XAI LIME Tool: Uma ferramenta que avalia explicabilidade em Inteligência Artificial

CENTRO DE INFORMÁTICA
UNIVERSIDADE FEDERAL DA PARAÍBA

Caio Vinícius Alves Lima

João Pessoa, 2024

Catalogação na publicação
Seção de Catalogação e Classificação

L732x Lima, Caio Vinicius Alves.
XAI LIME Tool: uma ferramenta que avalia explicabilidade em Inteligência Artificial / Caio Vinicius Alves Lima. - João Pessoa, 2024.
33 f. : il.

Orientação: Natasha Correia Queiroz Lino.
TCC (Graduação) - UFPB/CI.

1. Explicabilidade. 2. Inteligência artificial. 3. LIME. 4. Faithfulness. I. Lino, Natasha Correia Queiroz. II. Título.

UFPB/CI CDU 004.8

Elaborado por Michelle de Kassia Fonseca Barbosa - CRB-738

XAI LIME Tool: Uma ferramenta que avalia explicabilidade em Inteligência Artificial

Caio Vinícius Alves Lima

Centro de Informática – Universidade Federal da Paraíba (UFPB) Rua dos Escoteiros, Mangabeira VII, João Pessoa, Paraíba, Brasil CEP: 58058-600

caio81887@gmail.com

Abstract. This paper proposes a tool to evaluate the quality of explanations generated by the LIME algorithm, widely used to increase the interpretability of opaque machine learning models. The tool applies the *faithfulness* metric, which verifies if the explanations truly reflect the behavior of the original model, reinforcing transparency and reliability in predictions. The system was tested with the Pima Indians Diabetes Database and the Random Forest model. Based on this, local explanations of predictions were generated, allowing a detailed analysis of how the attributes influence the model's decisions.

Resumo. Este trabalho propõe uma ferramenta para avaliar a qualidade das explicações geradas pelo algoritmo LIME, amplamente utilizado para aumentar a interpretabilidade de modelos de aprendizado de máquina opacos. A ferramenta aplica a métrica *faithfulness*, que verifica se as explicações realmente refletem o comportamento do modelo original, reforçando a transparência e confiabilidade nas previsões. O sistema foi testado com a base de dados Pima Indians Diabetes Database e com o método Random Forest. A partir disso, foram geradas explicações locais das previsões, possibilitando uma análise detalhada sobre como os atributos influenciam nas decisões do modelo de IA.

1. Introdução

A inteligência artificial (IA) experimentou um crescimento exponencial na última década, consolidando-se especialmente em áreas críticas de tomada de decisão, tais como medicina, justiça, sistemas de recomendação, concessão de crédito e processos de emprego. Entretanto, o aumento da complexidade dos modelos de IA, como aqueles baseados em aprendizado de máquina e redes neurais profundas, tem reduzido a interpretabilidade desses sistemas. Esse fenômeno torna a tarefa de explicar a eficácia dos sistemas de IA cada vez mais desafiadora para os usuários [Dierle e Otávio 2023].

Com isso, surge uma demanda crescente por IA que seja responsável e transparente. Esse objetivo é alcançado através do fornecimento de explicações claras e compreensíveis, que assegurem os princípios de justiça, segurança e privacidade.

Neste contexto, este trabalho tem como **objetivo geral** propor uma ferramenta para avaliar a confiabilidade das explicações geradas pelo algoritmo LIME (*Local*

Interpretable Model-agnostic Explanations) [Guestrin, Singh e Ribeiro 2016], utilizado em sistemas de Inteligência Artificial, por meio da métrica de *faithfulness* [Lundberg e Lee 2017]. A proposta busca garantir que as explicações fornecidas pelo LIME realmente reflitam o comportamento do modelo original, assegurando maior transparência e confiança nas decisões geradas por algoritmos de aprendizado de máquina.

Para alcançar esse objetivo, foram delineados alguns **objetivos específicos**, que orientaram o desenvolvimento da ferramenta:

1. O primeiro objetivo consistiu em selecionar e aplicar um modelo de aprendizagem de máquina do tipo caixa-preta sobre uma base de dados previamente escolhida. Esse modelo, conhecido por sua complexidade e opacidade, será utilizado como objeto de estudo para a análise das explicações.
2. Em seguida, foi necessário definir e aplicar o algoritmo de explicabilidade LIME, que será responsável por gerar explicações locais para as previsões feitas pelos modelos selecionados na ferramenta desenvolvida.
3. Por fim, a métrica *faithfulness* foi aplicada para avaliar a qualidade das explicações geradas pelo LIME, verificando se elas são realmente consistentes e aderentes ao comportamento do modelo original. Com isso, buscou-se garantir que as explicações produzidas sejam tanto compreensíveis quanto fidedignas, assegurando a confiabilidade das mesmas.

2. Fundamentação Teórica

A fundamentação teórica deste trabalho explora os conceitos que sustentam a explicabilidade em IA, com foco na aplicação do método LIME e na importância da métrica *faithfulness* para assegurar que as explicações fornecidas sejam verdadeiramente representativas das decisões do modelo original. A partir disso, serão apresentados os resultados e as oportunidades na aplicação dessas técnicas, destacando sua relevância para a construção de sistemas de IA mais transparentes e confiáveis.

A diferença entre interpretabilidade e explicabilidade é um tema central no campo da Inteligência Artificial (IA), especialmente quando se trata de modelos de aprendizado de máquina. Embora esses dois conceitos estejam relacionados, eles possuem definições distintas que se aplicam a diferentes contextos.

A interpretabilidade refere-se à facilidade com que um ser humano pode compreender o funcionamento interno de um modelo de IA ou a lógica por trás de suas decisões. Em outras palavras, um modelo é interpretável quando o processo que leva às suas previsões

pode ser entendido diretamente, sem a necessidade de técnicas adicionais. Modelos como regressão linear ou logística e árvores de decisão são exemplos de modelos intrinsecamente interpretáveis, pois oferecem uma visão clara da contribuição de cada variável para o resultado final. Por exemplo, na regressão linear, os coeficientes das variáveis explicam diretamente a influência de cada fator sobre a previsão, enquanto em árvores de decisão, o caminho de decisão permite rastrear como cada instância foi classificada ou prevista, a explicabilidade trata da capacidade de gerar explicações compreensíveis para modelos que, por sua complexidade, não são interpretáveis por natureza. Isso é feito por meio de métodos externos (post-hoc), que analisam o comportamento do modelo e criam narrativas que auxiliam o usuário a entender as previsões. Ferramentas como LIME (Local Interpretable Model-agnostic Explanations) e SHAP (SHapley Additive exPlanations) exemplificam técnicas de explicabilidade. Essas ferramentas permitem que modelos complexos, como redes neurais profundas ou florestas aleatórias, sejam explicados ao destacar as variáveis mais influentes para cada previsão. Por exemplo, o LIME cria um modelo local simplificado para explicar previsões específicas, enquanto o SHAP utiliza a teoria dos jogos para calcular a importância de cada variável em relação à previsão do modelo, interpretabilidade e explicabilidade está na abordagem. A interpretabilidade é uma propriedade intrínseca de modelos simples e transparentes, que permite compreender diretamente sua lógica interna. Já a explicabilidade é um recurso adicional, necessário para tornar compreensíveis modelos de caixa-preta, que priorizam a precisão e a complexidade em detrimento da simplicidade. Este entendimento é crucial paras de IA que sejam tanto eficazes quanto confiáveis, especialmente em aplicações críticas, como saúde e justiça, onde as decisões precisam ser justificáveis e transparentes.

