposicionais, que extraem regras a partir dos neurônios; métodos pedagógicos, que tratam a rede como uma caixa-preta; e métodos ecléticos, que combinam as duas abordagens.

A explicabilidade de modo textual é outra abordagem importante, onde as decisões são explicadas por meio de descrições em linguagem natural. Esse tipo de explicabilidade é particularmente útil para usuários não técnicos, pois facilita a compreensão das decisões tomadas pelo modelo.

A explicabilidade visual utiliza representações gráficas para explicar como os modelos de IA tomam decisões. Este método é especialmente útil em áreas como o reconhecimento de imagem, onde mapas de calor e gráficos de gradiente podem ser usados para destacar as áreas de uma imagem que mais influenciam a decisão do modelo. Tem-se ainda a explicabilidade mista que combina múltiplas abordagens, como numérica, visual e textual, para fornecer uma explicação mais robusta e acessível.

## 3. Materiais e métodos

O desenvolvimento do trabalho compreendeu seis etapas, sendo elas: a) justificativa da escolha das ferramentas computacionais, b) definição de um estudo de caso, c) implementação de um modelo caixa-preta utilizando Redes Neurais, d) realização de testes com as três ferramentas de XAI, e) definição dos critérios de avaliação, f) comparação dos resultados utilizando MAH a fim de estabelecer um ranking das ferramentas. As seções seguintes detalham as etapas do trabalho.

### 3.1 Escolha das ferramentas computacionais

Neste estudo, as ferramentas SHAP (SHapley Additive exPlanations), LIME (Local Interpretable Model-agnostic Explanations) e InterpretML foram selecionadas para análise devido à sua popularidade e capacidade de fornecer diferentes abordagens para a explicação de modelos de IA (TREVISAN, 2022; VISANE, 2020).

SHAP é uma ferramenta de explicabilidade que foi criada por Lundberg e Lee (2017). Ela permite explicar como um modelo de IA obteve um resultado. A técnica baseia-se nos valores de *Shapley*, que usam a teoria dos jogos para atribuir crédito para a previsão de um modelo a cada recurso ou valor de recurso. A ideia chave da técnica SHAP é calcular os valores de *Shapley* para cada forma da amostra a ser interpretada, onde cada valor de Shapley representa o impacto que a forma à qual está associado, gera na predição (LOPEZ, 2021). Os valores de *Shapley* são um conceito do campo da teoria dos jogos cooperativos, cujo objetivo é medir a contribuição de cada jogador para o jogo. Os valores de *Shapley* surgem do contexto onde “n” jogadores participam coletivamente obtendo uma recompensa “p” que se pretende distribuir de forma justa em cada um dos “n” jogadores, de acordo com a contribuição individual. Esta contribuição é um valor de *Shapley*.

LIME é uma ferramenta popular de interpretabilidade de modelos que permite entender como um modelo faz as previsões. A técnica LIME funciona para qualquer tipo de modelo de aprendizado de máquina, e visa explicar apenas uma pequena parte da função de aprendizagem (VISANI, 2020). A saída do LIME é uma lista de explicações, refletindo a contribuição de cada recurso para a previsão de uma amostra de dados. Isso fornece interpretabilidade local e também permite determinar quais alterações de recursos terão mais impacto na previsão (ALVES, 2021).

InterpretML é um pacote de código aberto que incorpora técnicas de interpretabilidade. Com este pacote, pode-se treinar modelos interpretáveis e explicar sistemas opacos. InterpretML ajuda a entender o comportamento local e global do seu modelo ou entender as razões por trás das previsões individuais (ALVES, 2021). Nos métodos de interpretação global fornecem uma visão geral do modelo em relação a todo conjunto de dados, utilizando os seguintes métodos: importância geral de recursos, agregação de recursos, análise de correlação e análise de interações.

### 3.2 Definição do estudo de caso

O estudo caso utilizado neste trabalho esta inserido na temática da visão computacional. O *dataset* Image Segmentation do Vision Group da University of Massachusetts foi constituído através de um processo automatizado de segmentação de imagens, no qual cada imagem foi dividida em regiões distintas com base em características visuais, como cor, textura e forma. Os segmentos resultantes representavam imagens de sete categorias, sendo elas: paredes de tijolo (brickface), imagens do céu (sky), folhagens (foliage), cimento (cement), janelas (window), caminhos (path) e terrenos com grama (grass). A Figura 1 apresenta as sete classes após segmentação de imagem e suas respectivas imagens.

