# O uso de árvores de decisão para explicar os resultados de algoritmos de inteligência artificial opacos

Relatório final de trabalho semestral de Iniciação Científica na graduação de Sistemas de Informação da Escola de Artes, Ciências e Humanidades da USP. Este trabalho foi desenvolvido voluntariamente ao longo do primeiro semestre de 2023 sob a orientação do Prof. Dr. Luciano Antonio Digiampietri.

**Aluno:** Julia Machado Lechi
**Orientador:** Prof. Dr. Luciano Antonio Digiampietri
**Período:** 20 de março de 2023 a 19 de setembro de 2023

**Resumo:** O uso de algoritmos de Inteligência Artificial tem se adentrado na vida das pessoas nas últimas décadas, tornando-se mais comum o uso deles, entre outras coisas, para recomendações de músicas, vídeos ou produtos durante a navegação em alguns sites. Entretanto, o uso desses algoritmos está cada vez mais afetando diretamente o cotidiano das pessoas, alguns exemplos de uso são: concessão de créditos em bancos, avaliação de currículos e recrutamento de profissionais em departamentos de RH ou identificação de suspeitos usando reconhecimento facial. Muitos dos algoritmos de inteligência artificial que apresentam melhores resultados são os chamados de black-box (caixa-preta) ou opacos, pois não é visível ou entendível a forma como os algoritmos tomam decisões ou como são treinados. Dessa forma, o intuito deste trabalho é ajudar a entender o processo de decisão dos algoritmos de inteligência artificial usando árvores de decisão. Para isso, o processo utilizado foi da comparação do desempenho entre modelos caixa-pretas e interpretáveis, e também utilizando do algoritmo opaco no processo de treinamento dos algoritmos explicáveis. Para avaliar o desempenho dos modelos, foram utilizadas as métricas F1-score, acurácia e fidelidade tanto aos modelos caixa-preta quanto ao interpretável. Os resultados obtidos mostraram que os modelos de caixa-preta geralmente alcançam resultados superiores em relação ao modelo interpretável. No entanto, a abordagem de utilizar modelos opacos como assistentes no treinamento do modelo interpretável se mostrou promissora, aumentando seu desempenho.
**Palavras-chave:** Inteligência Artificial Explicável; Discriminação Algorítmica; Mineração de Dados

# Sumário

**1. Introdução..... 3**
**2. Trabalhos relacionados..... 5**
**3. Objetivo..... 11**
**4. Materiais e Métodos..... 12**
**5. Resultados e Discussão..... 17**
**5.1 Desempenho do Modelo..... 17**
**5.2 Fidelidade do Modelo Interpretável..... 19**
**5.3 Melhorias no Modelo Interpretável..... 21**
**6. Conclusão..... 24**
**7. Referências..... 26**

## 1. Introdução

Na sociedade atual, também conhecida como sociedade da informação, os algoritmos de inteligência artificial têm conquistado cada vez mais seu espaço, desempenhando um papel significativo na tomada de decisões. Esses algoritmos de inteligência são projetados para analisar a imensa quantidade de dados que produzimos para as grandes empresas em seus aplicativos e sites, para oferecer percepções e fornecer recomendações em várias áreas. Algumas dessas decisões são apresentadas de formas mais sutis, que muitas vezes não são percebidas pelas pessoas ou que afetam pouco seu dia a dia, porém há outras decisões que podem causar grande impacto. Um caso extremo foi um erro ocorrido em um carro autônomo que ocasionou um atropelamento de um ciclista que veio a óbito [1].

Este é um exemplo extremo que envolveu potenciais falhas de projeto, de desenvolvimento e de testes, porém, mesmo quando os algoritmos fazem exatamente o que foram implementados para fazer (o que, no caso de um algoritmo de inteligência artificial, tipicamente significa tomar uma decisão que maximize uma medida de desempenho) eles também podem influenciar negativamente a vida de pessoas. Por exemplo, um algoritmo que sugira a um funcionário da seção de recursos humanos a não contratar (ou não chamar para uma entrevista) uma dada pessoa, ou que sugira a não concessão de créditos a alguém que está solicitando, ou que indique que uma pessoa deva ser detida/investigada por meio do uso de reconhecimento facial ou da análise de características socioeconômicas [2]. No entanto, é necessário questionar a imparcialidade dos resultados, uma vez que os algoritmos são treinados por pessoas e podem refletir vieses ou comportamentos tendenciosos das empresas. *A transparência do algoritmo é um fator essencial para*

*responsabilizar as organizações por seus produtos, serviços e comunicação de informações* [3]. Adicionalmente, não só para empresas, mas pensando em um caso de utilização pública, em casos de saúde, por exemplo, a transparência torna mais segura a revisão e o uso por profissionais da área.

A utilização de algoritmos de inteligência artificial oferece diversas vantagens, como desempenho superior em certas tarefas e capacidade de processar grandes conjuntos de dados. Muitos dos algoritmos de inteligência artificial que atualmente apresentam os melhores desempenhos são chamados de opacos ou caixa-preta [4], significando que não apresentam de forma clara e inteligível por humanos a maneira como suas decisões são tomadas. Por exemplo, é muito difícil de entender o porquê um algoritmo de aprendizado profundo fez uma dada classificação (ou sugeriu um valor contínuo específico para um problema de regressão).

Há diferentes formas de se explicar resultados de algoritmos de classificação. Entre as mais comuns está o uso direto de algoritmos considerados inerentemente explicáveis ou interpretáveis, como árvores de decisão ou regressão linear; é possível tentar explicar um modelo apresentando exemplos e contraexemplos de classificações “próximas” a uma instância em avaliação (abordagem conhecida como contrafactual); e também é possível utilizar algoritmos explicáveis para tentar explicar local (em torno de uma instância) ou globalmente os resultados de algoritmos considerados opacos ou caixa-preta (abordagem também conhecida como uso de modelo substituto) [5].

Este trabalho visa a explorar a importância da transparência nos algoritmos de inteligência artificial inerentemente explicáveis, indo além da busca de ótimos resultados, mas considerando outros aspectos relevantes, como por exemplo a compreensão clara do caminho percorrido por um

algoritmo até o resultado final, permitindo que usuários e profissionais entendam o processo de tomada de decisão e identifiquem possíveis tendências ou erros. Isso proporciona uma maior confiança no processo de implementar a automatização em áreas que seriam úteis, minimizando erros e injustiças, e também viabilizando a interação mais efetiva entre humano-máquina.

