UNIVERSIDADE
FEDERAL RURAL
DE PERNAMBUCO

André Carlos Santos de Assis

Explainable Artificial Intelligence - Uma Análise dos *trade-offs* entre Desempenho e Explicabilidade

Recife

Setembro de 2023

André Carlos Santos de Assis

Explainable Artificial Intelligence - Uma Análise dos *trade-offs* entre Desempenho e Explicabilidade

Artigo apresentado ao Curso de Bacharelado em Sistemas de Informação da Universidade Federal Rural de Pernambuco, como requisito parcial para obtenção do título de Bacharel em Sistemas de Informação.

Universidade Federal Rural de Pernambuco – UFRPE
Departamento de Estatística e Informática
Curso de Bacharelado em Sistemas de Informação

Orientador: Dr. Ermeson Andrade
Coorientador: Dr. Douglas Véras

Recife
Setembro de 2023

Explainable Artificial Intelligence - Uma Análise dos *trade-offs* entre Desempenho e Explicabilidade

André Assis [1], **Ermeson Andrade** [2], **Douglas Véras** [2]

[1]Departamento de Estatística e Informática – Universidade Federal Rural de Pernambuco
Rua Dom Manuel de Medeiros, s/n, - CEP: 52171-900 – Recife – PE – Brasil

[2]Departamento de Computação – Universidade Federal Rural de Pernambuco
Rua Dom Manuel de Medeiros, s/n, - CEP: 52171-900 – Recife – PE – Brasil

{andre.assis, douglas.veras, ermeson.andrade}@ufrpe.br

Resumo. *A explicabilidade é essencial para que os usuários entendam, confiem e gerenciem com eficiência sistemas computacionais que utilizam inteligência artificial. Desta forma, assim como a assertividade, entender como se deu o processo decisório dos modelos é fundamental. Embora existam trabalhos que se concentrem na explicabilidade de algoritmos de inteligência artificial, é importante destacar que, até onde sabemos, nenhum deles analisou os trade-offs entre desempenho e explicabilidade de forma abrangente. Nesse sentido, esta pesquisa tem como objetivo preencher essa lacuna, investigando tanto algoritmos transparentes, como Árvore de Decisão e Regressão Logística, quanto algoritmos opacos, como Floresta Aleatória e Máquina de Vetores de Suporte, a fim de avaliar os trade-offs entre desempenho e explicabilidade. Os resultados revelam que os algoritmos opacos apresentam uma baixa explicabilidade e não têm uma boa performance quanto ao tempo de resposta devido à sua complexidade, contudo são mais assertivos. Em contra partida, os algoritmos transparentes possuem uma explicabilidade mais efetiva e uma melhor performance quanto ao tempo de resposta, porém, em nossos experimentos, observamos que a acurácia obtida foi menor do que a acurácia dos modelos opacos.*

Abstract. *Explainability is essential for users to efficiently understand, trust, and manage computer systems that use artificial intelligence. Thus, as well as assertiveness, understanding how the decision-making process of the models occurred is fundamental. While there are studies that focus on the explainability of artificial intelligence algorithms, it is important to highlight that, as far as we know, none of them have comprehensively analyzed the trade-offs between performance and explainability. In this sense, this research aims to fill this gap by investigating both transparent algorithms, such as Decision Tree and Logistic Regression, and opaque algorithms, such as Random Forest and Support Vector Machine, in order to evaluate the trade-offs between performance and explainability. The results reveal that opaque algorithms have a low explanability and do not perform well regarding response time due to their complexity, but are more assertive. On the other hand, transparent algorithms have a more effective explainability and better performance regarding response time, but in our experiments, we observed that accuracy obtained was lower than the accuracy of opaque models.*

1. Introdução

Com a ascensão da Tecnologia da Informação e Comunicação (TIC), grandes volumes de dados são gerados continuamente. Dessa forma, há uma grande necessidade de se obter modelos performáticos que consigam processar grandes volumes de dados. Além disso, os problemas do mundo real como previsão de eventos climáticos, análise do mercado financeiro, capacidade produtiva e distribuição de energia, dentre outros, exigem soluções cada vez mais robustas e complexas. Assim, os algoritmos de *Machine Learning* estão evoluindo cada vez mais para lidar com os novos desafios da sociedade e os modelos construídos através desses algoritmos possuem processos internos cada vez mais difíceis de serem compreendidos para os humanos. Explicações para as decisões e previsões dos modelos são, portanto, necessárias para justificar sua confiabilidade. Para isso, é fundamental que os algoritmos utilizados sejam interpretáveis, o que implica na necessidade de entender os mecanismos subjacentes aos mesmos. A *eXplainable Artificial Intelligence* (XAI) possui o propósito de tornar os resultados de sistemas de Inteligência Artificial (IA) mais compreensíveis aos humanos [Adabi and Berrada 2018].