Cada segmento foi descrito por um conjunto de atributos numéricos, incluindo métricas de cor, textura e propriedades geométricas, fornecendo uma representação detalhada das características visuais de cada categoria. Esses dados foram revisados e rotulados para garantir que cada segmento fosse representativo de sua classe, criando um conjunto de dados robusto e adequado para o treinamento e teste de algoritmos de aprendizado de máquina voltados para a segmentação e classificação de imagens.

**Figura 1 -** Ilustração de imagens que foram utilizadas para gerar o dataset.

**Fonte:** elaborado pelos autores (2024).

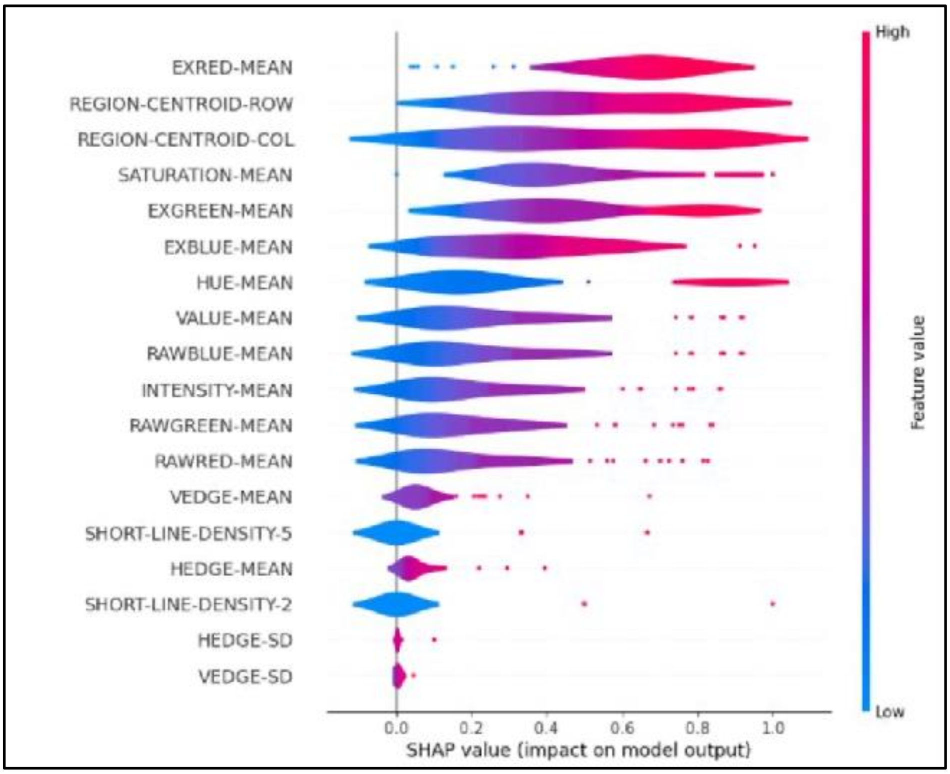
O conjunto contém 18 atributos para um conjunto de 2100 imagens. A combinação desses atributos permite capturar tanto as características globais (como cor e forma) quanto as locais (como textura e borda) dos segmentos de imagem, o que é essencial para realizar uma segmentação de imagem eficaz. Esses atributos são frequentemente usados como entrada em modelos de classificação para prever a classe de cada segmento de imagem com base em suas características visuais.

### 3.3 Treinamento da rede neural e testes de explicabilidade

A partir do conjunto de dados, projetou-se uma rede neural multicamadas apresentando: camada de entrada de 18 neurônios, camada intermediária de 13 neurônios e camada de saída de 7 neurônios. O algoritmo empregado no treinamento foi o Backpropagation. O método de amostragem utilizado foi a validação cruzada. O modelo gerado aprendeu a classificar as imagens das 7 classes com acurácia de 0.92.