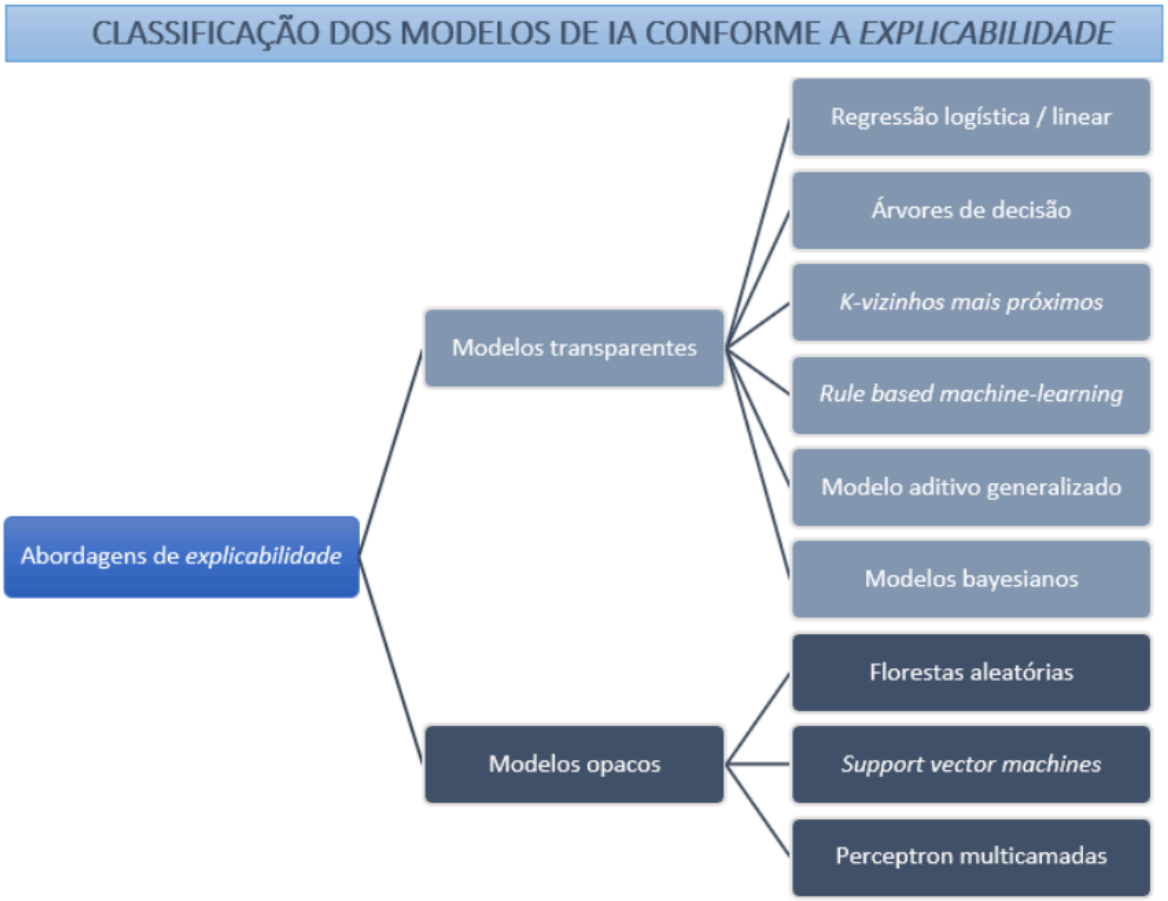
2.1. Modelos de Aprendizagem de Máquina

Alguns algoritmos são auto-interpretáveis, mas outros precisam ser explicados com algoritmos auxiliares de explicabilidade. Um modelo de aprendizagem de máquina é considerado um modelo interpretável ou transparente quando este não precisa de adições de técnicas para ser compreendido por um humano. Alguns exemplos desses sistemas são os modelos de regressão linear/logística; árvores de decisão, k-vizinhos que mais próximos, RBML(Rule-Based Machine Learning), GAM(Generalized Additive Models) e modelos de bayesianos. Entretanto, mesmo que os modelos de algoritmos sejam interpretáveis, isso não os isenta de explicabilidade. Pelo contrário, sua explicabilidade será ainda mais favorecida. Por exemplo, considere uma árvore de decisão. Modelos baseados em árvore cortam muitas vezes os dados com algum valor limite de um recurso.

Quando a árvore é dividida, diferenciam-se subconjuntos do conjunto de dados, cada instância pertence a um subconjunto. As previsões individuais de uma árvore de decisão podem ser explicadas decompondo-se o caminho de decisão em um componente por recurso. Pode-se rastrear uma decisão por meio da árvore e explicar uma previsão pelas contribuições adicionadas em cada nó de decisão.

A seguir vemos uma organização hierárquica de modelos de abordagens da explicabilidade:

Figura 1 - Classificação dos modelos de IA conforme as abordagens de explicabilidade



Fonte: [Dierle e Otávio 2023]

Em contraponto aos “modelos transparentes” de aprendizagem de máquina, ainda existem os “modelos opacos”, cuja compreensão irá requerer um processo de explicação adicional, chamado de “explicabilidade post-hoc”. Entre esses modelos opacos, é possível apontar as Support Vector Machines(SVM), o Perceptron multicamadas e as florestas de decisão aleatórias (Random Forests) [Doshi-Velez e Kim 2017], sendo este último utilizado no trabalho. Portanto será explicado mais detalhadamente na próxima sessão.

2.1.1. Random Forest

O Random Forest é um algoritmo de aprendizagem de máquina desenvolvido por Leo Breiman em 2001, amplamente utilizado tanto para tarefas de classificação quanto de regressão. Ele faz parte de uma classe de métodos conhecidos como aprendizado de conjunto (*ensemble learning*), que combinam vários modelos simples para construir um modelo mais robusto e preciso. A principal ideia por trás do Random Forest é a utilização de várias árvores de decisão individuais, cujas previsões são agregadas para gerar um resultado final mais confiável, o que reduz a variabilidade e o risco de *overfitting*, ou seja, quando o modelo se ajusta excessivamente aos dados de treinamento, resultando em baixa performance em novos dados. Como tal, o Random Forest se destaca por melhorar a precisão e generalização dos modelos, ao mesmo tempo em que se mantém resistente a ruídos nos dados [Breiman, L. 2001].

O funcionamento do Random Forest se baseia em três etapas principais. Primeiro, na fase de treinamento, o algoritmo constrói diversas árvores de decisão, onde cada uma é treinada em um subconjunto aleatório dos dados de treino. Esse subconjunto é gerado através do método de *bootstrap*, uma técnica de amostragem com reposição, o que significa que algumas instâncias podem ser repetidas no conjunto de treino de uma árvore, enquanto outras podem ser ignoradas. Segundo, em cada nó da árvore, o algoritmo seleciona aleatoriamente um conjunto limitado de variáveis para determinar a melhor divisão dos dados. Esse processo aumenta a diversidade das árvores e impede que elas se tornem muito parecidas. Por fim, a previsão final do Random Forest é obtida por meio de um processo de agregação. No caso de classificação, cada árvore "vota" em uma classe e a classe com mais votos é a previsão final. No caso de regressão, a média das previsões das árvores é utilizada [Breiman, L. 2001].

Esse algoritmo em questão é menos suscetível ao *overfitting*, pois a diversidade introduzida pela aleatoriedade na construção das árvores impede que o modelo se ajuste demais aos dados de treinamento. Além disso, o modelo é extremamente robusto e preciso, especialmente quando se lida com dados ruidosos ou incompletos, uma vez que a combinação de várias árvores reduz o impacto de ruídos. Outra vantagem importante é que o algoritmo permite a extração de importância das variáveis, ou seja, ele calcula a relevância de cada variável, o que ajuda a entender quais variáveis têm maior impacto nas previsões, além de ser conveniente para a métrica *faithfulness*, a qual utiliza essa relevância das variáveis para calcular a fidelidade das explicações [Breiman, L. 2001].

2.2. O que é Explicabilidade?

No campo da IA, explicabilidade refere-se à capacidade de entender os motivos e os detalhes por trás de uma decisão algorítmica. Portanto, sistemas de inteligência artificial explicável seriam aqueles “capazes de explicar seus fundamentos, caracterizar seus pontos fortes e fracos e transmitir uma compreensão acerca de suas condutas futuras”, isso se dá através de métricas estabelecidas e analisadas para cada tipo de situação [Lipton 2017], [Dierle e Otávio 2023]. Assim como ocorre com explicações em outras

áreas da ciência, a compreensão dos processos inerentes à IA precisa utilizar representações comunicáveis, como, por exemplo, expressões linguísticas ou lógicas, sentenças matemáticas e diagramas visuais, algo de forma que quem estiver fazendo uso possa compreender a explicação da decisão final da IA. Dessa forma, enquanto a opacidade cria uma "caixa preta" que limita o entendimento humano sobre as decisões de um sistema de IA, a explicabilidade resulta no oposto, ou seja, uma "caixa de vidro" ou "caixa branca" que permite a compreensão dos processos internos por trás de um resultado algorítmico [Pedreschi, Giannotti, Turini, Ruggieri, Monreale e Guidotti 2018].

2.2.1. Método LIME de Explicabilidade

O método LIME(*Local Interpretable Model-agnostic Explanations*) é uma técnica de explicação que visa aumentar a interpretabilidade dos modelos de aprendizado de máquina. O método foi introduzido para abordar o desafio da "caixa preta" presente em muitos modelos de aprendizado de máquina complexos, como redes neurais profundas, que têm um desempenho superior, mas cuja lógica de funcionamento é difícil de entender para humanos. LIME explica as previsões de qualquer classificador ao aprender um modelo interpretável localmente ao redor da previsão. Basicamente, ele cria uma versão simplificada do modelo em torno de uma previsão específica que está sendo explicada [Guestrin, Singh e Ribeiro 2016]. Algumas áreas que utilizam esse método de explicação em seu benefício, como Medicina: Para explicar previsões feitas por modelos diagnósticos; Marketing: Para entender quais características de um cliente são mais influentes em modelos de previsão de comportamento de compra; Finanças: Para a análise de risco e decisões de crédito.

O LIME gera um conjunto de amostras perturbadas do dado original que está sendo explicado. Por exemplo, se estivermos explicando a previsão de um classificador de texto, LIME pode gerar amostras perturbadas ao remover palavras de um documento. Após gerar esse conjunto de dados perturbados, LIME treina um modelo linear simples ou uma árvore de decisão nas amostras perturbadas usando as previsões do modelo original como rótulos. Esse modelo explicável local é ajustado para aproximar a superfície de decisão do modelo original perto do ponto específico de interesse. Por fim, na geração da explicação, os coeficientes do modelo local interpretável são então usados para entender a importância de cada característica na previsão do modelo original para o dado específico [Guestrin, Singh e Ribeiro 2016].

2.2.2. Métrica *faithfulness* de Avaliação de Explicabilidade

A métrica *faithfulness* é amplamente utilizada no campo da Inteligência Artificial Explicável para avaliar o grau em que uma explicação gerada reflete com precisão o funcionamento interno do modelo que tomou a decisão. Em termos simples,

uma explicação é considerada fiel se as razões apresentadas para uma previsão específica correspondem diretamente às operações e cálculos realizados pelo modelo. Essa métrica é especialmente relevante em modelos complexos de "caixa-preta", como redes neurais profundas e florestas aleatórias, que, embora altamente eficazes, são difíceis de interpretar diretamente devido à sua complexidade.

O principal objetivo da métrica faithfulness é verificar se a explicação fornecida realmente representa as características que o modelo utilizou para chegar a uma determinada previsão, e não apenas fornecer uma narrativa plausível para os humanos. Isso é feito analisando se a importância atribuída a cada variável na explicação é consistente com o impacto real que essa variável tem sobre a saída do modelo. Por exemplo, se uma explicação aponta que o nível de glicose no sangue é um fator importante para a previsão de um diagnóstico de diabetes, o cálculo da métrica de fidelidade verificará se a remoção ou alteração do valor dessa variável afeta significativamente a previsão do modelo. Caso contrário, a explicação não seria fiel, indicando que o fator destacado na explicação não reflete de fato o comportamento interno do modelo.