O termo XAI foi usado inicialmente por Vant Lent em 2004 para descrever a capacidade de explicação do comportamento de entidades controladas por IA em aplicativos de jogos de simulação [Van Lent et al. 2004]. Apesar de estar em evidência nos últimos anos, o "problema de explicabilidade" pode ser encontrado desde a década de 1970, durante o estudo da explicação para sistemas especialistas [Swartout and Moore 1988]. No entanto, a motivação dos pesquisadores da época se debruçava muito mais na implementação dos algoritmos e tinha um foco maior no poder preditivo, enquanto a interpretação dos processos decisórios ficou em segundo plano. Atualmente, os algoritmos estão cada vez mais sofisticados, possuindo um bom desempenho quanto a acurácia e tempo de execução. Em contrapartida, devido à complexidade dos processos decisórios, sua explicação é comprometida, tornando difícil a interpretação dos modelos. Nesse contexto, surge a seguinte questão: "como confiar no resultado de um modelo se não é claro como o processamento interno foi realizado para chegar a tal resultado?"Esta indagação é uma das principais motivações para o estudo de técnicas de XAI.

Os esforços da XAI se concentram no desafio de desmistificar as "caixas pretas", mas além disso, a XAI se preocupa com a IA responsável, pois auxilia na compreensão dos modelos opacos. O objetivo é que isso aconteça sem afetar a precisão dos modelos, portanto, na inteligência artificial em geral e em aprendizagem de máquina especificamente, muitas vezes compensações devem ser feitas entre precisão e interpretabilidade. Diante desse contexto, se faz necessário analisar os *trade-offs* entre desempenho e explicabilidade a fim de selecionar os algoritmos que mais se adequem as necessidades das empresas. A presente pesquisa busca analisar os *trade-offs* através da comparação entre algoritmos opacos (baixa explicabilidade) e algoritmos transparentes (alta explicabilidade), avaliando suas respectivas performances focando na acurácia, tempo médio de resposta e explicabilidade. Mais especificamente, os algoritmos opacos adotados para nossas análises foram Floresta Aleatória e Máquina de Vetores de Suporte. Por outro lado, os algoritmos transparentes adotados foram Árvore de Decisão e Regressão Logística.

O restante deste artigo está organizado como segue. A Seção 2 apresenta a fundamentação teórica dos assuntos abordados, como a XAI e os algoritmos de *Machine Learning* utilizados. A Seção 3 apresenta os trabalhos relacionados. A Seção 4 apresenta

a arquitetura experimental, onde é explicado como foi criado e configurado o ambiente de experimentos. A Seção 5 discute os resultados obtidos. Por fim, a Seção 6 discorre as considerações finais e as perspectivas para os trabalhos futuros.

2. Fundamentação

A presente seção aborda a XAI, os algoritmos transparentes (Árvore de Decisão e Regressão Logística) e os algoritmos opacos (Floresta Aleatória e Máquina de Vetores de Suporte).

2.1. XAI

A XAI é uma área de pesquisa que dedica seus esforços a estudar e propor métodos para a construção de soluções com inteligência artificial que possam ser explicadas aos seres humanos [Van Lent et al. 2004]. Não existe uma definição matemática de explicabilidade (ou interpretabilidade) dos algoritmos de IA. Contudo, uma definição dada por Miller diz que "a interpretabilidade é o grau em que um humano pode entender a causa de uma decisão" [Miller 2019]. Quanto à explicabilidade no contexto de *Machine Learning*, Kim descreve como "o grau em que um ser humano pode prever consistentemente o resultado do modelo" [Kim et al. 2016], isto é, quanto mais fácil for para uma pessoa entender o modelo construído a partir de um algoritmo de IA e compreender os processos até o seu resultado final, mais explicável ele é. Analogamente, um modelo é mais explicável que outro modelo, se o processo decisório do primeiro for mais compreensível que o segundo.

Outra definição descreve a explicabilidade como a "capacidade de explicar ou apresentar em termos compreensíveis para um ser humano" [Doshi-Velez and Kim 2017], e trazendo para o contexto de *Machine Learning*, Molnar constata que "o aprendizado de máquina interpretável refere-se a métodos e modelos que tornam o comportamento e as previsões dos sistemas de aprendizado de máquina compreensíveis para os humanos" [Molnar 2020]. Logo, a explicabilidade está diretamente relacionada à capacidade humana de assimilar informações sobre determinado fluxo algorítmico. Desta forma, os modelos de aprendizado de máquina podem ser categorizados com base em seu nível de interpretabilidade, que pode ser definido como o grau em que um humano pode entender a causa de uma decisão ou ser capaz de reproduzir exatamente o que o modelo faz [Miller 2019].

Quanto a explicabilidade, os modelos podem ser divididos em dois grupos principais: transparentes e opacos. Os modelos considerados transparentes possuem estruturas simples e são mais fáceis de serem compreendidos, como por exemplo a Árvore de Decisão e Regressão Logística. O processo decisório desses modelos costumam ser transparentes, porém a transparência, como propriedade, não é suficiente para garantir que um modelo seja prontamente explicável [Angelov et al. 2021]. Os modelos opacos possuem esse nome por possuírem uma natureza "caixa preta" devido a complexidade do algoritmo, dessa forma dificultando a compreensão, como por exemplo a Floresta Aleatória e a Máquina de Vetores de Suporte (mais conhecida como *Support Vector Machine* ou SVM).