A partir destes resultados prosseguiu-se para a etapa de testes com as ferramentas de explicabilidade. Realizou-se a implementação com as três ferramentas obtendo-se diversas visualizações e insights importantes sobre o funcionamento da rede neural. Para demonstrar os principais achados da pesquisa, apresenta-se algumas visualizações geradas para cada ferramenta. A Figura 2 apresenta um gráfico violino gerado pela biblioteca SHAP. Segundo Trevisan (2022), este gráfico é caracterizado por seu eixo horizontal, representado por um valor SHAP, enquanto a cor do ponto nos mostra se aquela observação (amostra ou instância) tem um valor maior ou menor, quando comparada a outras observações. As cores do gráfico Violino representam os valores médios das características naquela posição. As regiões vermelhas indicam que a maioria dos valores das características são altos, enquanto as regiões azuis indicam que a maioria dos valores das características são baixos. Pode-se observar que os atributos EXRED-MEAN (identifica a predominância da cor vermelha no segmento analisado), REGION-CENTROID-ROW e REGION-CENTROID-COL (indicam coordenadas de linha e coluna do centroide e ajudam a determinar a posição exata do segmento dentro da imagem) se destacam por possuírem valores mais altos.

**Figura 2** - Gráfico Violino gerado pela ferramenta SHAP

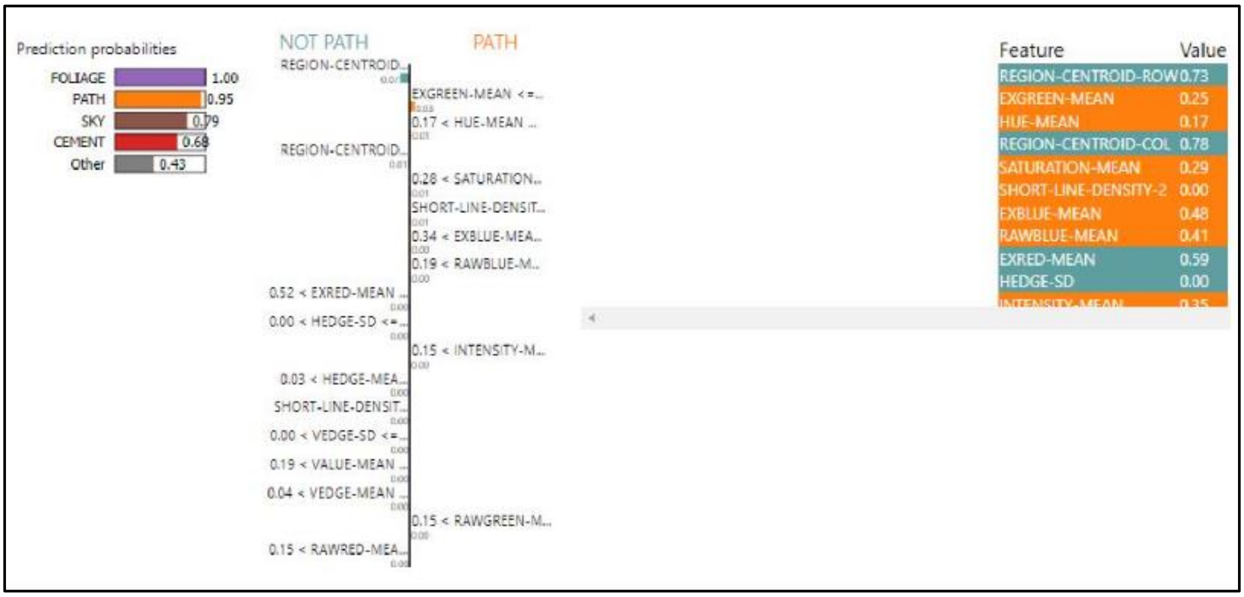



**Fonte**: elaborado pelos autores (2024)

A Figura 3 ilustra o gráfico LIME tabular explainer. Nesta visualização, para a classe de imagens de caminhos (path), os mesmos atributos EXRED-MEAN (identifica a predominância da cor vermelha no segmento analisado), REGION-CENTROID-

ROW e REGION-CENTROID-COL (indicam coordenadas de linha e coluna do centroide e ajudam a determinar a posição exata do segmento dentro da imagem) foram identificados como os mais relevantes.