O cálculo da fidelidade geralmente envolve a perturbação das variáveis de entrada para medir o impacto que essas alterações têm nas previsões do modelo. Métodos amplamente utilizados, como LIME e SHAP, seguem abordagens distintas para avaliar essa relação. No LIME, essas perturbações envolvem alterações controladas nas variáveis, como remover palavras em classificadores de texto ou ajustar valores em dados tabulares. A fidelidade é avaliada observando como as previsões mudam em resposta às alterações feitas. Por exemplo, se o LIME indica que uma variável é importante, a remoção ou modificação dessa variável deveria alterar significativamente a previsão do modelo. Caso contrário, a explicação seria considerada infiel. Já o SHAP utiliza a teoria dos jogos para calcular o impacto marginal de cada variável na previsão final. Ele distribui a importância de forma consistente, garantindo que as explicações sejam fiéis ao comportamento do modelo. A fidelidade, nesse caso, é verificada comparando a contribuição atribuída pelo SHAP a cada variável com o efeito real dessas variáveis nas saídas do modelo, considerando todas as possíveis combinações de variáveis.

Nesse estudo, o cálculo da fidelidade foi realizado da seguinte maneira:

1. **Inicialização:**
    - A função recebe como entrada os dados de teste (X_test, y_test), o modelo (model), o explicador (explainer) e o número de variáveis a serem consideradas (num_features).
    - Uma lista vazia (faithfulness_scores) é criada para armazenar as pontuações de fidelidade de cada instância.
2. **Iteração sobre as instâncias de teste:**
    - Para cada instância no conjunto de teste (X_test):

    - A previsão real do modelo para a instância é obtida usando (model.predict()).
    - Uma explicação para a instância é gerada pelo LIME usando (explainer.explain_instance()).
    - Uma variável (perturbation_faithfulness) é inicializada com 0 para rastrear a fidelidade para a instância atual.
3. **Perturbações e Avaliação da Fidelidade:**
    - Para um número fixo de perturbações (num_perturbations):
    - Uma cópia da instância é criada (perturbed_instance).
    - Uma variável aleatória é selecionada para ser perturbada.
    - Um valor aleatório é atribuído à variável perturbada.
    - A previsão do modelo para a instância perturbada é obtida.
    - Uma explicação para a instância perturbada é gerada pelo LIME.
    - Se a previsão real e a previsão perturbada forem iguais, a fidelidade é avaliada:
    - Se o sinal do peso da variável mais importante na explicação original e na explicação perturbada for o mesmo, (perturbation_faithfulness) é incrementado em 1, indicando que a explicação é consistente com a perturbação.
4. **Cálculo da Pontuação de Fidelidade:**
    - Após todas as perturbações, (perturbation_faithfulness) é dividido pelo número total de perturbações para obter a pontuação de fidelidade para a instância atual.
    - Essa pontuação é adicionada à lista (faithfulness_scores).
5. **Retorno da Fidelidade Média:**
    - A função retorna a média das pontuações de fidelidade de todas as instâncias.

A métrica de *faithfulness* é crucial para aumentar a confiança nas decisões dos modelos de IA, garantindo que as explicações fornecidas realmente correspondam ao processo de decisão do modelo. Ela é avaliada, principalmente, através da análise de perturbações nas variáveis de entrada ou pela comparação entre as importâncias atribuídas às variáveis pela explicação e pelo modelo [Alvarez-Melis, D., e Jaakkola, T. S. 2018].

2.3. Leis que regulamentam a Inteligência Artificial

Algumas limitações da inteligência artificial são abordadas no estudo da Scientific Foresight Unit (STOA) de 2020, relatando os impactos da Regulamentação Geral de Proteção de Dados (GDPR) [10] na inteligência artificial. O documento explora como o GDPR molda o desenvolvimento e a aplicação de IA na União Europeia. A regulamentação, centrada na proteção dos dados pessoais dos cidadãos, impõe desafios e oportunidades para o uso de IA, considerando que esta tecnologia depende significativamente do processamento de grandes volumes de dados.

Principais pontos abordados:

1. Desafios legais e éticos: O GDPR enfatiza a proteção da privacidade e impõe limites ao uso de dados pessoais, o que cria barreiras para IA, que frequentemente requer grandes conjuntos de dados para treinar algoritmos. Aspectos como o direito ao esquecimento, o consentimento explícito e a transparência no uso de dados desafiam a implementação de soluções de IA que operam com dados pessoais.
2. Transparência e *accountability*: O GDPR introduz o conceito de *accountability*, o que significa que as organizações que desenvolvem ou utilizam IA precisam demonstrar que seus sistemas cumprem com as regras de proteção de dados. A exigência de transparência também é destacada, o que implica que os modelos de IA precisam ser explicáveis para os usuários, especialmente quando decisões automatizadas são tomadas.
3. Automação e decisões algorítmicas: O GDPR regula o uso de decisões automatizadas baseadas em dados pessoais, permitindo que os cidadãos tenham o direito de contestar essas decisões e exigir intervenção humana em alguns casos. Isso limita o uso indiscriminado de algoritmos de IA em processos de tomada de decisão que afetam os indivíduos.

A Lei Geral de Proteção de Dados (LGPD), promulgada no Brasil em 2018, compartilha muitos princípios com o GDPR, sendo uma legislação fortemente inspirada no regulamento europeu. Tanto a LGPD quanto o GDPR exigem que os dados pessoais sejam processados para fins específicos e que o consentimento do titular dos dados seja obtido de forma clara e explícita. No caso de sistemas de IA, isso significa que as empresas devem informar de maneira clara como os dados serão utilizados, e os indivíduos devem ter controle sobre suas informações. A LGPD também segue o princípio da responsabilização (*accountability*), exigindo que as empresas demonstrem o cumprimento da lei. Além disso, a transparência no tratamento de dados é um requisito central, o que implica que, para IA, os algoritmos devem ser explicáveis e não atuar de forma discriminatória ou arbitrária.

Ambos os regulamentos têm um impacto significativo sobre o uso de inteligência artificial em seus respectivos territórios, exigindo que as tecnologias sejam desenvolvidas com foco na proteção da privacidade, transparência e respeito aos direitos dos titulares de dados.

3. Materiais e Métodos

Tendo em vista o objetivo deste trabalho que é de propor uma ferramenta para avaliar a qualidade das explicações geradas pelo algoritmo LIME de explicabilidade

implementado para modelos de aprendizagem de máquina, mais especificamente o *Random Forest*. Foi implementado um sistema em que o método de explicabilidade citado é aplicado junto à uma implementação da métrica *faithfulness* para avaliação da fidelidade desta explicação.

3.1. Proposta de Ferramenta para Aplicação do Método LIME de Explicabilidade

A ferramenta apresentada neste trabalho é uma interface gráfica interativa que visa facilitar a explicação de modelos de aprendizado de máquina utilizando a técnica LIME.

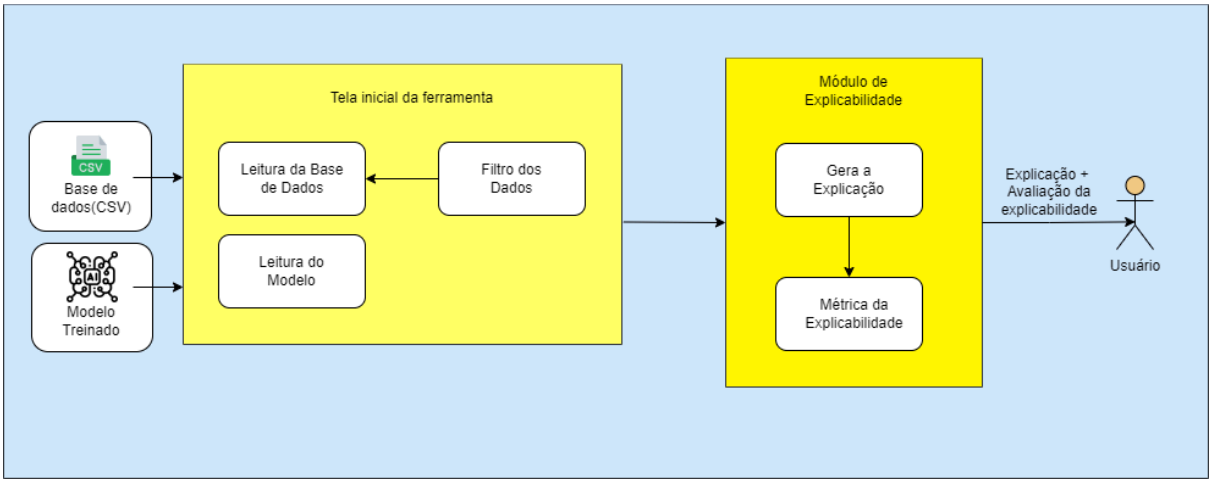
A interface proposta permite que os usuários carreguem modelos previamente treinados e as bases de dados associadas para análise e explicação. Com um design intuitivo e focado em etapas sequenciais, a ferramenta guia o usuário desde o carregamento dos dados até a geração de explicações usando o LIME, sem a necessidade de um profundo conhecimento técnico. A interface suporta tanto dados tabulares quanto dados de texto, possibilitando sua utilização em cenários como classificação de texto e análise de dados estruturados.

3.2. Pipeline da Ferramenta XAI LIME

O pipeline da ferramenta proposta é composto por duas fases principais: a fase de preparação dos dados e do modelo, realizada na tela inicial da ferramenta; e a fase de explicabilidade, que envolve a geração e a avaliação das explicações utilizando a técnica LIME e a métrica *faithfulness*.

A seguir podemos ver a figura de como funciona o fluxo do pipeline da ferramenta:

Figura 2 - Desenho ilustrativo do Pipeline da ferramenta



Fonte: Elaboração própria feita no [Draw.io]

O processo começa no módulo de leitura com o upload de dois arquivos por parte do usuário. O primeiro arquivo corresponde à base de dados (formato CSV), que contém os dados utilizados no treinamento do modelo de aprendizagem de máquina. O segundo arquivo é o modelo treinado, que pode estar em formatos como `.h5`, `.pkl`, `.model`, ou `.sav`. Após o upload, a ferramenta procede com a leitura desses arquivos.