2.2. Árvore de Decisão

A Árvore de Decisão é um algoritmo de aprendizado de máquina supervisionado utilizado para classificação e regressão. Ou seja, pode ser usado para previsão de

variáveis categóricas discretas, como por exemplo “sim” ou “não”, e valores numéricos [Quinlan 1996]. Como o próprio nome sugere, a Árvore de Decisão possui uma hierarquia de nós que se relacionam entre si. Os nós armazenam um conjunto de informações, o primeiro nó é chamado de “nó raiz”que é o ponto de partida, os nós intermediários são as possíveis decisões e são chamados de “ramos”, os últimos nós nos quais estão os resultados finais são chamados de “nós-folha”[Quinlan 1987]. A hierarquia da Árvore de Decisão possui um conjunto de instruções de controle condicional, onde os nós intermediários representam decisões e os nós folhas podem ser rótulos de classe (classificação) ou quantidades contínuas (regressão). De um modo geral, o modelo é considerado transparente, o que significa que é fácil entender como ele faz suas previsões com base nas características dos dados de entrada. No entanto, esse modelo pode ser limitado em termos de precisão preditiva, uma vez que pode tender a superajustar os dados de treinamento, o que dificulta a generalização para novos dados.

2.3. Regressão Logística

Este modelo estatístico é utilizado para determinar a probabilidade de um evento acontecer. Também é um modelo de aprendizado de máquina supervisionado, que além de criar previsões precisas, é utilizado para mostrar a relação entre as variáveis. Assim como a Árvore de Decisão, a Regressão Logística é usada para previsão de variáveis categóricas discretas. A Regressão Logística tem o objetivo de entender a relação entre variáveis independentes e dependentes, e assim, construir um modelo com as associações. Com esse modelo é possível prever o valor que a variável dependente vai ter de acordo com os valores das variáveis independentes, podendo assim, atuar como classificador [Oliveira 2016]. As variáveis dependentes podem ser categóricas (desde que sejam dicotomizadas após transformação) ou contínuas [Gonçalves 2005]. No entanto, para que os modelos mantenham suas características de transparência, seu tamanho deve ser limitado e as variáveis utilizadas devem ser compreensíveis por seus usuários.

2.4. Floresta Aleatória

Com características semelhantes à Árvore de Decisão, a Floresta Aleatória é um algoritmo de aprendizado de máquina supervisionado usado para realizar predições. Esse algoritmo gera aleatoriamente Árvores de Decisão (*ensemble*) mesclando os resultados dessas árvores para uma previsão mais precisa. Uma das principais diferenças é que, enquanto uma árvore de decisão tem a finalidade de construir uma estrutura completa a partir de um conjunto de dados, a Floresta Aleatória tem a finalidade de criar várias árvores de decisão baseada em um subconjunto de atributos selecionados aleatoriamente a partir do conjunto original [Teloken et al. 2016]. Tal modelo foi definido por Tim Kam Ho, em 1995, no artigo *“Random Decision Forests”* e pode ser usado tanto para classificação quanto para regressão. De maneira geral, as Florestas Aleatórias apresentam um bom desempenho em termos de acurácia na tarefa de prever valores de saída a partir de um conjunto de entradas. No entanto, a sua complexidade e natureza aleatória podem dificultar a interpretação dos resultados, prejudicando a sua explicabilidade.

2.5. Máquina de Vetores de Suporte

A Máquina de Vetores de Suporte (SVM, do inglês *Support Vector Machine*) é um algoritmo de aprendizagem de máquina supervisionado que pode ser usado tanto para regressão quanto para classificação. Em termos gerais, uma SVM funciona da seguinte

forma: dada duas classes e um conjunto de pontos dessas classes, uma SVM busca determinar o hiperplano que separa os pontos, de modo que seja colocado o maior número de pontos da mesma classe do mesmo lado, enquanto maximiza a distância entre cada classe e o hiperplano. A margem de separação da SVM é a distância entre a superfície de decisão (ou hiperplano) e os pontos de dados mais próximos de cada classe. A SVM busca maximizar a margem de separação entre as classes, pois quanto maior a margem, maior a capacidade de generalização do modelo. A margem também é uma importante medida que ajuda a controlar o sobreajuste (*overfitting*) do modelo aos dados de treinamento, o que pode levar a uma performance pior em novos dados. O hiperplano criado pela SVM é determinado através de um subconjunto dos pontos das duas classes, denominados de vetores de suporte [Gevert et al. 2010]. Sua alta dimensionalidade e explicação geométrica torna os modelos complexos e opacos.

3. Trabalhos Relacionados

O estudo da XAI é relativamente novo e tem atraído a atenção dos pesquisadores. Em [Angelov et al. 2021], é introduzido uma revisão do estado da arte sobre a explicabilidade da inteligência artificial no contexto de aprendizagem profunda, abordando a taxonomia e apresentando os principais desafios com relação a explicabilidade baseado nos princípios do *National Institute of Standards* em relação à explicação, significado, precisão e limites do conhecimento. Uma das principais contribuições do artigo é a classificação dos autores das técnicas de XAI em cinco categorias: explicações *post-hoc*, explicações intrínsecas, explicações transparentes, explicações interativas e explicações híbridas. Esse esquema de classificação fornece uma estrutura útil para entender as diferentes abordagens da XAI e as compensações envolvidas na seleção de uma técnica apropriada para uma determinada aplicação.