## 2. Trabalhos relacionados

Nos últimos anos, a criação de diferentes regulamentações relacionadas à proteção de dados pessoais dos usuários e ao direito de explicação (como GDPR na Europa e LGPD no Brasil) intensificou a discussão sobre o uso e o impacto dos algoritmos de inteligência artificial na sociedade. Em particular, o direito à explicação levantou questões sobre a viabilidade e a legalidade do uso de alguns algoritmos que afetam a vida cotidiana das pessoas. Embora essa discussão não seja nova, ela tem se tornado cada vez mais relevante na comunidade de desenvolvimento de software e na sociedade como um todo.

Em algumas áreas, como na medicina, sempre se questionou como poderiam ser utilizados os diagnósticos gerados por computador. Na área de desenvolvimento de carros autônomos [6], esse assunto também é amplamente discutido, incluindo a responsabilidade legal no caso de acidentes causados por problemas de projeto ou implementação. Além desses casos específicos em que vidas humanas podem estar diretamente em risco devido à ação de algoritmos de IA, existem diversas outras atividades que afetam diretamente a vida das pessoas. Por exemplo, na área financeira, existem diversas empresas que utilizam algoritmos para auxiliar no processo de concessão ou não de financiamentos. Também existem empresas que usam modelos de IA em seus setores de recursos humanos. Além disso, vários serviços públicos e de

segurança usam diferentes modelos de IA, incluindo algoritmos de reconhecimento facial, como parte de seus procedimentos. [7, 8, 9 10]

Apesar de não haver uma definição padronizada do que seja um algoritmo (ou modelo) interpretável ou explicável, juntamente com a subjetividade em várias definições, a maioria das definições compartilha algumas ideias. Em particular, um modelo interpretável é um modelo que permite que um ser humano entenda o que fez com que o modelo produzisse um determinado resultado. Vale ressaltar que pessoas diferentes, com formações diversas, precisarão de informações distintas para compreender efetivamente um resultado. Por exemplo, para um radiologista, um destaque em uma radiografia indicando uma mancha pode ser suficiente para explicar um diagnóstico enquanto, para um leigo, isso pode não fazer sentido.

Em termos de modelos de aprendizado de máquina, duas características principais são frequentemente usadas para indicar o quão interpretável um modelo é: (i) a natureza intrínseca de como o “conhecimento” é representado no modelo; e (ii) o tamanho do modelo. Por exemplo: em uma regressão linear, os coeficientes relacionados a cada atributo ou característica são aprendidos, já em uma árvore de decisão, os nós de decisão (condições e valores) são aprendidos.

O significado do valor de um coeficiente em uma regressão linear ou um nó de decisão em uma árvore de decisão geralmente é considerado mais fácil de entender do que os pesos em uma Rede Neural Profunda. Por outro lado, o tamanho do modelo também é importante. Entender a importância e o significado dos coeficientes em uma regressão linear que usa três atributos é considerado mais fácil do que em uma regressão que usa cem deles, assim como uma árvore de decisão de altura três é mais “inteligível” do que uma de altura 20.

Modelos cuja interpretação não é simples, costumam ser considerados opacos, mesmo para quem entende a lógica por trás da construção desses

modelos. Por exemplo, redes neurais profundas ou modelos produzidos a partir de grandes modelos de linguagem são frequentemente considerados opacos. Vale ressaltar que a opacidade pode ser atribuída diretamente com base no tipo de algoritmo que produziu o modelo, ou considerando os atributos que foram utilizados. Por exemplo, se os atributos usados para construir o modelo foram produzidos a partir de um processo de projeção (por exemplo, usando análise de componentes principais), o modelo resultante, mesmo que produzido por um algoritmo considerado inerentemente interpretável, será considerado opaco, porque os recursos usados não têm sentido claro para um ser humano.

O conceito de caixa-preta (black-box) é usado há décadas em Engenharia de Software. Considerando um sistema, um componente ou uma função como caixa-preta significa tratar esse recurso como se você não soubesse (ou não se importasse) com seu comportamento interno. Como um exemplo, testando um recurso considerando que dada uma entrada, você saberá a saída produzida, mas sem saber nada sobre o processo que levou a produção da saída, isso será considerado como um processo de caixa-preta.

No contexto da Inteligência Artificial Explicável, é normal observar o termo caixa-preta em duas situações: para se referir a algoritmos ou modelos opacos (frequentemente caixa-preta ou opaco são usados alternadamente); e na análise de modelos, ignorando o seu funcionamento interno, considerando apenas as saídas produzidas por suas respectivas entradas (aqui o termo tem o mesmo significado usado na Engenharia de Software). Em um último caso, independentemente se o modelo é interpretável ou não, ou se o acesso do código ou modelo está disponível ou não, isso será tratado como uma caixa-preta (considerando a última definição dada), pois não possuem o código-fonte ou mesmo as entradas originais que foram utilizadas para construir um sistema que, por exemplo, foi adquirido de uma empresa terceirizada.

O presente trabalho utiliza o termo caixa-preta neste segundo contexto: independente da natureza do algoritmo que produziu o modelo, será considerado uma caixa-preta se não tivermos acesso nem aos dados de treinamento utilizados para sua produção ou aos detalhes internos do modelo. Como será detalhado na próxima seção, em todos os experimentos realizados temos acesso a dados e modelos resultantes. No entanto, para avaliar os resultados da maneira que geralmente ocorre em condições do mundo real, as análises tratarão esses modelos como caixas-pretas.

A inteligência artificial explicável, por sua vez, corresponde a uma área que estuda, de diferentes formas, como a IA pode produzir resultados que podem ser interpretados pelos seres humanos. Normalmente, essa interpretação (ou compreensão) ocorre de três maneiras principais, apresentadas e detalhadas a seguir.

1. **Desenvolvimento de Modelos Inerentemente Explicáveis**: muitos dos primeiros modelos utilizados em sistemas de inteligência artificial foram considerados inerentemente interpretáveis como, por exemplo, modelos baseados em árvores de decisão, regressão linear ou modelos que utilizam regras de associação. Ao longo das décadas, modelos cada vez mais complexos foram desenvolvidos, e a compreensão da razão ou lógica que leva esses modelos a produzir seus resultados é cada vez mais nebulosa. São, por exemplo, soluções produzidas por redes neurais convolucionais, redes profundas em geral, soluções que utilizam projeções de atributos de entrada e soluções baseadas em grandes modelos de linguagem. Este ramo da XAI (*Explanaible IA*) visa a desenvolver novos algoritmos inerentes explicáveis ou melhorar os já existentes,

seja modificando algumas de suas características ou refinando o treinamento para produzir modelos melhores.