Na tela inicial da ferramenta, duas ações principais são realizadas: a leitura da base de dados e a leitura do modelo. A ferramenta processa a base de dados para entender sua estrutura e as variáveis presentes, enquanto o modelo é carregado e validado para garantir que está em um formato compatível e apto para ser analisado. Além disso, é oferecida ao usuário a possibilidade de realizar um filtro dos dados carregados, permitindo que ele selecione subconjuntos específicos de informações que deseja focar na explicação.

Com o modelo e os dados preparados, o pipeline avança para o módulo de explicabilidade, onde são realizadas duas operações fundamentais. Primeiro, o sistema aplica o algoritmo LIME para gerar explicações locais sobre as previsões feitas pelo modelo. Essas explicações são apresentadas na forma de gráficos ou representações visuais que detalham como as variáveis de entrada influenciam as previsões realizadas pelo modelo. Em seguida, a ferramenta utiliza a métrica faithfulness para avaliar a qualidade dessas explicações, verificando se elas realmente refletem o comportamento real do modelo, assegurando que as informações fornecidas são confiáveis.

Por fim, o sistema exibe as explicações geradas pelo LIME juntamente com a avaliação de sua qualidade por meio da métrica faithfulness, possibilitando ao usuário visualizar as explicações de forma clara e verificar o grau de confiabilidade das mesmas em relação ao comportamento real do modelo. Esse pipeline é projetado para ser

intuitivo, flexível e eficiente, proporcionando uma análise detalhada e fundamentada da explicabilidade dos modelos de aprendizagem de máquina.

3.3. Especificação da Ferramenta, Modelagem e Fluxo de Atividades

Para validar a ferramenta proposta, foi desenvolvido um estudo de caso que simula um cenário de interação de um usuário com a interface do sistema, cujo objetivo é carregar um modelo de aprendizagem de máquina treinado, gerar explicações por meio do algoritmo LIME, e avaliar a qualidade dessas explicações utilizando a métrica faithfulness. A fim de representar e organizar esse fluxo de interação, foram utilizados alguns recursos da UML (Unified Modeling Language) para ilustrar o comportamento do sistema de maneira clara e objetiva, focando nos aspectos mais relevantes para o estudo.

A opção escolhida foi utilizar apenas o Fluxo de Atividades, pois é o suficiente para descrever as principais interações e fluxos do sistema, sem adicionar complexidade desnecessária à modelagem. Esse elemento fornece uma visão clara e objetiva do comportamento da ferramenta, permitindo que o foco do estudo permaneça na análise da explicabilidade gerada pelo algoritmo LIME e na avaliação de sua qualidade pela métrica faithfulness.

Essa abordagem facilita a compreensão do funcionamento da ferramenta, mantendo a clareza necessária para explicar o processo de interação do usuário com a ferramenta, sem sobrecarregar o estudo com detalhes excessivos de modelagem.

Ator Principal:

- Usuário: Pessoa que deseja carregar e obter explicações sobre um modelo de aprendizagem de máquina treinado.

Objetivo:

- Permitir que o usuário faça upload de uma base de dados e um modelo de aprendizagem de máquina treinado, visualize explicações baseadas na técnica LIME e avalie a qualidade dessa explicação utilizando a métrica *faithfulness*.

Fluxo Principal:

1. Usuário acessa a interface da ferramenta.
   - O usuário abre o sistema e vê a opção de upload de arquivos na interface.

2. Usuário faz o upload de um modelo treinado.

- O usuário clica no botão "Escolher Arquivo", seleciona o arquivo do modelo treinado (formatos aceitos: ‘.h5’, ‘.pkl’, ‘.model’, ‘.sav’), e o nome do arquivo é exibido na interface.

- Pré-condição: O usuário deve ter um modelo treinado válido no formato aceito.

3. Usuário envia o arquivo.

- O usuário clica no botão "Enviar Modelo", e o sistema faz o upload do modelo para o servidor.

4. Sistema valida o arquivo.

- O sistema verifica se o arquivo é um modelo treinado válido e carrega o modelo para ser analisado.

- Regra de Negócio: O arquivo deve ser um modelo treinado em um dos formatos aceitos.

5. Usuário faz upload da Base de dados.

- O usuário faz o upload do arquivo contendo a base de dados utilizada.

6. Sistema exibe pergunta de tipo de dados.

- O sistema exibe a pergunta: “Que tipo de dados contém essa base?”, sendo uma única resposta dentre duas possíveis, sendo elas: “Dados tabulares”, “Dados textuais”.

7. Usuário escolhe uma opção.

- O usuário deve responder a pergunta apresentada escolhendo uma das opções apresentadas, seja a resposta “Dados tabulares” ou “Dados textuais”.

8. Sistema exibe o botão "Explicar com LIME".

- Após o carregamento bem-sucedido do modelo, o sistema apresenta o botão "Explicar com LIME", que permite ao usuário gerar explicações para as previsões do modelo.

9. Usuário clica em "Explicar com LIME".

- O usuário seleciona a opção de explicar o modelo usando a técnica LIME.

10. Sistema gera a explicação do modelo usando LIME.

- O sistema aplica a técnica LIME para gerar uma explicação visual (gráfico) que descreve como as variáveis de entrada influenciam as previsões do modelo.

- Regra de Negócio: O sistema deve ser capaz de executar a técnica LIME no modelo treinado.

11. Sistema exibe o gráfico explicativo.

- Um gráfico gerado pelo LIME é exibido na interface, mostrando a contribuição de cada variável nas previsões.

12. Sistema apresenta avaliação da explicação (faithfulness).

- O sistema também calcula e exibe a métrica faithfulness, que avalia a qualidade da explicação baseada em quanto as características explicadas afetam diretamente as previsões do modelo.

- A métrica é apresentada ao lado do gráfico, indicando se a explicação é fiel ao comportamento real do modelo.

Fluxo Alternativo:

1A. Falha no Upload do Arquivo:

- Se o arquivo enviado não for um modelo válido ou estiver corrompido, o sistema exibe uma mensagem de erro ("Modelo inválido") e solicita um novo arquivo.

- O usuário retorna ao passo 2.

Pós-condição:

- O usuário consegue ver a explicação gerada pela técnica LIME e avaliar sua fidelidade usando a métrica faithfulness.

Diagrama de Caso de Uso (especificações em texto):

1. Usuário → [Carrega modelo treinado]
2. Sistema ← Valida o modelo carregado
3. Usuário → [Carrega a base de dados]
4. Usuário → [Responde a pergunta do tipo de dados utilizado]
5. Usuário → Clica em [Explicar com LIME]

6. Sistema ← Gera gráfico explicativo usando LIME

7. Sistema ← Avalia a explicação usando a métrica faithfulness

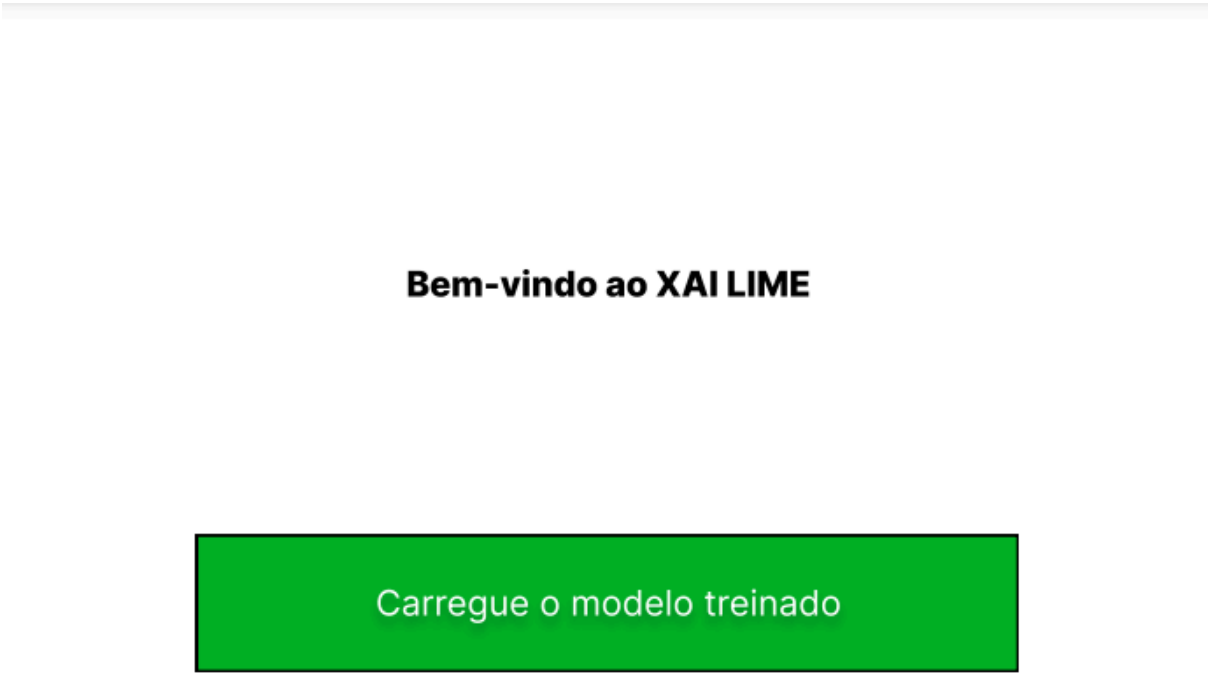
8. Usuário → Visualiza gráfico e avaliação de explicação

3.4. Prototipagem da Ferramenta

Para uma compreensão mais realista de como a ferramenta proposta poderia atuar e ser desenvolvida, foi feito uma prototipagem das telas através do Figma [https://www.figma.com/], onde foi possível demonstrar as funcionalidades da ferramenta e também ilustrar o passo a passo de seu uso além do resultado esperado.