Semelhantemente, [Vilone and Longo 2021] apresentaram uma revisão sistemática agrupando os estudos científicos através de um sistema hierárquico que classifica teorias e noções relacionadas a abordagens de avaliação para métodos de XAI. As análises identificaram requisitos que uma explicação deve conter para que seja facilmente compreendida aos usuários finais e também sugeriram abordagens para avaliar o grau das explicações. Os autores identificam três dimensões principais de explicabilidade: transparência, interpretabilidade e causalidade. Transparência refere-se ao grau em que o processo de tomada de decisão de um sistema de IA é visível e compreensível para usuários humanos. A interpretabilidade refere-se à capacidade dos usuários humanos de entender o raciocínio por trás das decisões do sistema. A causalidade refere-se ao grau em que as decisões do sistema podem ser rastreadas até entradas ou fatores específicos.

Por outro lado, [Confalonieri et al. 2021] fizeram uma análise através de uma perspectiva histórica fazendo um contraponto entre a forma como a XAI foi concebida e a forma que é compreendida atualmente e quais são as perspectivas futuras. O trabalho apresenta a história da XAI, começando com as primeiras tentativas de explicar o comportamento dos sistemas especialistas nas décadas de 1970 e 1980. Os autores descrevem como o foco mudou de sistemas baseados em regras para algoritmos de aprendizado de máquina, que muitas vezes são considerados "caixas pretas" devido ao seu complexo funcionamento interno. Eles destacam marcos importantes no desenvolvimento da XAI, como a introdução de árvores de decisão e classificadores baseados em regras na década de 1990 e o surgimento de métodos agnósticos de modelo, como *Local Interpre-*

table Model-agnostic Explanations (LIME) e *SHapley Additive exPlanations* (SHAP) na década de 2010. Também propõe critérios para explicações e apresenta a necessidade do desenvolvimento de sistemas explicáveis aos humanos.

Em [Islam et al. 2022], primeiramente, é introduzido o conceito de XAI, que se refere à capacidade dos sistemas de IA de fornecer explicações transparentes e compreensíveis para suas decisões e ações. Eles discutem a importância da XAI em vários domínios de aplicação, incluindo saúde, finanças e veículos autônomos. Em seguida, o artigo apresenta uma análise detalhada das diferentes abordagens e técnicas utilizadas para XAI, como sistemas baseados em regras, árvores de decisão e redes neurais profundas. Os autores também discutem os desafios e limitações dessas técnicas, como os *trade-offs* entre precisão e interpretabilidade. Também fornece uma visão abrangente do estado da arte sobre a XAI e destaca a importância da transparência e interpretabilidade em sistemas de IA para garantir seu uso ético e responsável em vários domínios de aplicação.

Em [Samek and Müller 2019], é discutido a importância da interpretabilidade e explicabilidade na IA, especialmente em aplicações críticas como a medicina. Eles argumentam que a falta de interpretabilidade é um problema significativo para as técnicas de aprendizado profundo, que são frequentemente usadas em sistemas de IA. Os modelos de aprendizagem profunda podem ser extremamente precisos em prever resultados, mas sua estrutura interna é frequentemente difícil de compreender e explicar. O artigo também explora a ideia de que a interpretabilidade é um processo contínuo e dinâmico, de modo que as tecnologias para explicar modelos de IA devem estar em constante evolução. Os autores enfatizam a importância da colaboração entre especialistas em IA e usuários finais para criar modelos que sejam precisos e interpretáveis, concluindo que a interpretabilidade deve ser considerada uma característica essencial de qualquer sistema de IA que se destina a ser usado em aplicações críticas.

Na literatura, também encontramos uma série de estudos que exploram o uso da proveniência como metadados para fornecer explicações sobre o processamento de dados, identificar as vantagens dessas explicações e determinar em quais configurações elas são úteis. Por exemplo, em [Wu et al. 2020], foi desenvolvida uma abordagem baseada em proveniência para a atualização incremental de parâmetros em modelos de aprendizado de máquina. O objetivo dessa abordagem é auxiliar nas atividades de depuração e melhoria desses modelos. Os resultados obtidos demonstraram a eficácia da abordagem PrIU, assim como sua versão otimizada, sendo comprovada sua correção, convergência e validação por meio de experimentos. Outro estudo relevante é o trabalho de [Grafberger et al. 2023], que visa fornecer suporte automatizado e de baixo esforço para *pipelines* de aprendizado de máquina em cenários do mundo real. Nesse contexto, uma abordagem baseada em proveniência foi proposta, utilizando registros de entrada para calcular uma saída específica. Além disso, foram desenvolvidos protótipos para extrair resultados intermediários e identificar problemas comuns, contribuindo para a reconstrução dos modelos.

Este trabalho apresenta algumas semelhanças e diferenças com os trabalhos acima citados. Embora os trabalhos anteriormente listados tenham realizado análises de modelos opacos e transparentes e desenvolvido técnicas de XAI, é importante observar que nenhum deles investigou de forma abrangente a relação entre tempo de resposta e acurácia nos algoritmos opacos e transparentes. Além disso, esses estudos não consideraram diferentes

fatores na arquitetura experimental, como diferentes cargas de trabalho. O diferencial da presente pesquisa é comparar algoritmos transparentes (Árvore de Decisão e Regressão Logística) e opacos (Floresta Aleatória e Máquinas de Vetores de Suporte), com a finalidade de analisar *trade-offs* entre desempenho e explicabilidade através da arquitetura cliente e servidor. Além disso, espera-se que o trabalho permita que desenvolvedores possam avaliar de forma mais precisa as vantagens e desvantagens de cada modelo em relação à explicabilidade e desempenho, possibilitando a seleção do modelo mais adequado para cada situação.