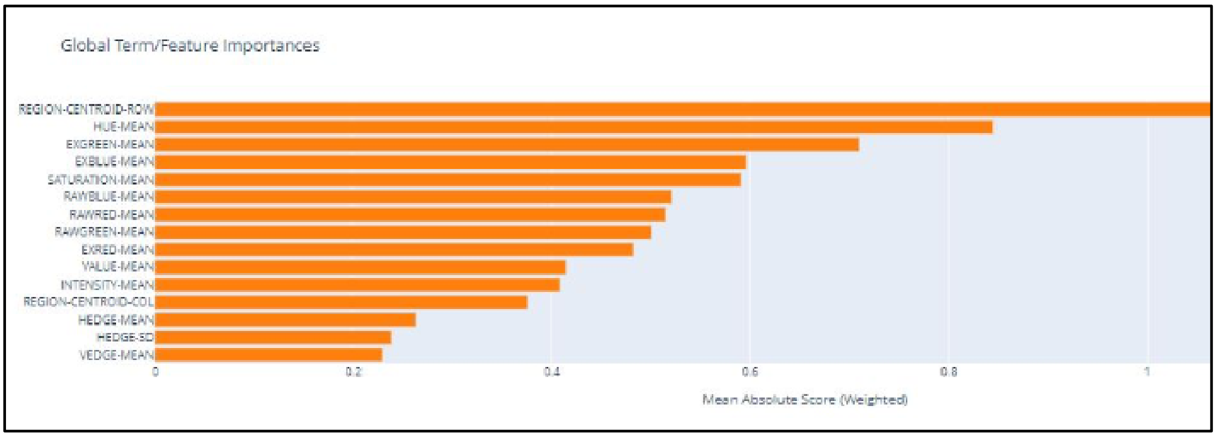
**Figura 3** - Gráfico tabular explainer gerado pela ferramenta LIME




**Fonte:** elaborado pelos autores (2024)

Já a ferramenta InterpretML (Figura 4) identificou com maior relevância na tarefa de classificação os atributos: REGION-CENTROID-ROW, HUE-MEAN (identifica a tonalidade geral do segmento: vermelho, azul ou amarelo) e EXGREEN-MEAN (representa a média do excesso de verde nos pixels do segmento).

**Figura 4** - Histograma vertical gerado pela ferramenta InterpretML




**Fonte:** elabora pelos autores (2024)

Diversas visualizações foram produzidas e estão detalhamente descritas no trabalho de conclusão de curso sobre o tema (BAVARESCO, 2023).

### 3.4 Critérios de avaliação

Para assegurar uma avaliação abrangente e relevante das ferramentas de explicabilidade, foram definidos 18 critérios principais, compreendo quatro capacidades: explicabilidade, flexibilidade, eficiência computacional, e facilidade de uso. Entende-se que os seguintes critérios permitem avaliar a capacidade das ferramentas computacionais em fornecer recursos e funcionalidades para desenvolver sistemas de IA transparentes:

1. Transparência: A biblioteca fornece métodos para explicar os resultados e processos in- ternos do modelo de maneira compreensível?

2. Interpretabilidade: A biblioteca oferece recursos para entender e interpretar o modelo, permitindo a análise dos fatores que influenciam as decisões tomadas?

3. Métodos de Explicação Numérica: A biblioteca fornece métodos específicos para explicar as decisões do modelo de forma numérica, identificando a contribuição individual de cada feature para o resultado final?

4. Explicabilidade por Regras: A biblioteca suporta a geração automática de regras lógicas ou heurísticas que explicam o comportamento do modelo, permitindo que os usuários compreendam as condições e padrões subjacentes às decisões tomadas?

5. Explicabilidade por Árvores de Decisão: A biblioteca possui recursos para extrair e interpretar árvores de decisão geradas pelo modelo, permitindo uma explicação mais intuitiva e baseada em regras?

6. Explicabilidade por Modelos Lineares: A biblioteca suporta a interpretação de modelos lineares, como regressão linear ou regressão logística, permitindo a análise dos coeficien- tes e efeitos de cada feature no resultado?

7. Explicabilidade por Aproximação de Modelos Complexos: A biblioteca oferece métodos para aproximar modelos complexos, como redes neurais, por modelos mais simples, como modelos lineares ou árvores de decisão, com o objetivo de fornecer explicações mais interpretáveis?

8. Métricas de Importância de Características: A biblioteca fornece métricas que medem a importância relativa das features no modelo, como importância baseada em ganho de informação, coeficientes de correlação, etc.?

9. Avaliação de Robustez: A biblioteca permite avaliar a robustez da explicabilidade fornecida pelo modelo, por meio de análises de sensibilidade, avaliação de perturbações nas features e quantificação do impacto nas explicações?

10. Comparação de Métodos de Explicação: A biblioteca oferece a possibilidade de comparar diferentes métodos de explicação disponíveis, permitindo que os usuários avaliem a consistência e confiabilidade das explicações geradas?