2. **Explicação dos Modelos Opaco/Caixa-Preta:** devido ao fato de vários modelos de caixa-preta terem atingido o estado da arte para algumas famílias específicas de problemas, muitos pesquisadores consideram relevante tentar explicar por que esses modelos produzem seus resultados. A explicação de um modelo pode ser global (ou seja, tentando explicar o modelo como um todo) ou local (ou seja, a explicação em torno de um exemplo de entrada específica). Existem duas abordagens principais para explicar tais modelos [11]. Na primeira, um modelo explicável é construído para imitar o comportamento do modelo de caixa-preta. Esta abordagem é conhecida como uso de modelo substituto). Para isso, durante o treinamento do modelo explicável, a saída do modelo caixa-preta é usada como os valores da variável de destino. Na segunda abordagem, o modelo caixa-preta é explicado com base na importância estimada de cada uma de suas características de entrada [12]. Normalmente, várias saídas são produzidas variando o valor de diferentes recursos e o impacto no atributo alvo é analisado em relação a essas variações.
3. **Apresentação de Exemplos ao Utilizador:** este ramo da XAI assume que o entendimento humano, e em particular aqueles que não são especialistas em IA, pode beneficiar da apresentação de um conjunto de exemplos. A justificativa é que, a partir de um conjunto de exemplos de classificação (dados de entrada e respectivas saídas), uma pessoa pode ter uma ideia melhor de como o modelo funciona. Em particular, este ramo trabalha com exemplos contrafactuais [13]. Ou seja, exemplos em que a saída foi diferente de algum exemplo de entrada específico. Esse tipo de

> abordagem geralmente é usada não apenas para explicar a lógica de um modelo, mas também para orientar o usuário sobre o que ele pode fazer para obter uma saída diferente. Por exemplo, se um pedido de empréstimo foi negado, é possível, a partir de exemplos contrafactuais “próximos” dos dados inseridos pelo usuário, apresentar o que seria necessário para que o pedido fosse aceito (por exemplo, solicitar uma quantidade menor de dinheiro ou propor o pagamento em um número diferente de parcelas).

Finalmente, a área de Justiça Algorítmica, que está intimamente relacionada ao XAI, também teve um desenvolvimento significativo nos últimos anos. Embora também careça de uma definição precisa e comum, *Algorithmic Fairness* normalmente se refere a tentativas de corrigir o viés algorítmico em processos automatizados de tomada de decisão [14 ,15]

Em seu trabalho, Suresh e Guttag [16] identificaram fontes potenciais podem levar a sete vieses em algoritmos de aprendizado de máquina: viés histórico, viés de apresentação, viés de medição, viés de agregação, viés de aprendizado, avaliação viés e viés de implantação. Embora cada um desses vieses tenha características específicas e é produzido de diferentes maneiras (por exemplo, se historicamente houve um preconceito na sociedade que eventualmente é reproduzido pelo algoritmo ou o algoritmo está sendo usado incorretamente), a XAI possui ferramentas para auxiliar na detecção e possível correção desses vieses. A compreensão geral de um modelo de IA, seja a partir de um modelo interpretável ou de diferentes tipos de explicação de modelos opacos, permite uma avaliação além das métricas de desempenho tradicionais. Isso, se feito por um especialista comprometido em minimizar vieses, permite que o modelo seja revisto ou melhorado antes mesmo de sua implantação, para que possa ser considerado mais justo.

## 3. Objetivo

No contexto dos algoritmos opacos, muito se tem discutido sobre sua real necessidade, principalmente em problemas que requerem explicações. Um dos principais argumentos nessa discussão é que é impossível explicar efetivamente esses modelos, seja com base em modelos mais simples ou através da identificação da importância de alguns atributos de entrada para o modelo construído [17, 18].

O presente trabalho visou a contribuir para essas discussões, partindo de três questões de pesquisa:

1. Quão fiel um modelo interpretável pode ser a uma contraparte opaca?
2. Existe uma diferença significativa entre os resultados dos modelos produzidos por um algoritmo interpretável e os produzidos por algoritmos considerados caixas-pretas?
3. É possível construir modelos explicáveis melhores, com base nas saídas produzidas por um algoritmo caixa-preta?

O presente trabalho trata destes três aspectos da XAI. Inicialmente, este trabalho avalia a explicação de modelos de caixa-preta por meio de modelos inerentemente explicáveis (Questão de Pesquisa 1). Neste caso, usamos árvores de decisão de altura três. Este trabalho também investiga se é possível construir modelos explicáveis com base nos resultados de modelos de caixa-preta (questão de Pesquisa 3). A análise dos resultados produzidos visa a responder à Questão de Pesquisa 2, contribuindo para a discussão sobre a efetiva necessidade de modelos opacos em problemas de diversas naturezas.

Em relação à avaliação, existem diferentes maneiras de avaliar o resultado de uma explicação [1,8,14]. Uma das mais robustas é verificar com um grande conjunto de usuários o quão adequadas são as explicações dadas sobre um modelo. No entanto, devido à complexidade e custo de questionar um

grande número de usuários, também existem métricas que podem ser obtidas automaticamente, para avaliar a simplicidade de um modelo e a qualidade de alguns métodos de explicações. Devido à sua simplicidade, duas das métricas mais utilizadas são o tamanho do modelo, que verifica quão simples (ou seja, o quão fácil de entender) ele é; e a fidelidade de algum modelo quando usado para explicar outro.

Como uma instância de um modelo interpretável e simples de entender, neste trabalho aplicamos árvores de decisão de altura três. A qualidade explicativa deste modelo, ao explicar sua contraparte de uma caixa-preta, foi medida por sua fidelidade, ou seja, a taxa de coincidência entre as previsões feitas pela caixa-preta e o modelo interpretável. Embora essa medida possa ser considerada equivalente à precisão, em vez de comparar a saída do modelo interpretável com a variável alvo do conjunto de teste, essa comparação é feita com a saída do modelo de caixa-preta

## 4. Materiais e Métodos

Oitos conjuntos de dados públicos comumente usados para a construção e validação de modelos de inteligência artificial foram arbitrariamente selecionados a partir dos repositórios UCI[1] e Kaggle [20][2]. Todos os conjuntos de dados passaram por uma etapa de pré-processamento semelhante, na qual os dados categóricos foram convertidos em dados numéricos e os valores ausentes foram substituídos por valores padrão. Problemas com mais de duas classes foram reduzidos para duas classes, em particular, havia conjuntos de dados com quatro classes seguindo uma gradação (por exemplo: muito ruim, ruim, bom e muito bom), caso em que as classes foram mapeados para duas

[1] https://archive.ics.uci.edu
[2] https://www.kaggle.com/

(ruim e bom) para tratar apenas problemas binários neste trabalho. A Tabela 1 resume as características dos conjuntos de dados usados após esta etapa de pré-processamento.