4. Plano Experimental

Para analisar os *trade-offs* entre desempenho e explicabilidade, utilizamos a base de dados MNIST *(Modified National Institute of Standards and Technology)* disponível no *Tensorflow* que é uma biblioteca de código aberto para aprendizado de máquina que pode ser utilizada em uma variedade de tarefas [Abadi et al. 2015]. A base de dados MNIST é um vasto conjunto de dígitos manuscritos que possui uma coleção de dados de treinamento contendo 60.000 exemplos e uma coleção de dados de teste contendo 10.000 exemplos. O MNIST é um subconjunto do *NIST Special Database 3* (dígitos escritos por funcionários do United States Census Bureau) e *Special Database 1* (dígitos escritos por alunos do ensino médio) que contém imagens monocromáticas de dígitos manuscritos [LeCun et al. 2010].

Nos experimentos, selecionamos dois algoritmos transparentes (Árvore de Decisão e Regressão Logística) e dois algoritmos opacos (Floresta Aleatória e Máquina de Vetores de Suporte). Na construção dos modelos utilizamos o *RandomizedSearchCV*, disponível no *scikit-learn* [Pedregosa et al. 2011], para selecionar os melhores parâmetros. Para avaliar a assertividade utilizamos a acurácia e para analisar o desempenho utilizamos o tempo médio de resposta de cada modelo. No processo de envio das imagens utilizamos a variação da carga em 1.0, 0.5 e 0.1, sendo classificadas como baixa, média e alta, respectivamente. Os valores da carga correspondem ao intervalo de tempo em segundos entre os envios de imagens do cliente para o servidor. Dessa forma, realizamos o envio de mil imagens em cada carga para cada modelo.

4.1. Ambiente de testes

Para realizar os testes de performance, foi montada uma estrutura com dois computadores, sendo um cliente e um servidor com os modelos treinados implantados, ambos conectados a mesma rede local via WI-FI. Nessa estrutura, o cliente envia as imagens via *socket* para o servidor e este prontamente responte tais solicitações, sendo possível assim, calcular o tempo médio de resposta de acordo com a variação da carga. As configurações do servidor e do cliente estão detalhadas na Tabela 1.

4.2. Execução de Experimentos

Antes de realizar o treinamento dos modelos, utilizamos a biblioteca *RandomizedSearchCV* para a seleção dos hiperparâmetros. Essa biblioteca realiza treinamentos combinando os possíveis parâmetros e retorna o conjunto de hiperparâmetros que obteve a melhor avaliação. A principal vantagem do *RandomizedSearchCV* é que ele pode economizar tempo e recursos computacionais, pois testa apenas uma amostra aleatória de combinações de hiperparâmetros. Desta forma, para cada modelo realizamos testes com

Tabela 1. Configuração das Plataformas Adotadas

Configuração	Servidor	Cliente
CPU	Intel(R) Core(TM) i5-8500T	Intel(R) Core(TM) i7-8565U
Memória RAM	24 Gb DDR4 2666 Mhz	12 Gb DDR4 2400 Mhz
Armazenamento em Disco	500 Gb SSD	500 Gb SSD
GPU	Intel(R) UHD Graphics 630	Intel(R) UHD Graphics 620

o objetivo de obter o melhor conjunto de parâmetros que obteve a melhor acurácia. A Tabela 2 apresenta os hiperparâmetros selecionados através do *RandomizedSearchCV*.

Tabela 2. Hiperparâmetros selecionados através do *RandomizedSearchCV*

Algoritmos	Hiperparâmetros
Árvore de Decisão	*criterion: gini, max_depth: None*
Regressão Logística	*penalty: l2, dual: False*
Floresta Aleatória	*criterion: gini, max_depth: None*
Máquina de Vetores de Suporte	*kernel: rbf, degree: 3, gamma: scale*

Uma vez realizado o treinamento dos modelos, o experimento foi realizado. Tal experimento consistiu em enviar as imagens do cliente para o servidor variando a carga em 1.0, 0.5 e 0.1 (baixa, média e alta, nesta ordem). Com a execução do teste coletamos a acurácia dos modelos e o tempo de resposta das requisições. Vale salientar que o tempo de resposta é a quantidade de tempo que leva para uma determinada operação ou processo ser concluído, desde o momento em que a solicitação é feita até o momento em que a resposta é recebida, ou seja, é uma métrica de desempenho responsável por informar o tempo que um sistema leva para dar o retorno desejado para o usuário.

4.3. Análises

Para avaliar a assertividade dos modelos, utilizamos a acurácia que é uma métrica usada para avaliar o desempenho de modelos de *Machine Learning*. Esta métrica mede a proporção de predições corretas que o modelo faz em relação ao total de predições. Como a base de dados MNIST é um conjunto de dados equilibrado, ou seja, as classes estão distribuídas de maneira uniforme, a acurácia é uma boa métrica de avaliação para esse experimento. Para obter a acurácia dos modelos utilizamos a biblioteca *sklearn.metrics* que é uma função da biblioteca *scikit-learn* em *Python* que calcula métricas de avaliação de modelos de *Machine Learning*.