11. Personalização de Explicações: A biblioteca permite a personalização das explicações, permitindo que os usuários ajustem parâmetros de relevância, filtrem informações desnecessárias e adaptem as explicações aos seus requisitos específicos?

12. Documentação e Suporte: A biblioteca possui documentação abrangente que explique os métodos de explicação disponíveis, exemplos de uso e casos de uso? Além disso, a biblioteca possui uma comunidade ativa que fornece suporte e compartilha conhecimentos sobre a aplicação da explicabilidade em Python?

13. Visualização das Explicações: A biblioteca oferece recursos de visualização para tornar as explicações mais compreensíveis e intuitivas, como gráficos, diagramas, mapas de calor, ou outras representações visuais que ajudem a transmitir a lógica e o impacto das features no modelo?

14. Compatibilidade com Frameworks: A biblioteca é compatível com os principais fra- meworks de Machine Learning em Python, como TensorFlow, PyTorch, Scikit-learn, etc.?

15. Facilidade de Uso: A biblioteca é fácil de usar e possui uma documentação abrangente e clara?

16. Flexibilidade: A biblioteca oferece flexibilidade suficiente para adaptar e personalizar as técnicas de explicabilidade de acordo com as necessidades específicas?

17. Eficiência Computacional: A biblioteca é eficiente em termos de recursos computacionais, especialmente para modelos de grande escala?

18. Suporte à Comunidade: A biblioteca possui uma comunidade ativa de desenvolvedores e usuários que podem fornecer suporte e compartilhar conhecimento?

### 3.5. Método analítico hierárquico

O Método Analítico Hierárquico (MAH) foi escolhido como a principal metodologia para a avaliação comparativa das ferramentas SHAP, LIME e InterpretML. O MAH é uma técnica de tomada de decisão multicritério que permite comparar diferentes alternativas com base em vários critérios, tanto quantitativos quanto qualitativos (SANTOS, 2017). Os critérios e as alternativas são comparados entre si, dois a dois (pares), em termos de sua importância ou preferência relativa. Essa comparação é feita usando uma escala de julgamento, geralmente de 1 a 9, onde 1 significa igual importância e 9 significa que um critério é extremamente mais importante que o outro.

A partir das comparações pares, o MAH calcula os pesos relativos de cada critério e alternativa. Isso é feito por meio de métodos matemáticos que envolvem a normalização das comparações e a obtenção de vetores próprios. Os pesos indicam a importância relativa de cada critério em relação ao objetivo e de cada alternativa em relação aos critérios.

A hierarquia do MAH foi estruturada em três níveis principais: o objetivo de identificar a melhor ferramenta de XAI, os critérios de avaliação divididos nas quatro categorias descritas previamente, e as alternativas, que são as três ferramentas de XAI avaliadas.

As ferramentas foram comparadas par a par em relação a cada critério. Por exemplo, a clareza das explicações fornecidas por SHAP foi comparada com as de LIME e InterpretML. Essa comparação foi realizada com base em experimentos práticos e na análise de literatura especializada, garantindo uma avaliação precisa e fundamentada.

Os pesos atribuídos a cada critério foram determinados com base em sua importância para o ensino e a prática de IA em um contexto de compreensão por estudantes e explicabilidade para usuários. Esses pesos foram estabelecidos após consultas a especialistas na área e análise de estudos anteriores, assegurando que as prioridades educacionais e éticas fossem devidamente refletidas.

## 4. Resultados da avaliação das ferramentas de explicabilidade

A síntese dos resultados foi realizada utilizando as matrizes de comparação do MAH, que geraram um ranking das ferramentas com base na pontuação agregada de cada uma em relação aos critérios avaliados. Esta etapa não apresenta os resultados, mas descreve o processo que permitiu determinar a ferramenta mais adequada para ser utilizada no ensino de computação, focado em promover um consumo consciente e ético das tecnologias de IA.

As ferramentas foram submetidas a um processo de testes rigoroso, que envolveu a implementação de cada uma delas em um ambiente controlado. Para cada ferramenta, foram realizados múltiplos testes que envolveram a aplicação de modelos de IA aos conjuntos de dados escolhidos, seguidos da utilização das ferramentas de explicabilidade para gerar explicações sobre as decisões dos modelos. Esses testes foram replicados em diferentes cenários para avaliar a consistência dos resultados.