Tabela 1: Descrição dos conjuntos de dados

| **Nome** | **N° de Linhas** | **Porcentagem de elementos na classe majoritária** |
|---|---|---|
| Câncer de mama | 286 | 70,3% |
| Votação do Congresso | 435 | 61,4% |
| Diabetes | 768 | 65,1% |
| Crédito Alemão | 1000 | 70% |
| Insuficiência cardíaca | 918 | 55,3% |
| Popularidade das notícias on-line | 39644 | 50,7% |
| Compradores on-line | 12330 | 84,5% |
| Desempenho do aluno | 649 | 50,5% |

Embora saibamos que seria possível aplicar outras tarefas de pré-processamento mais sofisticadas, decidimos manter esta etapa o mais simples possível, produzindo dados adequados para as próximas etapas do processo, uma vez que utilização de técnicas mais complexas e específicas para cada conjunto de dados não fez parte dos principais objetivos deste trabalho e não ajudaria a responder às questões de pesquisa.

Sete algoritmos foram selecionados para criar os modelos. Esta seleção foi feita com objetivo de usar modelos que são construídos sob diferentes princípios. A Tabela 2 mostra os algoritmos utilizados e seus parâmetros. Notavelmente, Árvores de Decisão foram usadas em dois contextos diferentes: como modelo interpretável com altura limitada a três e como um modelo de caixa-preta com valores padrão para todos os seus parâmetros. Valores de

parâmetros padrão foram usados para todos os modelos considerados caixa-preta exceto o número máximo de interações para regressão logística e Perceptron multicamadas, aumentamos este número para garantir a convergência desses modelos. SVM também teve duas implementações, ambas usadas como caixa-preta: uma usando um kernel linear e a outra usando um kernel polinomial. Neste trabalho, usamos as implementações desses algoritmos disponíveis na Biblioteca Python Scikit-Learn[3].

Tabela 2: Algoritmos e parâmetros

| **Algoritmo** | **Parâmetros** | **Uso neste trabalho** |
|---|---|---|
| Árvore de Decisão | altura máxima = 3 | Interpretável |
| Árvore de Decisão | (padrão) | Caixa-preta |
| Random Forest | (padrão) | Caixa-preta |
| SVM | (padrão) | Caixa-preta |
| SVM | kernel = polinomial | Caixa-preta |
| Regressão Logística | iteração máxima = 5000 | Caixa-preta |
| Multilayer Perceptron | iteração máxima = 5000 | Caixa-preta |
| Gaussian Naive Bayes | (padrão) | Caixa-preta |
| KNN | (padrão) | Caixa-preta |

Após o pré-processamento, cada conjunto de dados foi dividido aleatoriamente em três subconjuntos, através de amostragem aleatória estratificada a partir de uma distribuição uniforme. Como resultado, 40% dos dados construíram o primeiro conjunto de treinamento e outros 40% o segundo, com os 20% restantes dos dados sendo considerados como um conjunto de testes, conforme ilustrado na Figura 1.

[3] https://scikit-learn.org/stable/


Figura 1: Estratégia de divisão estratificada do conjunto de dados

Essa divisão se justifica pela necessidade de diferentes conjuntos de treinamento para modelos de caixa-preta e pelos modelos interpretáveis que pretendem explicar essas caixa-pretas. Em particular, quando uma empresa usa um sistema de caixa-preta, é comum não ter acesso aos dados de treinamento originais para esses sistemas. Esta divisão de conjuntos torna então possível simular esta situação.

Depois de dividir os dados, adotamos uma abordagem de quatro etapas para esta pesquisa, como ilustrado na Figura 2. Vale ressaltar, neste ponto, que este processo foi repetido para cada combinação de modelo de caixa-preta e conjunto de dados. Na primeira etapa (Figura 2A), o conjunto de Treinamento 1 é usado para treinar tanto um modelo interpretável (Modelo Interpretável 1) e o modelo de caixa-preta. O objetivo do Modelo Interpretável 1 não é explicar a caixa-preta, mas sim para ser usado na análise comparativa dos modelos para responder à Questão de Pesquisa 2.

Na segunda etapa (Figura 2B), o conjunto de treinamento 2 é aplicado ao modelo de caixa-preta, fazendo suas previsões para esses dados (*predição*). Em seguida, no terceiro estágio (Figura 2C), dois modelos interpretáveis adicionais (modelos interpretáveis 2 e 3) são treinados usando o conjunto de treinamento 2. Um deles tem como valores para a variável alvo os valores do conjunto de treinamento, enquanto o outro assume a saída produzida pelo modelo de caixa-preta (*predição*). Finalmente, na quarta etapa (Figura 2D), todos os quatros modelos fazem suas previsões para o conjunto de testes. Essas previsões são, por sua vez, usadas para avaliar o desempenho dos modelos.