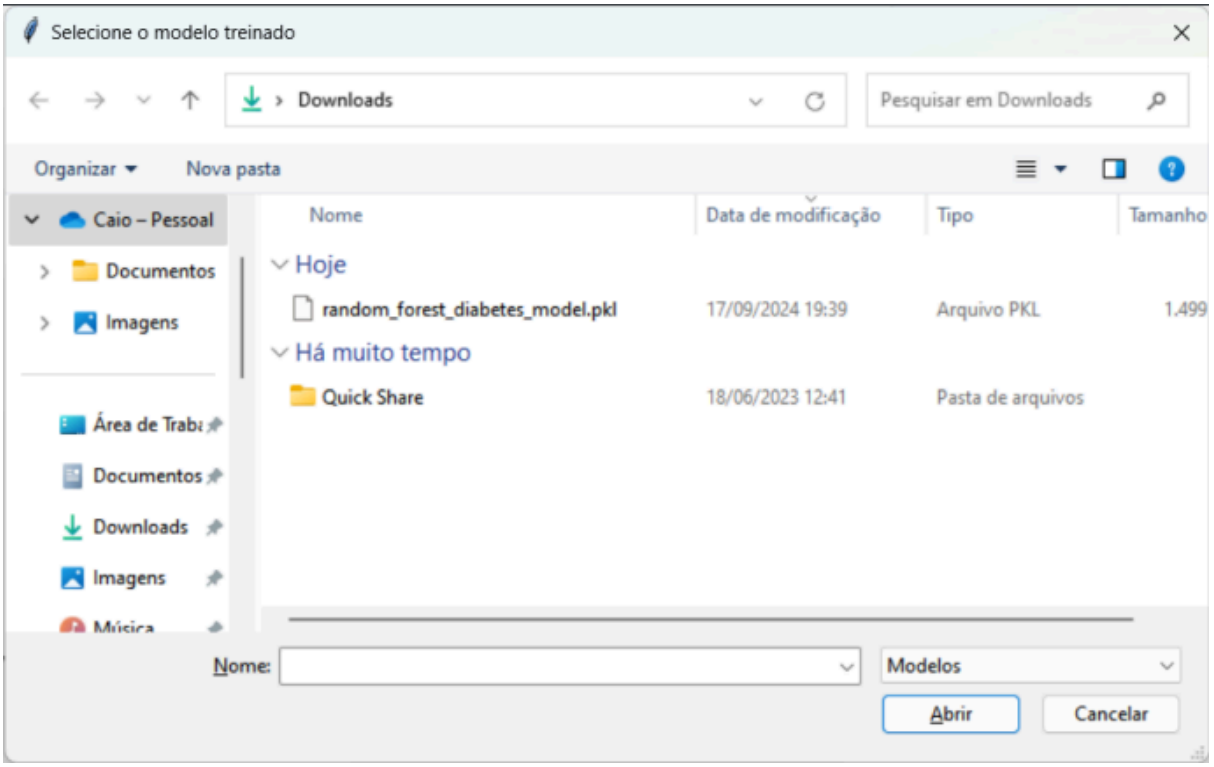
Na figura 3 temos a tela inicial apresentada ao acessar a ferramenta.

Figura 3 - Tela inicial da ferramenta



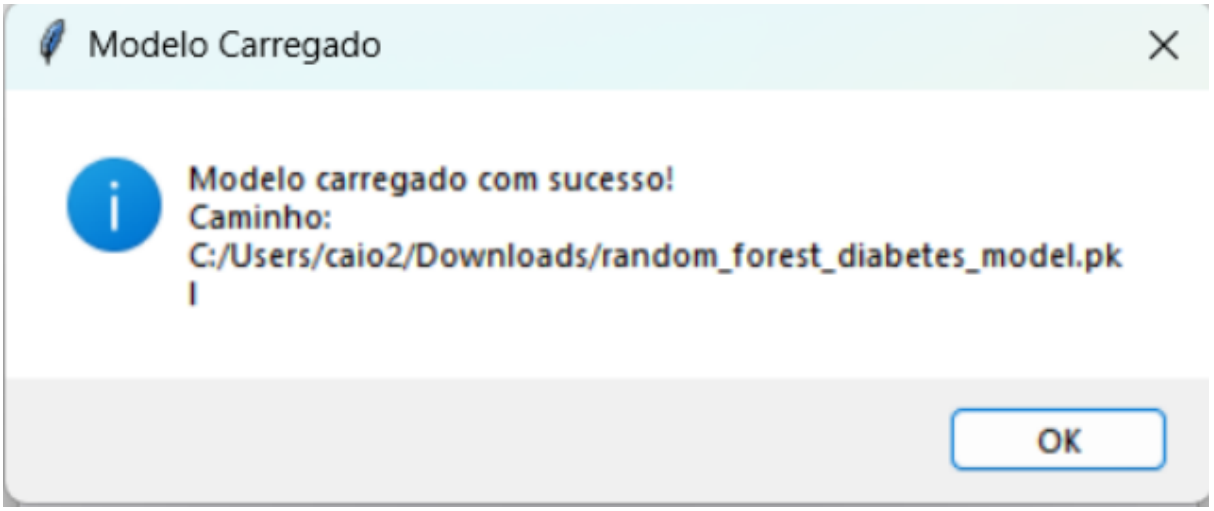
A figura 4 mostra a tela de seleção do modelo de aprendizagem de máquina já treinado.

Figura 4 - Tela de seleção do modelo treinado



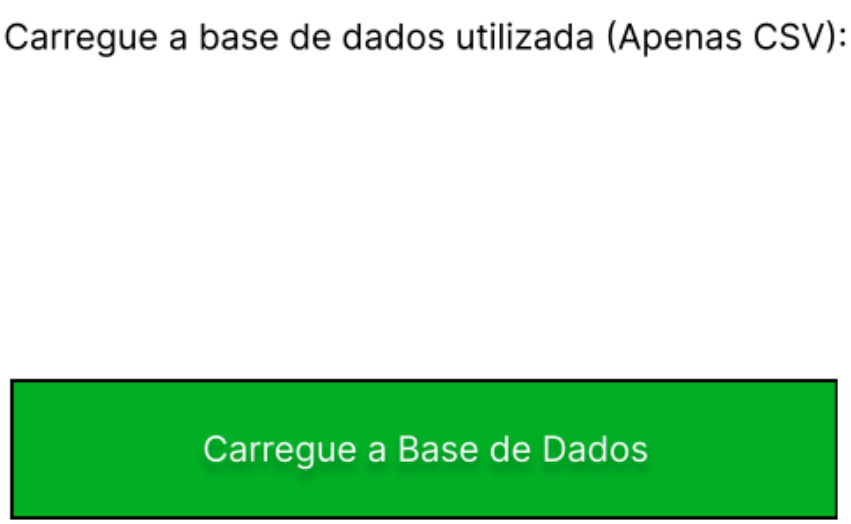
A figura 5 apresenta uma mensagem ao carregar o modelo treinado com sucesso.

Figura 5 - Tela com mensagem de sucesso



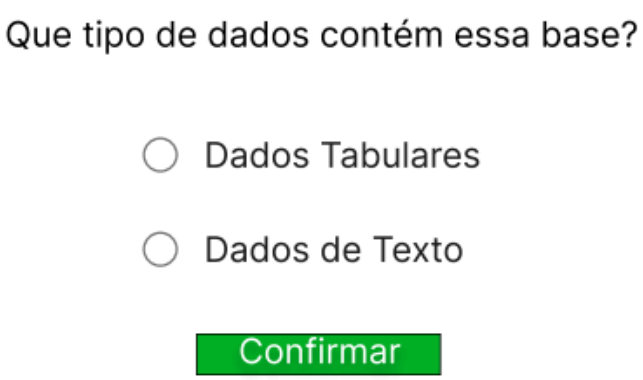
A figura 6 mostra a tela em que o usuário irá carregar a base de dados.

Figura 6 - Tela para carregamento da base de dados



A figura 7 mostra a tela onde o usuário seleciona que tipo de dados a base usa.

Figura 7 - Tela de seleção de tipo de dados usados

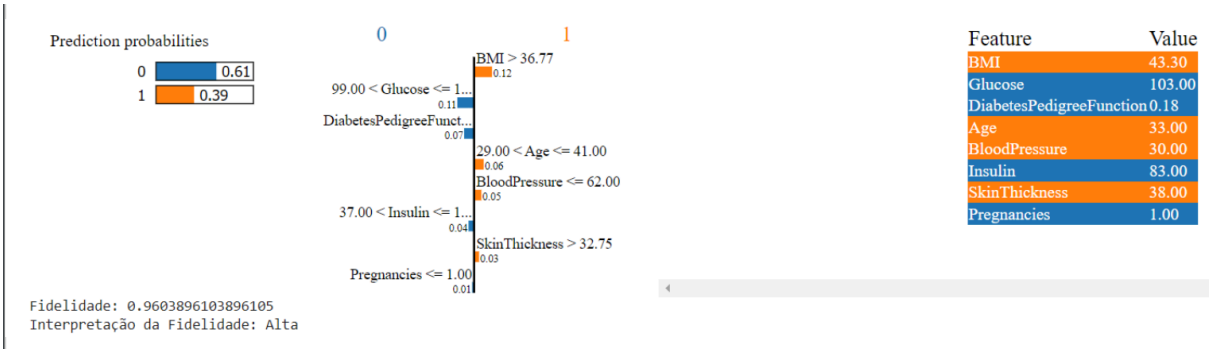


A figura 8 mostra a tela em que o usuário vai gerar a explicação com o LIME.

Figura 8 - Tela para geração da explicação com o LIME

A figura 9 mostra a explicação gerada para a predição de uma instância escolhida aleatoriamente pela ferramenta. Acompanha também a probabilidade das classes previstas além da avaliação da confiabilidade da explicação.