Com a finalidade de mensurar o desempenho dos modelos, calculamos as médias dos tempos de respostas de cada modelo de acordo com a variação das cargas. O tempo médio de resposta pode variar dependendo da complexidade do modelo, da entrada fornecida e da capacidade de processamento do servidor. Calcular o tempo de resposta é uma parte importante da avaliação do desempenho do modelo, da garantia da eficiência e da usabilidade de uma aplicação, principalmente para garantir que os modelos sejam eficientes, escaláveis e capazes de gerar resultados em tempo real.

5. Resultados

Nesta seção, serão apresentados os resultados que foram obtidos através da execução dos experimentos. Após a realização do envios das imagens do cliente para o servidor com as variações de cargas mencionada na seção anterior, coletamos o tempo de resposta de cada modelo e suas respectivas acurácias.

Com relação a assertividade, os algoritmos transparentes Árvore de Decisão e Regressão Logística obtiveram uma acurácia de 0.83 e 0.93 respectivamente, enquanto os algoritmos opacos Floresta Aleatória e Máquina de Vetores de Suporte atingiram uma acurácia de 0.97 e 0.98, nesta ordem. Desta forma, neste experimento os algoritmos opacos foram mais assertivos que algoritmos transparentes, como é apresentado na Tabela 3. Analisando as acurácias percebe-se que a Floresta Aleatória obteve um desempenho melhor do que a Árvore de Decisão e isso ocorre porque a Floresta Aleatória é menos propensa a *overfitting* e pode produzir resultados mais precisos e robustos, graças à agregação de previsões de várias árvores de decisão construídas a partir de amostras e subconjuntos aleatórios dos dados. O algoritmo Máquina de Vetores de Suporte obteve a melhor acurácia, pois o conjunto de dados MNIST é bastante robusto e linearmente separável. A grande quantidade de variáveis preditoras favoreceu o algoritmo Regressão Logística, porém devido a robustez do conjunto de dados, teve sua acurácia comprometida, pois este é preferível quando o conjunto de dados é pequeno.

Tabela 3. Resultado Geral dos Modelos

Modelo	Acurácia	Explicabilidade
Árvore de Decisão	0.83	Transparente
Regressão Logística	0.93	Transparente
Floresta Aleatória	0.97	Opaco
Máquina de Vetores de Suporte	0.98	Opaco

A Figura 1 apresenta os resultados obtidos nesse experimento referente ao tempo médio de resposta de cada modelo de acordo com as cargas adotadas. O eixo x apresenta as cargas agrupadas por modelo e o eixo y o tempo médio de resposta em segundos. Podemos notar que o tempo médio de resposta cresce conforme o aumento da carga em todos os modelos. Esse aumento é mais expressivo nos algoritmos opacos (Floresta Aleatória e Máquina de Vetores de Suporte). Os algoritmos transparentes (Árvore de Decisão e Regressão Logística) se mantiveram mais estáveis, com um tempo médio de resposta bastante próximo na carga baixa e uma diferença pouco mais significante nas cargas média e alta. Vale ressaltar a diferença expressiva entre os algoritmos opacos e transparentes na carga alta. Enquanto a Árvore de Decisão atingiu 0.261383 segundos e a Regressão Logística 0.238978 segundos, a Floresta Aleatória alcançou um tempo médio de 0.414797 segundos e a Máquina de Vetores de Suporte 0.383682 segundos.

Para analisar se as amostras são estatisticamente diferentes, utilizamos o teste de Kruskal-Wallis [Elliott and Hynan 2011] e o teste post-hoc de Dunn [McKight and Najab 2010]. O teste de Kruskal-Wallis é realizado inicialmente para determinar se há diferenças significativas entre os grupos e o teste Post-Hoc de Dunn é aplicado para comparar os grupos dois a dois e obter diferenças significativas. Realizamos os testes para as três cargas (baixa, média e alta). Os resultados do teste de Kruskal-Wallis

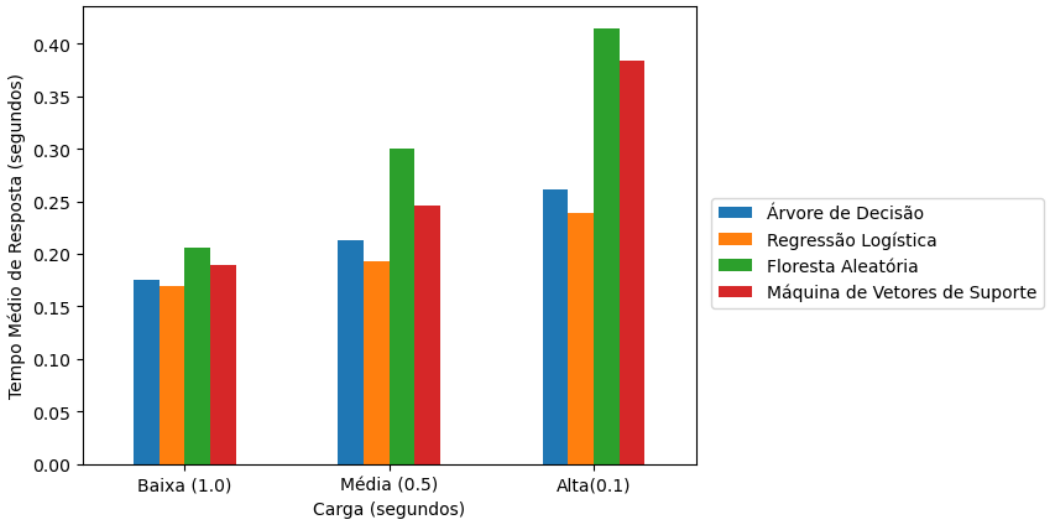

Figura 1. Tempo Médio de Resposta por Carga

Fonte: Elaborado pelos autores.