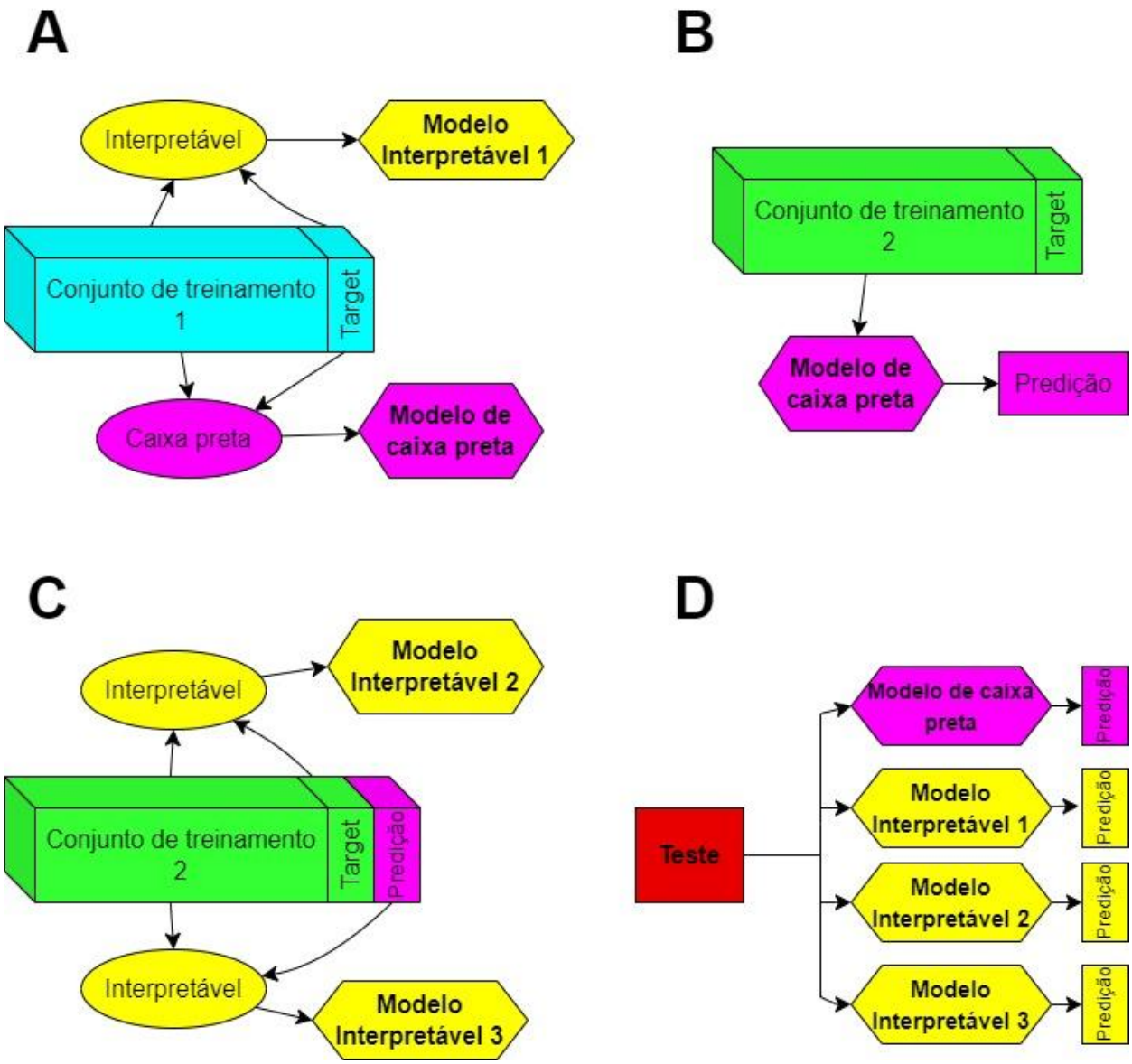


Figura 2: As quatros partes desta pesquisa.

No presente trabalho, todos os modelos foram avaliados usando precisão e medida F1-macro. A capacidade explicativa do modelo interpretável 3 em relação aos modelos de caixa preta foi avaliada por meio da medida de fidelidade, a fim de responder à Questão de Pesquisa 1. A Questão de Pesquisa 3, por sua vez, será respondida com base na avaliação das saídas do modelo

interpretável 3, quando comparado com os valores da variável de destino do conjunto de testes.

## 5. Resultados e Discussão

Iniciamos nossa discussão dos resultados obtidos apresentando o desempenho dos modelos treinados no conjunto de treinamento 1 (seção 5.1). Em seguida, na seção 5.2, analisamos nossos resultados em relação à fidelidade. Finalmente, na seção 5.3, descrevemos nossas descobertas relacionadas à melhoria do treinamento do modelo interpretável, com base nas saídas dos modelos de caixa-preta.

### 5.1 Desempenho do Modelo

A precisão do modelo alcançada ao treinar com o conjunto de treinamento 1 pode ser vista na tabela 3, com as pontuações F1 correspondentes sendo mostradas na tabela 4. Para facilitar a visualização, nessa tabela, as células com fundo verde mais intenso correspondem aos valores mais altos. Os valores mais baixos são indicados por células em vermelho mais intenso

Tabela 3: Acurácia para o Modelo Interpretável 1 e para cada modelo de caixa-preta

| Dataset | Interpretable Model 1 | Decision Tree | Random Forest | SVM (linear) | SVM (polynomial) | Logistic Regression | Multilayer Perceptron | Gaussian Naive Bayes | KNN |
|---|---|---|---|---|---|---|---|---|---|
| Breast Cancer | 0.707 | 0.690 | 0.759 | 0.690 | 0.690 | 0.741 | 0.690 | 0.690 | 0.741 |
| Congressional Voting | 0.954 | 0.954 | 0.954 | 0.954 | 0.931 | 0.931 | 0.943 | 0.954 | 0.908 |
| Diabetes | 0.721 | 0.727 | 0.792 | 0.760 | 0.773 | 0.779 | 0.753 | 0.766 | 0.740 |
| German Credit | 0.685 | 0.700 | 0.725 | 0.710 | 0.700 | 0.715 | 0.505 | 0.725 | 0.690 |
| Heart Failure | 0.848 | 0.793 | 0.908 | 0.701 | 0.734 | 0.897 | 0.864 | 0.897 | 0.636 |
| Online News Popularity | 0.618 | 0.567 | 0.661 | 0.542 | 0.555 | 0.605 | 0.499 | 0.528 | 0.571 |
| Online Shoppers Purchasing | 0.878 | 0.864 | 0.905 | 0.841 | 0.842 | 0.878 | 0.877 | 0.822 | 0.861 |
| Student Performance | 0.631 | 0.615 | 0.731 | 0.723 | 0.762 | 0.715 | 0.685 | 0.754 | 0.708 |

Tabela 4: Resultados do F1 score do Modelo Interpretável 1 e de cada modelo de caixa-preta

| Dataset | Interpretable Model 1 | Decision Tree | Random Forest | SVM (linear) | SVM (polynomial) | Logistic Regression | Multilayer Perceptron | Gaussian Naive Bayes | KNN |
|---|---|---|---|---|---|---|---|---|---|
| Breast Cancer | 0.61 | 0.61 | 0.67 | 0.41 | 0.53 | 0.66 | 0.41 | 0.64 | 0.59 |
| Congressional Voting | 0.95 | 0.95 | 0.95 | 0.95 | 0.93 | 0.93 | 0.94 | 0.95 | 0.9 |
| Diabetes | 0.68 | 0.7 | 0.77 | 0.72 | 0.72 | 0.74 | 0.7 | 0.74 | 0.69 |
| German Credit | 0.57 | 0.65 | 0.61 | 0.5 | 0.47 | 0.57 | 0.5 | 0.64 | 0.54 |
| Heart Failure | 0.84 | 0.79 | 0.9 | 0.7 | 0.72 | 0.89 | 0.86 | 0.89 | 0.63 |
| Online News Popularity | 0.62 | 0.57 | 0.66 | 0.53 | 0.53 | 0.6 | 0.35 | 0.41 | 0.56 |
| Online Shoppers Purchasing | 0.69 | 0.75 | 0.79 | 0.46 | 0.46 | 0.7 | 0.69 | 0.69 | 0.61 |
| Student Performance | 0.62 | 0.61 | 0.72 | 0.72 | 0.75 | 0.71 | 0.68 | 0.75 | 0.71 |

Pode-se observar que, para os modelos e conjuntos de dados selecionados, Random Forest foi o modelo que apresentou a melhor precisão em sete dos oito conjuntos de dados (empatando com outros quatro modelos no conjunto de dados Congressional Voting), perdendo sua posição para SVM com kernel polinomial apenas no conjunto Student Perfomance, onde ficou em terceiro lugar (cerca de 4% pior que o SVM).