Figura 9 - Tela com a explicação e avaliação geradas



3.5. Implementação da Ferramenta

A interface da ferramenta foi desenvolvida em Python [Downey 2016] utilizando bibliotecas como “sklearn”, “matplotlib”, “lime” e a biblioteca “Tkinter”, com foco na simplicidade e na interação do usuário, permitindo que modelos previamente treinados e bases de dados sejam carregados e analisados. Após a definição do tipo de dados (tabulares ou texto), a ferramenta facilita a geração de explicações utilizando o LIME, tornando o processo mais acessível e compreensível para pesquisadores e usuários não especialistas.

3.6. Base de Dados e Modelo de Aprendizagem de Máquina Usados

Na fase de desenvolvimento e validação da ferramenta de explicabilidade, utilizamos a base de dados Pima Indians Diabetes Database, disponível no [Kaggle](https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database). Essa base contém informações de exames médicos de um grupo de mulheres da etnia Pima, com idades a partir de 21 anos, e é amplamente utilizada para prever a ocorrência de diabetes. A base é composta por 768 instâncias e 9 atributos, sendo 8 variáveis preditoras e 1 variável de classe binária que indica a presença ou ausência de diabetes.

Os atributos preditores incluem:

- Pregnancies (Número de gravidezes)
- Glucose (Nível de glicose no sangue)
- BloodPressure (Pressão arterial diastólica)
- SkinThickness (Espessura da dobra cutânea)
- Insulin (Insulina)
- BMI (Índice de Massa Corporal (IMC))
- DiabetesPedigreeFunction (Função hereditária de diabetes)
- Age (Idade)

Para realizar os testes com a ferramenta de explicabilidade, foi utilizado o modelo de Random Forest [Breiman, L. 2001], uma técnica de aprendizagem de máquina que se caracteriza por ser um modelo "caixa-preta".

Após o treinamento do modelo de Random Forest utilizando a base de dados, a ferramenta foi aplicada para gerar explicações locais das previsões por meio do LIME.

O algoritmo permitiu identificar as contribuições de cada variável nas previsões individuais, oferecendo uma visão mais clara e interpretável sobre como o modelo chegou a cada decisão. Por exemplo, foi possível visualizar como o nível de glicose ou o índice de massa corporal influenciaram a decisão do modelo sobre o diagnóstico de diabetes.

Além da geração das explicações, a ferramenta também avaliou a qualidade das explicações geradas utilizando a métrica *faithfulness* [Lundberg e Lee 2017]. Essa métrica verificou o quanto as variáveis destacadas nas explicações realmente refletiam o comportamento real do modelo, garantindo que as explicações fossem confiáveis e fiéis ao funcionamento interno do Random Forest.

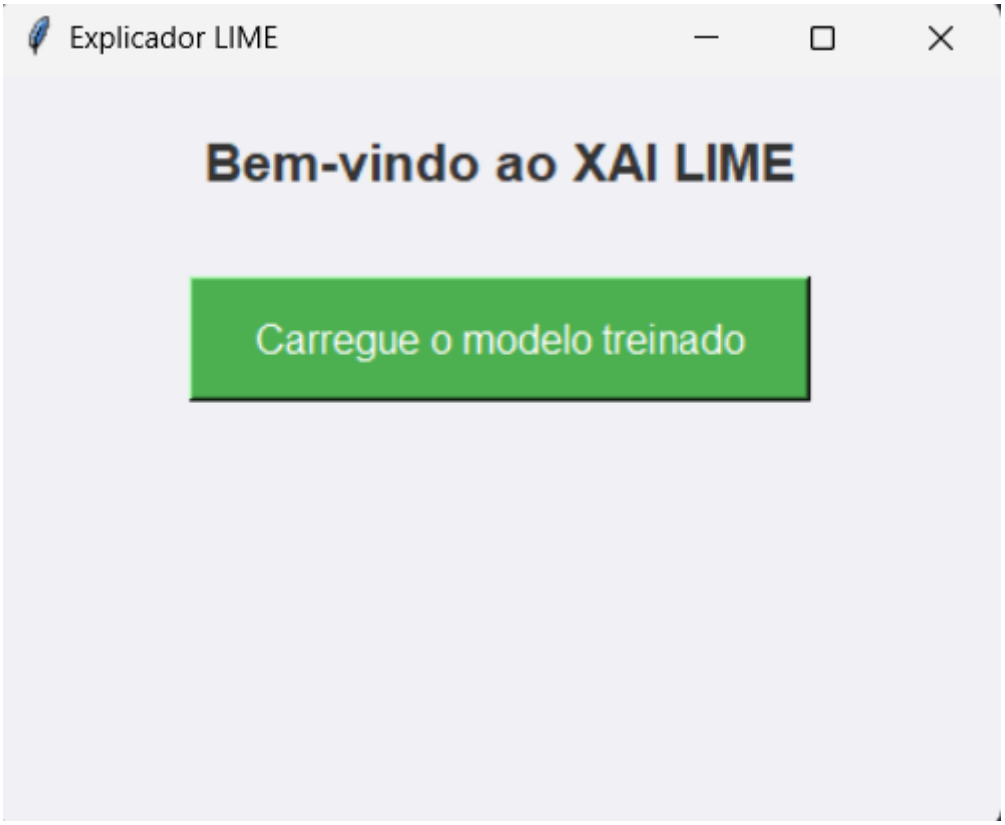
A utilização dessa base de dados e do modelo Random Forest foi fundamental para testar a ferramenta em um contexto de predição de doenças, onde a interpretabilidade dos resultados é crucial para a confiança nos modelos de IA. A capacidade de explicar e avaliar a fidelidade das previsões ajuda a aumentar a transparência e a confiabilidade das aplicações de aprendizagem de máquina na área da saúde.

4. Resultados

Como resultado da pesquisa, estudo e implementação, temos uma proposta de ferramenta apresentada a seguir, que tem por finalidade fornecer explicações baseadas na base de dados apresentada e no modelo de aprendizagem de máquina utilizado, por fim avaliando a qualidade da própria explicação.

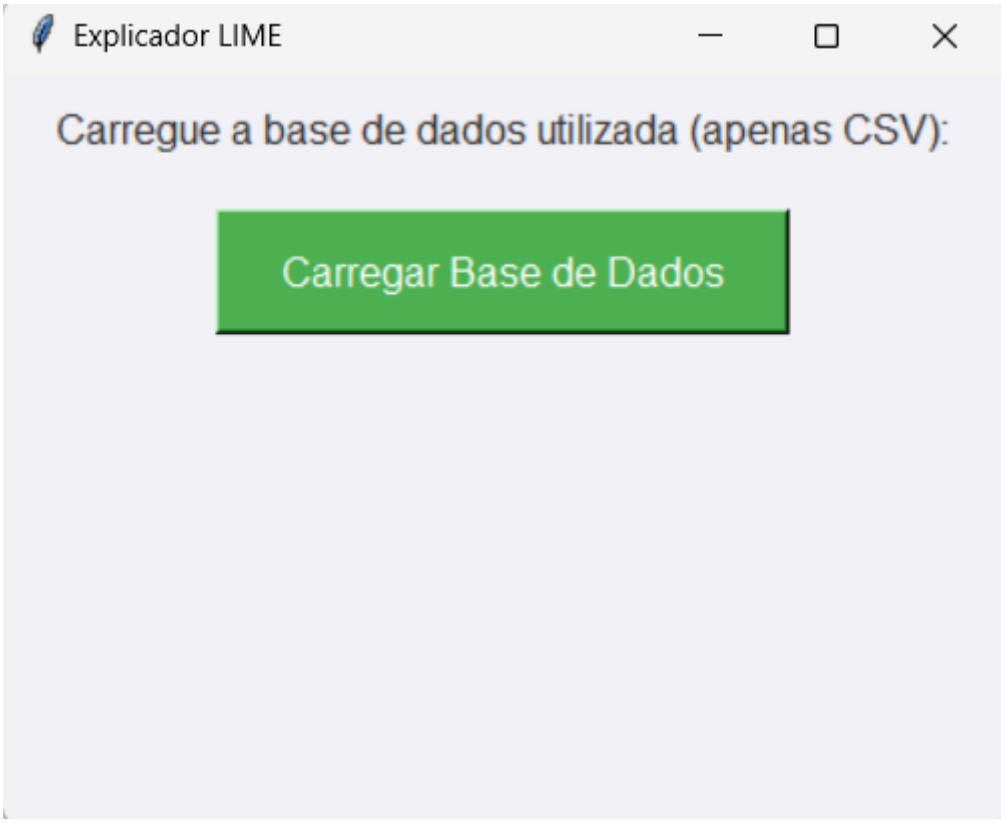
A tela inicial apresentada na figura 10 é uma interface com instruções para o usuário inserir o modelo de aprendizagem de máquina treinado.