Durante a execução dos testes, foram coletados dados referentes ao tempo de execução, uso de recursos computacionais, clareza e qualidade das explicações geradas, entre outros fatores relevantes para a avaliação dos critérios estabelecidos.

O uso do MAH como metodologia central assegurou que a comparação entre as ferramentas fosse conduzida de maneira estruturada e criteriosa, refletindo as necessidades da área. A Tabela 1 apresenta a avaliação de cada ferramenta segundo os critérios.

**Tabela 1** - Resultados da comparação entre ferramentas para explicabilidade.

| Critérios de avaliação | SHAP | LIME | InterpretML |
|---|---|---|---|
| 1.Transparência | S | S | S |
| 2.Interpretabilidade | S | S | S |
| 3.Métodos de Explicação Numérica | S | S | S |
| 4.Explicabilidade por regras | N | S | S |
| 5.Explicabilidade por árvore de decisão | N | S | S |
| 6.Explicabilidade por Modelos Lineares | S | S | S |
| 7.Explicabilidade por Aproximação de Modelos Complexos | S | S | S |
| 8.Métricas de importância de Features | S | S | S |
| 9.Avaliação de Robustez | S | S | S |
| 10.Comparação de Métodos de Explicação | S | S | S |
| 11.Personalização de Explicações | S | S | S |
| 12.Documentação e Suporte | S | S | S |
| 13.Visualização das Explicações | S | S | S |
| 14.Compatibilidade com Frameworks | S | S | S |
| 15.Facilidade de uso | S | S | S |
| 16.Flexibilidade | S | S | N |
| 17.Eficiência computacional | N | S | S |
| 18.Suporte à Comunidade | S | S | S |

**Fonte:** elaborado pelos autores (2024).

Na etapa final foi estabelecido um ranking de ferramentas de XAI em termos dos critérios de avaliação utilizados (Tabela 2). Com isso, foi possível observar qual das ferramentas possui a maior pontuação e ver quais são as que melhor atendem os critérios de avaliação.

**Tabela 2 -** Ranking final e pontuação das ferramentas para a XAI.

| Ranking | Biblioteca | Pontuação Total | Critérios Atendidos |
|---|---|---|---|
| 1 | LIME | 6,42 | 18 |
| 2 | InterpretML | 6,02 | 17 |
| 3 | SHAP | 5,22 | 15 |

**Fonte**: elaborado pelos autores (2024).

A ferramenta que obteve a maior pontuação total foi a biblioteca LIME. Em linhas gerais, com um somatório de 6,42, a LIME foi a ferramenta que obteve o melhor desempenho geral. Sua combinação de alta explicabilidade, flexibilidade, e eficiência computacional a torna ideal para utilização em XAI. A SHAP obteve um somatório de 6,02, destacando-se particularmente em diversidade de modos de explicação, mas sendo penalizada por sua complexidade e maior exigência de recursos computacionais. Por fim, a InterpretML, com um somatório de 5,22, InterpretML se mostrou adequada para cenários específicos, mas menos versátil e eficiente em comparação com LIME e SHAP, especialmente em contextos mais complexos.

Em termos de requisitos, a explicabilidade é um dos aspectos centrais na avaliação de ferramentas de XAI, especialmente no contexto deste trabalho, onde a clareza das explicações é essencial para a compreensão de usuários não especialistas. LIME foi a ferramenta mais bem avaliada nesse critério, recebendo uma pontuação de 9. Sua abordagem, que se baseia em perturbações nas entradas para gerar explicações locais, torna as previsões do modelo altamente compreensíveis, o que é extremamente útil para estudantes. SHAP, por outro lado, recebeu uma pontuação de 8. Embora suas explicações sejam detalhadas e robustas, a complexidade de seus cálculos pode ser um desafio para alunos iniciantes. InterpretML ficou atrás com uma pontuação de 7, sendo mais adequada para modelos mais simples e menos eficaz em cenários complexos.

Quando se trata de flexibilidade, LIME mais uma vez se destacou, com uma pontuação de 9, devido à sua capacidade de se adaptar a diferentes tipos de modelos e dados, incluindo dados tabulares, textuais e de imagens. Essa versatilidade é um fator crucial no ensino de IA, que muitas vezes abrange diversas abordagens e aplicações. SHAP também obteve uma boa pontuação (8) em flexibilidade, mas sua complexidade técnica pode limitar sua aplicabilidade em certos cenários. InterpretML, com uma pontuação de 7, mostrou-se mais restrita em termos de compatibilidade com diferentes tipos de dados e modelos, sendo mais adequada para explicações de árvores de decisão e modelos baseados em regras.