Em termos de precisão, o desempenho do modelo interpretável foi, em média[4], 93,2% do desempenho do melhor modelo para cada um dos conjuntos de dados (variando de 82,3% para o conjunto de dados Student Perfomance a 100%, ou seja, igual ao melhor resultado obtido, para o conjunto de dados Congressional Voting).

Considerando a medida F1, o modelo com melhor desempenho foi novamente o Random Forest, obtendo o melhor resultado para seis dos oitos conjuntos de dados. O modelo Naive Bayes Gaussiano alcançou a melhor pontuação F1 para os conjuntos German Credit e Student Perfomance, empatando, neste último caso, com o SVM com kernel polinomial.

Em média, a pontuação F1 do modelo interpretável correspondeu a 90,5% dos valores obtidos pelos modelos de melhor desempenho (variando de 82,7% para o conjunto Student Perfomance a 100% para o conjunto Congressional Voting). Com base nesses resultados, pode-se dizer que o modelo interpretável utilizado neste trabalho teve desempenho pior (em média) do que o melhor modelo para cada um dos conjuntos de dados. Assim a Questão de Pesquisa 2 foi respondida confirmando que há diferença de desempenho entre eles.

## 5.2 Fidelidade do modelo interpretável

Nesta seção, avaliamos a capacidade do Modelo Interpretável 3 em mimetizar os resultados produzidos pelos modelos de caixa-preta, calculada por meio da medida de fidelidade. A tabela 5 apresenta a fidelidade dos resultados do modelo interpretável 3 em relação aos resultados dos modelos de caixas pretas.

[4] Todas as médias neste trabalho correspondem à média aritmética

Tabela 5: Resultado da fidelidade do Modelo Interpretável 3

| Dataset | Decision Tree | Random Forest | SVM (linear) | SVM (polynomial) | Logistic Regression | Multilayer Perceptron | Gaussian Naive Bayes | KNN |
|---|---|---|---|---|---|---|---|---|
| Breast Cancer | 0.741 | 0.810 | 1.000 | 0.897 | 0.879 | 1.000 | 0.948 | 0.966 |
| Congressional Voting | 1.000 | 1.000 | 0.989 | 0.977 | 0.977 | 0.966 | 1.000 | 0.977 |
| Diabetes | 0.675 | 0.929 | 0.935 | 0.935 | 0.935 | 0.870 | 0.857 | 0.831 |
| German Credit | 0.715 | 0.865 | 0.995 | 1.000 | 0.940 | 0.880 | 0.945 | 0.930 |
| Heart Failure | 0.826 | 0.924 | 0.891 | 0.929 | 0.918 | 0.913 | 0.897 | 0.832 |
| Online News Popularity | 0.617 | 0.745 | 0.976 | 0.931 | 0.831 | 0.993 | 0.975 | 0.711 |
| Online Shoppers Purchasing | 0.881 | 0.952 | 1.000 | 1.000 | 0.981 | 0.978 | 0.899 | 0.980 |
| Student Performance | 0.715 | 0.792 | 0.785 | 0.785 | 0.715 | 0.708 | 0.762 | 0.685 |

A fidelidade média geral foi de 88,9%. Ao considerar cada modelo de caixa-preta, observa-se que, em média, a maior fidelidade do modelo interpretável 3 ocorreu com o SVM utilizando kernel linear (94,6%). Por outro lado, o pior desempenho médio foi de 77,1%, com a Árvore de Decisão sem limite de altura. Este último resultado é bastante interessante, pois indica que uma árvore menor (altura três) não poderia imitar bem os resultados de uma árvore que não tivesse limite de altura.

Aqui vale ressaltar que o modelo de árvore de decisão tratado como caixa-preta obteve, em média, o pior valor para o F1 score e um dos piores resultados de acurácia, ambos valores inferiores aos resultados do Modelo Interpretável 3 (ver seção 5.1). Isso indica que o modelo não foi capaz de construir regras que generalizasse o conhecimento aprendido (não limitar a altura da árvore pode ter produzido um modelo excessivamente específico, ou seja, super ajustado, e essa característica não pode ser mimetizada pelo modelo interpretável)

Além da fidelidade média (88,9%), bem como a menor e a maior média por modelo (respectivamente 77,1% e 94,6%), consideramos adequado comparar esses valores com os valores equivalentes produzidos pelo Modelo

Interpretável 1. É importante lembrar que o Modelo Interpretável 1 não foi construído para imitar os modelos de caixa-preta, mas sim para resolver o mesmo problema que os modelos de caixa-preta estavam resolvendo. Em média, a interseção dos resultados do modelo interpretável com os modelos de caixa-preta (que equivale à fidelidade) foi de 79,2%, variando de 70,5% para o modelo baseado em Árvores de Decisão a 86,4% para o SVM com kernel linear.

Assim, em média, o Modelo Interpretável 3 conseguiu imitar satisfatoriamente os modelos caixa-pretas, significativamente melhor do que o modelo correspondente (Modelo Interpretável 1) que não foi treinado para esse fim, respondendo assim à Questão de Pesquisa 1. Vale ressaltar que, conforme apresentado na literatura (por exemplo, em [18]), mesmo altos valores de fidelidade podem não significar que a mesma lógica “aprendida” pelo modelo de caixa-preta também foi “aprendida” pelo modelo interpretável, e isso é uma das principais limitações da explicação deste tipo específico de modelos.