Figura 10 - Tela inicial da ferramenta



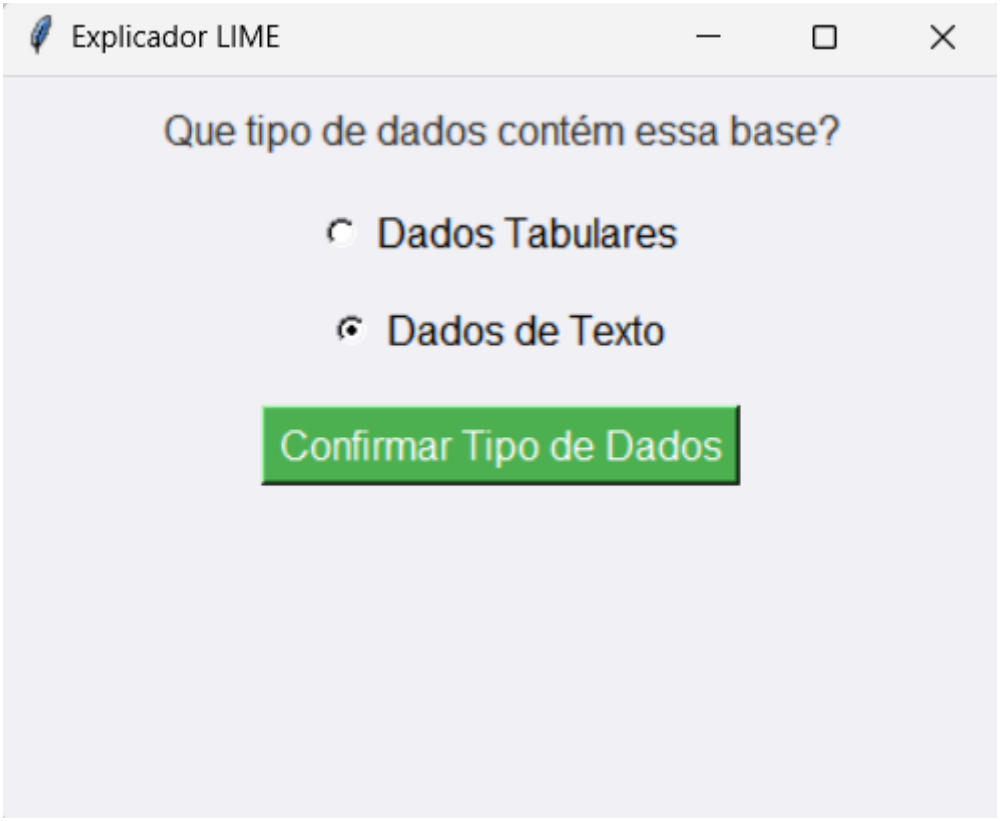
A segunda tela apresentada na figura 11 é uma interface com instruções para o usuário carregar a base de dados utilizada.

Figura 11 - Tela para carregamento da base de dados



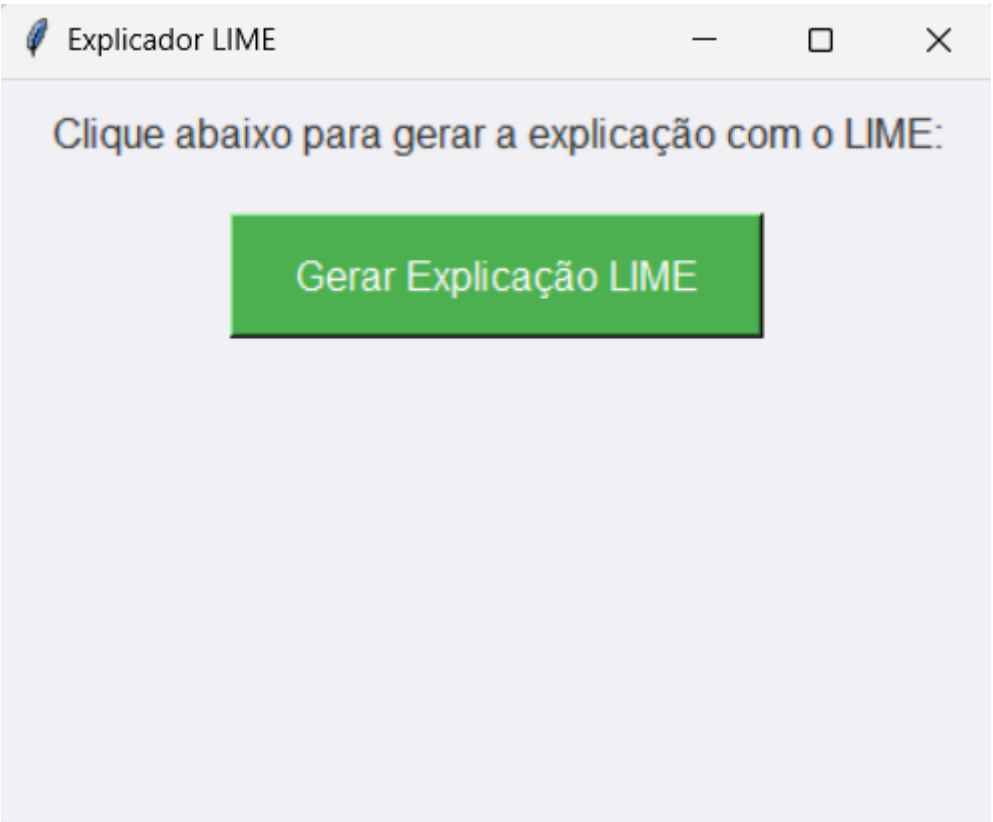
A terceira tela apresentada na figura 12 é uma interface com opções para o usuário escolher qual tipo de dado (Tabulares ou de Texto) está presente na base de dados.

Figura 12 - Tela de seleção de tipo de dados usados



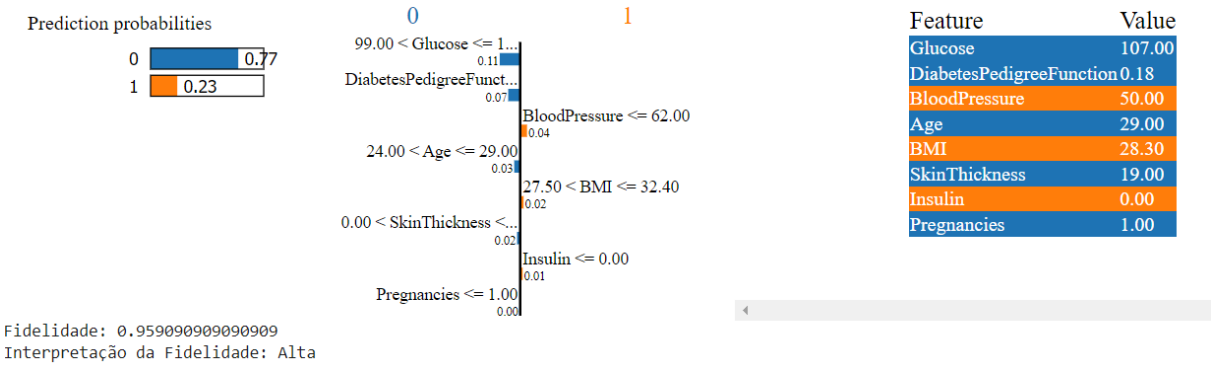
A quarta tela apresentada na figura 13 mostra a opção para que possa ser gerado a explicação com o LIME.

Figura 13 - Tela para geração da explicação com o LIME



A quinta tela ilustrada na figura 14 já apresenta a explicação gráfica além da avaliação da explicação.

Figura 14 - Tela com a explicação e avaliação geradas



O resultado da explicação gerada com o LIME acima é apresentada de forma que as probabilidades da previsão aparecem no lado esquerdo, sendo a probabilidade desta paciente de não ter diabetes é 77% e a de ter diabetes é de 23%. Ao centro vemos

os parâmetros utilizados para definir se a variável está influenciando favorável à diabetes ou contra. Mais a direita podemos observar os valores de cada variável atribuída a esta paciente.

5. Trabalhos Relacionados

5.1 Artigo “Explainable AI in Cybersecurity Operations: Lessons Learned from xAI Tool Deployment - Nyre-Yu, M., Morris, E., Smith, M. R., Moss, B., & Smutz, C. (2022)”

O artigo "Explainable AI in Cybersecurity Operations: Lessons Learned from xAI Tool Deployment" aborda a aplicação de ferramentas de IA explicável (xAI) nas operações de segurança cibernética, especialmente para ajudar analistas a identificar e justificar atividades suspeitas na rede. Em um estudo de caso, os autores analisaram como uma ferramenta xAI, usando o método TreeSHAP, poderia influenciar a confiança e a eficiência dos analistas nas tomadas de decisão. No entanto, os resultados mostraram que os analistas usaram pouco a ferramenta, mesmo após receberem treinamentos, o que indica que a explicabilidade não melhorou a precisão das decisões nem aumentou a confiança nas previsões dos modelos de IA. O estudo destacou a importância de alinhar o design da ferramenta xAI ao fluxo de trabalho específico dos analistas e recomendou métodos de coleta de dados que não interfiram nas tarefas diárias para facilitar a adoção dessa tecnologia.

5.2 Artigo “XAITK: The explainable AI toolkit - Hu, B., Tunison, P., Vasu, B., Menon, N., Collins, R., & Hoogs, A. (2021)”

Este trabalho introduz o XAITK, uma plataforma criada para reunir e compartilhar os avanços em IA explicável desenvolvidos pelo programa DARPA XAI. O objetivo é tornar os modelos de IA mais transparentes e fáceis de entender, consolidando algoritmos, frameworks e outros recursos em um repositório único e acessível. O XAITK facilita a aplicação dessas tecnologias em organizações interessadas, incluindo não só componentes de software voltados para análise e automação, mas também publicações, dados e orientações práticas. Dessa forma, o kit de ferramenta apoia o uso ético e confiável da IA em áreas como saúde e segurança, atendendo tanto pesquisadores quanto profissionais operacionais que dependem de uma IA confiável e transparente.

5.3 Artigo “Can Explainable AI Explain Unfairness? A Framework for Evaluating Explainable AI. Alikhademi, K., Richardson, B., Drobina, E., & Gilbert, J. E. (2022).”

Neste artigo, os autores exploram a importância de ferramentas de IA explicável (XAI) que não apenas tornam o comportamento dos modelos de aprendizado de máquina mais compreensível, mas também auxiliam na detecção e mitigação de vieses e injustiças. Eles propõem um framework de avaliação para essas ferramentas, destacando

como tecnologias XAI, em alguns casos, podem esconder vieses ao gerar explicações que induzem os usuários a confiar em modelos potencialmente injustos, um fenômeno conhecido como "fairwashing". O estudo examina ferramentas populares como LIME e AI Explainability 360, apontando falhas em suas funcionalidades para detectar vieses. Assim, os autores apresentam uma rubrica prática para orientar desenvolvedores a aprimorar suas ferramentas, tornando-as mais transparentes e confiáveis, com foco na justiça e na eficácia dos modelos de IA para os usuários.