No critério de eficiência computacional, LIME novamente se destacou, recebendo a maior pontuação (9). Sua rápida execução e o baixo consumo de recursos tornam-na ideal para uso em ambientes educacionais, onde o tempo e a infraestrutura podem ser limitados. SHAP, embora poderoso, teve uma pontuação menor (7) devido ao seu maior tempo de execução e consumo de recursos, especialmente em modelos mais complexos. InterpretML teve uma pontuação intermediária (8), apresentando bom desempenho em modelos simples, mas perdendo eficiência à medida que a complexidade aumenta.

Por fim, a facilidade de uso é um critério importante para a adoção de qualquer ferramenta em um ambiente educacional. LIME recebeu uma pontuação de 8, sendo relativamente fácil de aprender e usar, com uma boa documentação e uma comunidade ativa que oferece suporte adicional. SHAP, por outro lado, recebeu uma pontuação de 7. Embora tenha uma curva de aprendizado mais suave para usuários avançados, sua complexidade pode ser um obstáculo para iniciantes. InterpretML também recebeu uma pontuação de 7, destacando-se por sua simplicidade, mas com limitações em aplicações mais avançadas.

A combinação da matriz de pontuação com a análise global dos resultados nos permite concluir que LIME é a ferramenta mais indicada atualmente para uso em contextos educacionais, especialmente em cenários voltados para a explicabilidade de IA. Sua clareza, flexibilidade, eficiência computacional e facilidade de uso a tornam uma escolha sólida para professores e estudantes que buscam compreender e aplicar técnicas de IA explicável de forma prática e acessível. SHAP, apesar de sua complexidade, é uma excelente escolha para cenários onde a profundidade das explicações é crucial, enquanto InterpretML é mais adequada para cursos introdutórios ou para a explicação de modelos mais simples.

## 5. Considerações finais

Neste estudo, foi analisada a eficácia das ferramentas de XAI (SHAP, LIME, InterpretML) em promover a transparência e a compreensão dos modelos de inteligência artificial, com foco na redução da assimetria de informação nas relações de consumo. Através da aplicação do MAH foi possível avaliar e comparar essas ferramentas com base em critérios de explicabilidade, facilidade de uso, e impacto potencial na educação e no consumo consciente.

SHAP se destaca na literatura por sua capacidade de fornecer explicações tanto globais quanto locais. Baseado na teoria dos valores de Shapley, SHAP calcula a contribuição de cada característica para a saída do modelo, tornando-o uma ferramenta robusta e versátil para entender o comportamento geral e específico dos modelos de IA. Estudos têm mostrado que SHAP é especialmente útil em contextos onde é necessário interpretar as decisões do modelo em nível detalhado, permitindo uma visão clara de como cada entrada afeta a previsão.

LIME, por sua vez, é amplamente reconhecido por sua eficácia em gerar explicações locais. Ao criar modelos substitutos simples para uma pequena região do espaço de entrada, LIME fornece insights sobre previsões individuais, o que é crucial em aplicações onde a justificativa para uma decisão específica do modelo precisa ser compreendida, como em diagnósticos médicos ou avaliações de crédito. A literatura destaca LIME por sua flexibilidade em funcionar com qualquer modelo, sem necessidade de modificações no modelo original.

InterpretML integra múltiplas técnicas de explicabilidade, oferecendo uma abordagem abrangente que combina modelos intrinsecamente interpretáveis e métodos pós-hoc. Essa capacidade de combinar diferentes técnicas faz de InterpretML uma ferramenta valiosa no contexto educacional, onde é importante proporcionar aos alunos uma compreensão ampla das diferentes abordagens de XAI. Além disso, a literatura aponta que InterpretML é eficaz para explorar tanto explicações locais quanto globais, facilitando a adoção de XAI em uma variedade de aplicações.

No contexto do ensino de IA e das práticas de consumo consciente, as ferramentas de XAI desempenham um papel crucial. A inclusão dessas tecnologias no currículo educacional é vital para a formação de futuros profissionais que possam desenvolver sistemas de IA que não apenas sejam eficazes, mas também transparentes e éticos. Ao capacitar estudantes e profissionais com o conhecimento de XAI, criamos as bases para um mercado mais informado e equilibrado, onde consumidores e reguladores podem tomar decisões baseadas em informações claras e confiáveis.