### 5.3 Melhorias no modelo interpretável

Conforme explicado na seção 4, os modelos interpretáveis foram construídos usando a saída dos modelos de caixa-preta (o que foi chamado de Modelo Interpretável 3). A fidelidade desses modelos foi apresentada na seção anterior. Além da fidelidade, este trabalho levantou a hipótese de que modelos interpretáveis construídos a partir da saída dos modelos caixa-preta podem superar os modelos interpretáveis treinados usando a variável de destino do conjunto de treinamento. Essa hipótese foi baseada no fato de que os modelos de caixa-preta correspondem a simplificações do mundo e, portanto, pode ser mais fácil para o modelo interpretável aprender com essa simplificação em vez de aprender diretamente com os dados originais.

Para realizar esta análise, o desempenho dos modelos interpretáveis 2 e 3 foram comparados (ambos foram construídos com base no conjunto de treinamento 2). Nesse caso, enquanto o modelo 2 utilizou como variável alvo os

dados do conjunto de treinamento, o modelo interpretável 3 usou a saída de cada modelo de caixa-preta como alvo. As tabelas 6 e 7 apresentam os resultados de acurácia e medida F1. Nessas tabelas, a primeira coluna indica o conjunto de dados, a segunda mostra os resultados do modelo interpretável 2 e as demais colunas contêm os resultados do modelo interpretável 3, de acordo com a saída de cada modelo de caixa-preta.

Tabela 6: Comparação da acurácia do Modelo Interpretável 2 e 3

| Dataset | Interpretable Model 2 | Decision Tree | Random Forest | SVM (linear) | SVM (polynomial) | Logistic Regression | Multilayer Perceptron | Gaussian Naive Bayes | KNN |
|---|---|---|---|---|---|---|---|---|---|
| Breast Cancer | 0.707 | 0.638 | 0.776 | 0.690 | 0.690 | 0.793 | 0.690 | 0.672 | 0.741 |
| Congressional Voting | 0.943 | 0.954 | 0.954 | 0.943 | 0.954 | 0.954 | 0.931 | 0.954 | 0.908 |
| Diabetes | 0.727 | 0.675 | 0.760 | 0.747 | 0.747 | 0.753 | 0.740 | 0.753 | 0.714 |
| German Credit | 0.685 | 0.635 | 0.740 | 0.705 | 0.700 | 0.745 | 0.535 | 0.710 | 0.710 |
| Heart Failure | 0.859 | 0.859 | 0.853 | 0.690 | 0.728 | 0.891 | 0.842 | 0.870 | 0.641 |
| Online News Popularity | 0.631 | 0.619 | 0.625 | 0.542 | 0.546 | 0.586 | 0.499 | 0.522 | 0.578 |
| Online Shoppers Purchasing | 0.894 | 0.888 | 0.897 | 0.841 | 0.841 | 0.881 | 0.879 | 0.853 | 0.861 |
| Student Performance | 0.654 | 0.623 | 0.615 | 0.662 | 0.638 | 0.615 | 0.654 | 0.638 | 0.577 |

Como pode ser visto na Tabela 6, o Modelo Interpretável 2 obteve apenas três dos melhores resultados de acurácia, mostrando que os modelos treinados com a saída dos modelos de caixa-preta apresentaram melhores resultados para cinco dos conjuntos de dados. Destaca-se o modelo interpretável construído a partir da saída do modelo Random Forest, alcançando cinco resultados superiores aos do Modelo Interpretável 2.

Tabela 7: Comparação do F1 score do Modelo Interpretável 2 e 3

| Dataset | Interpretable Model 2 | Decision Tree | Random Forest | SVM (linear) | SVM (polynomial) | Logistic Regression | Multilayer Perceptron | Gaussian Naive Bayes | KNN |
|---|---|---|---|---|---|---|---|---|---|
| Breast Cancer | 0.47 | 0.49 | 0.72 | 0.41 | 0.41 | 0.73 | 0.41 | 0.60 | 0.56 |
| Congressional Voting | 0.94 | 0.95 | 0.95 | 0.94 | 0.95 | 0.95 | 0.93 | 0.95 | 0.90 |
| Diabetes | 0.65 | 0.67 | 0.72 | 0.68 | 0.69 | 0.70 | 0.68 | 0.72 | 0.67 |
| German Credit | 0.61 | 0.62 | 0.57 | 0.48 | 0.47 | 0.61 | 0.53 | 0.59 | 0.53 |
| Heart Failure | 0.85 | 0.85 | 0.85 | 0.68 | 0.72 | 0.89 | 0.83 | 0.86 | 0.64 |
| Online News Popularity | 0.63 | 0.62 | 0.62 | 0.53 | 0.51 | 0.58 | 0.34 | 0.40 | 0.56 |
| Online Shoppers Purchasing | 0.80 | 0.73 | 0.78 | 0.46 | 0.46 | 0.71 | 0.69 | 0.70 | 0.59 |
| Student Performance | 0.63 | 0.61 | 0.59 | 0.64 | 0.61 | 0.60 | 0.64 | 0.58 | 0.58 |

Em relação à medida F1, o Modelo Interpretável 2 obteve o melhor desempenho apenas em dois conjuntos de dados (Online News Popularity e Online Shoppers Purchasing). Os modelos interpretáveis construídos usando a saída do Modelo de Regressão Logística tiveram desempenho igual ou melhor que o Modelo Interpretável 2 em cinco dos oitos conjuntos de dados. O mesmo ocorreu com os modelos que utilizam como entrada a saída do modelo baseado em Árvore de Decisão sem limite de altura.

Esses resultados ajudam a fornecer uma resposta inicial para a Questão de Pesquisa 3, sobre como tentar construir modelos explicáveis com base nos resultados de modelos de caixa-preta, indicando que treinar modelos interpretáveis com base nos valores da variável alvo do conjunto de treinamento nem sempre oferece os melhores desempenhos, quando comparados a modelos treinados usando a saída de outros modelos de caixa-preta. Investigações adicionais são necessárias, no entanto, bem como o cálculo de diferentes medidas estatísticas. Nossos resultados, no entanto, indicam que este é um local interessante para investigação.

## 6. Conclusões

O aprendizado de máquina revolucionou a maneira como o conhecimento pode ser descoberto a partir dos dados. Desde o surgimento da Inteligência Artificial, diferentes modelos foram criados permitindo não apenas a resolução automatizada de problemas, mas também uma melhor compreensão dos dados.

Embora o desejo de entender os processos que envolvem as decisões tomadas por modelos criados por IA não seja algo novo e seja fundamental em algumas áreas, como a saúde, algumas das técnicas mais recentes possuem uma complexidade inerente que impossibilita o entendimento completo do processo que leva a cada uma de suas saídas. Embora muitos desses algoritmos sejam o estado da arte para diferentes problemas, há uma necessidade crescente de explicar as decisões tomadas. Essa necessidade, além de ser um desejo da sociedade, tornou-se recentemente um direito, garantido por lei.