5.4 Artigo “Uma Arquitetura de Referência para Explicabilidade como Serviço na Saúde H-XAIaaS: Health - eXplainable Artificial Intelligence as a Service - Thiago Montenegro (2024)”

Um trabalho relacionado relevante para esta pesquisa é o de Thiago Montenegro, que propõe a arquitetura H-XAIaaS (Health eXplainable Artificial Intelligence as a Service). Esse estudo apresenta uma solução voltada para o setor da saúde, com o objetivo de integrar explicabilidade em modelos de aprendizado de máquina aplicados a cenários clínicos. O principal foco da arquitetura é fornecer maior transparência e interpretabilidade nas decisões tomadas pelos modelos de IA, o que é essencial para aumentar a confiança no uso dessas tecnologias em ambientes críticos, como diagnósticos médicos.

A dissertação explora dois estudos de caso que validam a abordagem H-XAIaaS. O primeiro estudo utiliza dados tabulares, e o segundo trabalha com dados de imagens médicas, demonstrando a flexibilidade da arquitetura em lidar com diferentes tipos de dados. Os resultados mostram como a explicabilidade e a transparência podem ser aplicadas de forma eficaz em contextos variados de saúde, contribuindo para aumentar a confiança e a segurança no uso de modelos de IA na prática clínica.

A principal diferença entre os três trabalhos reside no domínio de aplicação e na forma como a explicabilidade é implementada e integrada ao fluxo de trabalho. Enquanto o XAITK se concentra em fornecer uma base geral para a XAI, "Explainable AI in Cybersecurity Operations" e "H-XAIaaS" focam em contextos específicos (segurança cibernética e saúde, respectivamente), abordando os desafios da adaptação da explicabilidade ao ambiente de uso e ao usuário final. Em relação ao seu trabalho, a aplicação direta ao contexto de segurança cibernética coloca "Explainable AI in Cybersecurity Operations" como a pesquisa mais relevante, enquanto o XAITK e o H-XAIaaS trazem insights sobre estruturas modulares e adaptabilidade que podem contribuir para a escalabilidade e aplicação prática em ambientes complexos. Já o “Can Explainable AI Explain Unfairness? A Framework for Evaluating Explainable AI.” foca em analisar e propor uma avaliação sobre a explicabilidade, o que se assemelha ao presente trabalho apresentado neste artigo.

A seguir vemos a tabela 1 que compara os trabalhos relacionados com o deste artigo.

Tabela 1 - Tabela comparativa de trabalhos relacionados

	Algoritmos XAI Usados	Métricas de XAI usadas na ferramenta	Domínio	Implementação
Art 1	TreeSHAP, Saliency Maps e explicabilidade em aprendizado por reforço	N/A	Cibersegurança	N/A
Art 2	Condensa diversas técnicas	N/A	Geral	N/A
Art 3	LIME e AI Explainability 360	Rubrica de avaliação para ferramentas de IA explicável	Avaliação de Explicabilidade	N/A
Art 4	CAM, GRAD-CAM, LIME e SHAP	Recall, Precisão, Acurácia, Cohen	Saúde	N/A
Este Trabalho	LIME	Faithfulness	Geral	Phyton

6. Conclusão

6.1. Contribuições

O presente trabalho atingiu o objetivo geral proposto, que era desenvolver uma ferramenta capaz de avaliar a confiabilidade das explicações geradas pelo algoritmo LIME em modelos de aprendizado de máquina, utilizando a métrica de *faithfulness*, porém a parte de integração dos módulos de leitura e de explicabilidade foi parcialmente desenvolvida, uma vez que a ferramenta não chegou na proposição final imaginada e o desenvolvimento estagnou no ponto de integração da interface com os resultados apresentados pela explicação. Ao longo do estudo, foram realizados testes em modelos caixa-preta, aplicando o *Random Forest*, utilizando a base de dados *Pima Indians Diabetes Database*, o que possibilitou verificar a aplicabilidade e eficiência da

ferramenta em contextos de predição de um possível diagnóstico de diabetes. Além disso, o trabalho atendeu aos objetivos específicos ao selecionar e aplicar modelos de aprendizado de máquina, gerar explicações locais com LIME e avaliar a fidelidade dessas explicações por meio da métrica *faithfulness*. Como resultado, foi possível fornecer aos usuários previsões mais interpretáveis e confiáveis, reforçando a transparência e a confiança nas decisões algorítmicas.

6.2. Limitações

Durante o desenvolvimento deste trabalho, percebe-se algumas limitações na ferramenta proposta. Uma delas é a necessidade de adaptar explicações que sejam claras e compreensíveis para diferentes públicos, o que pode ser complicado quando lidamos com a complexidade de modelos do tipo "caixa-preta". Além disso, é importante garantir que esses modelos estejam em sintonia com o contexto e as práticas de trabalho de quem vai utilizá-los. Outro desafio é que algumas técnicas de explicabilidade não funcionam tão bem quando aplicadas a grandes volumes de dados ou a modelos mais complexos, o que torna difícil gerar explicações que sejam claras e fáceis de entender em larga escala. Por fim, oferecer explicações detalhadas muitas vezes exige acesso a dados sensíveis, o que levanta preocupações quanto à privacidade e à segurança dessas informações.

6.3. Trabalhos Futuros

Como proposta para trabalhos futuros, sugere-se a ampliação da ferramenta para incorporar outros métodos de explicabilidade além do LIME, como o SHAP [Lundberg e Lee 2017] e *Anchors* [Ribeiro, Singh e Guestrin 2018], visando aumentar a flexibilidade da ferramenta em diferentes cenários e modelos de aprendizado de máquina. Além disso, a inclusão de novas métricas de avaliação, como *fidelity* [Alvarez-Melis, D., e Jaakkola, T. S. 2018] e *stability* [Alvarez-Melis, D., e Jaakkola, T. S. 2018], pode permitir uma análise mais robusta das explicações geradas, garantindo uma avaliação mais abrangente. Outra possibilidade de desenvolvimento seria a adaptação da ferramenta para funcionar com dados sensíveis, implementando medidas de segurança e anonimização de dados, alinhadas às regulamentações de proteção de dados, como a LGPD no Brasil e o GDPR na Europa.

Referências

1. Loredana Coroama, Adrian Groza (2022). *Evaluation metrics for Explainable Artificial Intelligence techniques: State of the Art Review and Challenges*. https://link.springer.com/chapter/10.1007/978-3-031-20319-0_30

2. Zachary C. Lipton (2017). *The Mythos of Model Interpretability*. https://dl.acm.org/doi/pdf/10.1145/3236386.3241340

3. Dierle José Coelho Nunes, Otávio Morato de Andrade (2023). *O USO DA INTELIGÊNCIA ARTIFICIAL EXPLICÁVEL ENQUANTO FERRAMENTA PARA COMPREENDER DECISÕES AUTOMATIZADAS: POSSÍVEL CAMINHO PARA AUMENTAR A LEGITIMIDADE E CONFIABILIDADE DOS MODELOS ALGORÍTMICOS?* *https://periodicos.ufsm.br/revistadireito/article/view/69329*

4. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). *Why Should I Trust You? Explaining the Predictions of Any Classifier*. https://dl.acm.org/doi/abs/10.1145/2939672.2939778

5. Ribeiro, M. T., Singh, S., & Guestrin, C. (2018). Anchors: High-precision model-agnostic explanations. https://ojs.aaai.org/index.php/aaai/article/view/11491

6. Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., Pedreschi, D. (2018). *A survey of methods for explaining black box models*. https://dl.acm.org/doi/abs/10.1145/3236009

7. Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html

8. Doshi-Velez, F., Kim, B. (2017). *Towards a rigorous science of interpretable machine learning*. https://arxiv.org/abs/1702.08608

9. Scientific Foresight Unit (STOA) (2020). *The impact of the General Data Protection Regulation (GDPR) on artificial intelligence*.

10. Miller, T. (2019). Explanation in Artificial Intelligence: Insights from the Social Sciences. https://www.sciencedirect.com/science/article/pii/S0004370218305988

11. EUROPEAN PARLIAMENT. *EU AI Act: First regulation on artificial intelligence*. European Parliament, 1 jun. 2023. Disponível em: https://www.europarl.europa.eu/topics/en/article/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence

12. Alvarez-Melis, D., & Jaakkola, T. S. (2018). On the robustness of interpretability methods. Proceedings of the 32nd International Conference on Neural Information Processing Systems (NeurIPS 2018), 2651-2662.

13. Alvarez-Melis, D., & Jaakkola, T. S. (2018). Towards robust interpretability with self-explaining neural networks. Advances in Neural Information Processing Systems, 31, 7786-7795.

14. Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5-32. https://doi.org/10.1023/A:1010933404324

15. Montavon, G., Samek, W., & Müller, K. R. (2018). Methods for interpreting and understanding deep neural networks. Digital Signal Processing, 73, 1-15.

16. MONTENEGRO, Thiago Cunha (2024). Uma arquitetura de referência para explicabilidade como serviço na saúde: H-XAIaaS: Health eXplainable Artificial Intelligence as a Service.

17. Downey, Allen B. (2016). Pense em Python: Pense Como um Cientista da Computação.

18. Nyre-Yu, M., Morris, E., Smith, M. R., Moss, B., & Smutz, C. (2022). Explainable AI in Cybersecurity Operations: Lessons Learned from xAI Tool Deployment. Usable Security and Privacy (USEC) Symposium, 28 March 2022. Sandia National Laboratories. https://dx.doi.org/10.14722/usec.2022.23014

19. Hu, B., Tunison, P., Vasu, B., Menon, N., Collins, R., & Hoogs, A. (2021). XAITK: The explainable AI toolkit. Applied AI Letters, 2(4), e40. https://doi.org/10.1002/ail2.40

20. Alikhademi, K., Richardson, B., Drobina, E., & Gilbert, J. E. (2022). Can Explainable AI Explain Unfairness? A Framework for Evaluating Explainable AI. University of Florida. https://arxiv.org/abs/2106.07483