indicam que há diferenças significativas dentro dos grupos, sugerindo que pelo menos um par de grupos é estatisticamente diferente. Os resultados do teste Post-Hoc de Dunn mostram através de uma matriz de comparações múltiplas entre os grupos que todos os pares de grupos têm diferenças estatisticamente significativas com base nas análises realizadas. Dessa forma, há evidências estatísticas suficientes para afirmar que as distribuições dos valores nos diferentes grupos são diferentes entre si.

Embora o algoritmo Máquina de Vetores de Suporte tenha produzidos resultados altamente precisos, a lógica por trás de como ele chega a esses resultados pode ser difícil de entender, pois trabalha em um espaço de alta dimensionalidade e, portanto, a interpretação dos pesos e características de entrada pode não ser intuitiva para um ser humano. A Floresta Aleatória também obteve uma boa acurácia, porém é considerada um modelo de aprendizado de máquina opaco ou de "caixa preta" em relação à interpretação de como exatamente ele toma suas decisões. Esse modelo combina várias árvores de decisão individuais para criar uma previsão final. Cada árvore é construída usando um subconjunto aleatório de recursos e amostras de dados, o que torna difícil interpretar as decisões do modelo como um todo.

A Regressão Logística obteve o melhor tempo médio de resposta entre os demais algoritmos utilizados nesse experimento e a terceira melhor acurácia. Por ser considerado um algoritmo linear, é relativamente fácil interpretar a relação entre as variáveis de entrada e a saída. Isso torna o modelo bastante transparente e fácil de interpretar, especialmente em comparação com modelos opacos. No entanto, é importante notar que a transparência da Regressão Logística depende da qualidade dos dados e da seleção das variáveis de entrada. Já a Árvore de Decisão, atingiu a pior acurácia entre os algoritmos utilizados, contudo dependendo da dimensão da árvore gerada, é possível facilmente entender como as decisões são tomadas e quais variáveis são mais importantes na previsão do resultado.

De maneira geral, devido a sua complexidade, os modelos construídos a partir de algoritmos opacos alcançaram uma acurácia maior do que os modelos transparentes.

Em contrapartida, os modelos transparentes atingiram um tempo de resposta menor que os modelos opacos em todas as variações de carga. Desse modo, é possível destacar a existência de um trade-off entre o desempenho (acurácia e tempo de resposta) e a explicabilidade dos modelos. Modelos que utilizam algoritmos opacos tendem a ter uma maior assertividade, porém apresentam um tempo de resposta mais longo e um processo decisório de baixa explicabilidade devido à sua complexidade. Contudo, os algoritmos transparentes possuem características opostas aos algoritmos opacos. Eles podem apresentar uma menor assertividade em comparação aos algoritmos opacos, porém, têm um tempo médio de resposta mais rápido. Além disso, devido à sua transparência, esses algoritmos oferecem um nível mais elevado de explicabilidade.

Os *trade-offs* entre desempenho e explicabilidade é uma questão importante nos algoritmos de aprendizagem de máquina. De um lado, temos a busca pelo melhor desempenho possível (acurácia e tempo de resposta) nos modelos, ou seja, a capacidade de fazer previsões precisas e acuradas em um período de tempo reduzido. Por outro lado, há a necessidade de entender como essas previsões são feitas, ou seja, a capacidade de explicar os porquês das decisões tomadas. Os algoritmos opacos de maneira geral podem oferecer uma alta precisão, mas são difíceis de explicar, pois envolvem um grande número de parâmetros e barreiras que podem obscurecer a compreensão da tomada de decisão. Os algoritmos transparentes, no entanto, são mais fáceis de explicar, mas podem não alcançar o mesmo nível de precisão que os modelos mais complexos. Além disso, em nosso experimento os algoritmos transparentes apresentaram um tempo de resposta mais rápido do que os algoritmos opacos. Assim, escolher entre desempenho e explicabilidade depende das necessidades e objetivos específicos do projeto. Se a transparência e interpretabilidade do modelo são críticas, pode ser mais adequado usar um modelo mais simples e fácil de entender. Todavia, se a precisão da previsão é mais importante, pode ser necessário sacrificar a explicabilidade em favor de um modelo mais complexo.

6. Conclusões

A *eXplainable Artificial Intelligence* é uma área de estudo promissora para melhorar a confiabilidade dos sistemas computacionais, que busca desenvolver técnicas e abordagens para tornar as decisões tomadas por algoritmos de IA mais transparentes e compreensíveis. Embora a IA seja capaz de aprender e tomar decisões de maneira eficaz, ela é frequentemente vista como uma "caixa preta" devido à complexidade dos modelos e da opacidade de suas decisões. Assim, se faz necessário que os usuários possam entender como as decisões são tomadas pelos sistemas e quais são os fatores que influenciam essas decisões.