Assim, o desenvolvimento de algoritmos que sejam capazes de explicar seu processo de tomada de decisão, ou de estratégias para tentar explicar as decisões tomadas por algoritmos de caixa-preta, tornou-se fundamental. Neste trabalho, comparamos o desempenho de modelos produzidos por um algoritmo interpretável com modelos considerados como caixas-pretas, produzidos por algoritmos de diferentes tipos. Além de comparar o desempenho entre os algoritmos, também foi medida a fidelidade (a capacidade do modelo interpretável de imitar parte do comportamento dos modelos de caixa-preta). Por fim, verificou-se que é possível utilizar modelos de caixa-preta para auxiliar no treinamento de modelos interpretáveis.

Em suma, verificou-se a esperada diferença de desempenho entre os modelos de caixa-preta e o modelo interpretável que, para muitos problemas, serve de justificativa para o uso de modelos de caixa-preta. Uma capacidade satisfatória por parte do modelo interpretável para imitar o comportamento do

modelo de caixa-preta também foi observada. A fidelidade, em média, foi maior do que a precisão do modelo interpretável para resolver o problema em questão.

Como limitações do presente trabalho, destacamos que os resultados e conclusões são baseados apenas nos materiais e métodos usados, ou seja, oito conjuntos de dados, um algoritmo usado para construir os modelos interpretáveis e oito modelos tratados como caixa-pretas, construídos a partir de setes algoritmos diferentes. Como indicações para trabalhos futuros, pretendemos expandir o estudo, incluindo mais conjuntos de dados, diferentes algoritmos interpretáveis e mais algoritmos de caixa-preta.

## 7. Referências


- [1] Favarò, F. M., Nader, N., Eurich, S., Tripp, M., Varadaraju, N.. (2017). Examining accident reports involving autonomous vehicles in California. PLoS ONE, v.12(9): e0184952
- [2] Buolamwini, J., Gebru, T. Friedler, S. A., Wilson, C. (Eds.). (2018). Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. Proceedings of the 1st Conference on Fairness, Accountability and Transparency, PMLR, v. 81, pp. 77-91.
- [3] Sina Mohseni, Niloofar Zarei, and Eric D. Ragan. 2020. A Multidisciplinary Survey and Framework for Design and Evaluation of Explainable AI Systems. https://doi.org/10.1145/3387166)
- [4] Papadopoulos, P., Walkinshaw, N.. (2015). Black-Box Test Generation from Inferred Models. Proceedings of the Fourth International Workshop on Realizing Artificial Intelligence Synergies in Software Engineering, IEEE Press, pp. 19-24.
- [5] Vieira, C., Digiampietri, L.. (2020). A study about Explainable Artificial Intelligence: using decision tree to explain SVM. Revista Brasileira de Computação Aplicada, v.12, pp. 113- 121
- [6] Nyholm, S., Smids, J.: The ethics of accident-algorithms for self-driving cars: an applied trolley problem? Ethical Theory and Moral Practice Article 19, 1275–1289 (2016). https://doi.org/10.1007/s10677-016-9745-2 1

- [7] Angwin, J., Larson, J., Mattu, S., Kirchner, L.: Machine Bias, p. 11. Auerbach Publications, New York, NY (2022), https://doi.org/10.1201/9781003278290
- [8] Coelho, J., Burg, T.: Uso de inteligência artificial pelo poder publico (2020)
- [9] Francisco, P.A., Hurel, L.M., Rielli, M.M.: Regulacao do reconhecimento facial no setor publico (2020), https://igarape.org.br/regulacao-do-reconhecimento-facialno-setor-publico/
- [10] Ramos, S.: Retratos da violência - cinco meses de monitoramento, análises e descobertas. Rede de Observatorios da Seguranca/CESeC 1(1), 1–72 (11 2019), https://cesecseguranca.com.br/textodownload/retratos-da-violencia-cincomeses-de-monitoramento-analises-e-descobertas/
- [11] Vieira, C.P., Digiampietri, L.A.: Machine learning post-hoc interpretability: A systematic mapping study. In: XVIII Brazilian Symposium on Information Systems. SBSI, Association for Computing Machinery, New York, NY, USA (2022). https://doi.org/10.1145/3535511.3535512
- [12] Ribeiro, M., Singh, S., Guestrin, C.: "Why should I trust you?": Explaining the predictions of any classifier. In: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations. pp. 97–101. Association for Computational Linguistics, San Diego, California (Jun 2016). https://doi.org/10.18653/v1/N16-3020, https://aclanthology.org/N16-3020
- [13] Byrne, R.M.J.: Counterfactuals in explainable artificial intelligence (XAI): Evidence from human reasoning. In: Proceedings of the Twenty-Eighth International Joint Conference on Artificial

Intelligence, IJCAI-19. pp. 6276–6282. International Joint Conferences on Artificial Intelligence Organization (7 2019). https://doi.org/10.24963/ijcai.2019/876

- [14] Aggarwal, A., Lohia, P., Nagar, S., Dey, K., Saha, D.: Black box fairness testing of machine learning models. In: Proceedings of the 2019 27th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering. p. 625–635. ESEC/FSE 2019, Association for Computing Machinery, New York, NY, USA (2019). https://doi.org/10.1145/3338906.3338937
- [15] Mitchell, S., Potash, E., Barocas, S., D'Amour, A., Lum, K.: Algorithmic fairness: Choices, assumptions, and definitions. Annual Review of Statistics and Its Application 8(1), 141–163 (2021). https://doi.org/10.1146/annurev-statistics-042720- 125902
- [16] Suresh, H., Guttag, J.: A framework for understanding sources of harm throughout the machine learning life cycle. In: Equity and Access in Algorithms, Mechanisms, and Optimization. EAAMO '21, Association for Computing Machinery, New York, NY, USA (2021)
- [17] Miller, T.: Explanation in artificial intelligence: Insights from the social sciences. Artificial Intelligence 267, 1–38 (2019). https://doi.org/https://doi.org/10.1016/j.artint.2018.07.007, https://www.sciencedirect.com/science/article/pii/S0004370218305988
- [18] Rudin, C., Radin, J.: Why Are We Using Black Box Models in AI When We Don't Need To? A Lesson From an Explainable AI Competition. Harvard Data Science Review 1(2) (nov 22 2019), https://hdsr.mitpress.mit.edu/pub/f9kuryi8