Por fim, este estudo contribui para o entendimento das ferramentas de XAI como facilitadoras da explicabilidade em IA, destacando a importância de sua aplicação tanto no campo educacional quanto nas relações de consumo. As conclusões aqui apresentadas reforçam a necessidade de continuar explorando e desenvolvendo tecnologias que promovam a transparência e a equidade na era digital, conforme indicado pela literatura.

## Referências

ALVES, O. M. D. A. M. A. S. Da “caixa-preta” à “caixa de vidro”: o uso da explainable artificial intelligence (xai) para reduzir a opacidade e enfrentar o enviesamento em modelos algorítmicos. Dossiê – Inteligência Artificial, Ética e Epistemologia, v. 18, p. 373, 2021.

ALVEY, B.J., ANDERSON, D.T., KELLER, J.M. Linguistic Comparisons of Black Box Models. 2024 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), Yokohama, Japan, 2024, pp. 1-9.

BAVARESCO,M.Z. Avaliação de Bibliotecas para a Explicabilidade da Inteligência Artificial. Caxias do Sul: Trabalho de conclusão de curso, 2023. 79p.

DOMINGUES, Juliana Oliveira e SILVA, Alaís Aparecida Bonelli da e SOUZA, Henrique Monteiro Araujo de. **Inteligência artificial nas relações de consumo: reflexões à luz do histórico recente.** Inteligência artificial : sociedade, economia e Estado. Tradução . São Paulo, SP: Thomson Reuters Brasil, 2021.

PHILLIPS, P. Jonathon; et al. **Four principles of explainable artificial intelligence**. 2021. Disponível em: https://nvlpubs.nist.gov/nistpubs/ir/2021/NIST.IR.8312.pdf. Acesso em 03-2024.

LUGER, G. F. Inteligência Artificial. 6. ed. [S.l.]: PEARSON, 2013. 632 p.

LUNDBERG, S. A unified approach to interpreting model predictions. Paper with Code, 2017.

LÓPEZ, F. Shap: Shapley additive explanations. Towards Data Science, 2021.

MISHRA, P. Explainable AI Recipes. 1. ed. [S.l.]: Apress Berkeley, CA, 2023. 254 p.

RIBEIRO, C. G. M. T.; SAMEER, S. Local interpretable model-agnostic explanations (LIME): An introduction. O'Reilly, 2016.

ROSSETTI, R.; ANGELUCI, A. Ética algorítmica: questões e desafios éticos do avanço tecnológico da sociedade da informação. Galáxia (São Paulo), Programa de Estudos Pós-graduados em Comunicação e Semiótica - PUC-SP, n. 46, p. e50301, 2021. ISSN 1982-2553.

SANTOS, V. **O que é AHP ou Processo Hierárquico Analítico e seus usos?** FM2S, 2017.

SETZU, M. et al. **GlocalX - From local to global explanations of black box AI models**. Artificial Intelligence, Elsevier, v. 294, p. 103457, 2021.

STIGLITZ, J. E. **The Theory of ‘Screening, ’Education, and the Distribution of Income.** The American Economic Review, 65(3), 283-300, 1975.

TREVISAN, V. **Using SHAP values to explain how your machine learning model works**. Towards Data Science, 2022.

VELASCO, F. L. **Shap: Explicações aditivas de Shapley**. Ichi Pro, 2020.

VILONE, G.; LONGO, L. **Classification of explainable artificial intelligence methods through their output formats**. Machine Learning and Knowledge Extraction, v. 3, n. 3, p. 615–661, 2021. ISSN 2504-4990.

VISANI, G. **LIME: Explain machine learning predictions**. Towards Data Science, 2020.

WEBER, S.; LAPUSCHKIN, A.; WICK, A.; SAMEK, W.; BINDER, A. **Beyond explaining: Opportunities and challenges of XAI-based model improvement**. Information Fusion, v. 92, p. 154–176, 2023. ISSN 1566-2535.

WEBER, P., CARL, K.V.; HINZ, O. **Applications of Explainable Artificial Intelligence in Finance -** a systematic review of Finance, Information Systems, and Computer Science literature. Manag Rev Q 74, 867–907, 2024.