Este trabalho analisou algoritmos transparentes (Árvore de Decisão e Regressão Logística) e opacos (Floresta Aleatória e Máquinas de Vetores de Suporte) visando avaliar os *trade-offs* entre desempenho (acurácia e tempo de resposta) e explicabilidade, utilizando a base de dados MNIST. Os modelos transparentes avaliados possuem uma melhor explicabilidade e um tempo médio de resposta mais rápido, porém possuem uma acurácia menor. Em contrapartida, os modelos opacos possuem uma alta acurácia, mas por serem inerentemente opacos, devido a sua natureza "caixa preta", não possuem uma boa explicabilidade e, em todas as cargas adotadas, obtiveram um tempo médio de resposta maior do que os modelos transparentes. Desta forma, se a explicabilidade e tempo de resposta forem essenciais, de maneira geral, os modelos transparentes serão mais recomendados.

Contudo, se a necessidade for assertividade, os modelos opacos terão um melhor desempenho. Com objetivo de unir assertividade e explicabilidade, em nosso trabalho futuro, planejamos utilizar algoritmos de *Deep Learning* como *Convolutional Neural Network* (CNN) e outras bases de dados.

Referências

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viégas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., and Zheng, X. (2015). TensorFlow: Large-scale machine learning on heterogeneous systems. Software available from tensorflow.org.

Adabi, A. and Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence. *IEEE Access*, 6:52138–52160.

Angelov, P. P., Soares, E. A., Jiang, R., Arnold, N. I., and Atkinson, P. M. (2021). Explainable artificial intelligence: an analytical review. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 11(5):e1424.

Confalonieri, R., Coba, L., Wagner, B., and Besold, T. R. (2021). A historical perspective of explainable artificial intelligence. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 11(1):e1391.

Doshi-Velez, F. and Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.

Elliott, A. C. and Hynan, L. S. (2011). A sas® macro implementation of a multiple comparison post hoc test for a kruskal–wallis analysis. *Computer methods and programs in biomedicine*, 102(1):75–80.

Gevert, V. G., da Silva, A. C. L., Gevert, F., and Ales, V. T. (2010). Modelos de regressão logística, redes neurais e support vector machine (svm s) na análise de crédito a pessoas jurídicas. *RECEN-Revista Ciências Exatas e Naturais*, 12(2):269–293.

Gonçalves, E. B. (2005). *Análise de risco de crédito com o uso de modelos de regressão logística, redes neurais e algoritmos genéticos*. PhD thesis, Universidade de São Paulo.

Grafberger, S., Groth, P., and Schelter, S. (2023). Provenance tracking for end-to-end machine learning pipelines. In *Companion Proceedings of the ACM Web Conference 2023*, pages 1512–1512.

Islam, M. R., Ahmed, M. U., Barua, S., and Begum, S. (2022). A systematic review of explainable artificial intelligence in terms of different application domains and tasks. *Applied Sciences*, 12(3):1353.

Kim, B., Khanna, R., and Koyejo, O. O. (2016). Examples are not enough, learn to criticize! criticism for interpretability. *Advances in neural information processing systems*, 29.

LeCun, Y., Cortes, C., and Burges, C. (2010). Mnist handwritten digit database. *ATT Labs [Online]. Available: http://yann.lecun.com/exdb/mnist*, 2.

McKight, P. E. and Najab, J. (2010). Kruskal-wallis test. *The corsini encyclopedia of psychology*, pages 1–1.

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence*, 267:1–38.

Molnar, C. (2020). *Interpretable machine learning*. Lulu. com.

Oliveira, P. H. M. A. (2016). *Detecção de fraudes em cartões: um classificador baseado em regras de associação e regressão logística*. PhD thesis, Universidade de São Paulo.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Quinlan, J. R. (1987). Simplifying decision trees. *International journal of man-machine studies*, 27(3):221–234.

Quinlan, J. R. (1996). Learning decision tree classifiers. *ACM Computing Surveys (CSUR)*, 28(1):71–72.

Samek, W. and Müller, K.-R. (2019). Towards explainable artificial intelligence. *Explainable AI: interpreting, explaining and visualizing deep learning*, pages 5–22.

Swartout, W. and Moore, J. (1988). Explanation in expert systems: A survey. *University of southern California*.

Teloken, A. et al. (2016). Estudo comparativo entre os algoritmos de mineração de dados random forest e j48 na tomada de decisão. *Simpósio de Pesquisa e Desenvolvimento em Computação*, 2(1).

Van Lent, M., Fisher, W., and Mancuso, M. (2004). An explainable artificial intelligence system for small-unit tactical behavior. In *Proceedings of the national conference on artificial intelligence*, pages 900–907. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999.

Vilone, G. and Longo, L. (2021). Notions of explainability and evaluation approaches for explainable artificial intelligence. *Information Fusion*, 76:89–106.

Wu, Y., Tannen, V., and Davidson, S. B. (2020). Priu: A provenance-based approach for incrementally updating regression models. In *Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data*, pages 447–462.