Revista Brasileira de Computação Aplicada, July, 2025

DOI: 10.5335/rbca.v17i2.16459
Vol. 17, Nº 2, pp. 46–53
Homepage: seer.upf.br/index.php/rbca/index

ORIGINAL PAPER

Explaining the black-box or using the black-box to develop better interpretable solutions?

Vinicius Alves Matias [1], Julia Machado Lechi[1], Norton Trevisan Roman [1], Luciano Antonio Digiampietri [1]

[1]School of Arts, Sciences and Humanities, University of São Paulo

*viniciusmatias@usp.br; julia.lechi@usp.br; norton@usp.br; digiampietri@usp.br

Received: 2024-11-07. **Revised:** 2025-07-24. **Accepted:** 2025-07-31.

Abstract

Understanding the decision-making processes behind Artificial Intelligence models became a crucial aspect of AI. This paper describes a study that compares the performance of models produced by both interpretable and black-box algorithms and evaluates if it is possible to use black-box models to assist in interpretable models' training. We verified a significant difference in performance between the two types of models. However, the interpretable model was able to mimic the behavior of the black-box models to a satisfactory degree. The promising initial results obtained from using black-box models to aid in interpretable models' training suggest the potential efficacy of this approach.

Keywords: Black Box Models; Explainable AI; Interpretable Models; Post-hoc Interpretability.

Resumo

Compreender os processos de tomada de decisão por trás dos modelos de Inteligência Artificial tornou-se um aspecto crucial da IA. Este artigo descreve um estudo que compara o desempenho de modelos produzidos por algoritmos interpretáveis e de caixa-preta, avaliando se é possível utilizar modelos de caixa-preta para auxiliar no treinamento de modelos interpretáveis. Verificamos uma diferença significativa de desempenho entre os dois tipos de modelos. No entanto, o modelo interpretável foi capaz de imitar o comportamento dos modelos de caixa-preta de maneira satisfatória. Os resultados iniciais promissores obtidos ao usar modelos de caixa-preta para auxiliar no treinamento de modelos interpretáveis sugerem a potencial eficácia dessa abordagem.

Palavras-Chave: IA Explicável; Interpretabilidade Post-hoc; Modelos Caixa-Preta; Modelos Interpretáveis.

1 Introduction

In recent decades, we have experienced a great advance in the development of artificial intelligence. More and more sophisticated models were developed and reached the state-of-the-art in solving different problems. As the complexity of these models increased, they also became increasingly opaque, i.e., it is increasingly difficult (and often even impossible) for a human being to understand the reason or logic that led the algorithm to produce its output.

With Artificial Intelligence increasingly present in people's daily lives in various areas and the advancement of the discussion on the protection of personal data and the right to explanation, an area of Artificial Intelligence (AI) that is receiving considerable attention is Explainable Artificial Intelligence (XAI). This area has several subfields, which share the objective of developing solutions to allow the results of AI algorithms to be understood by human beings (Miller, 2019; Mohseni et al., 2021). Among the main existing approaches to this end are (i) the production of inherently explainable models, (ii) the explanation of opaque models or those considered black-boxes, and (iii) the presentation of some examples to the user to enable them to understand the context of a result, including counterfactual examples.

In the context of opaque algorithms, much is being discussed about their real need, especially in problems that require explanations. One of the main arguments in this discussion is that it is impossible to effectively explain these models (Lipton, 2018; Rudin and Radin, 2019), either based on simpler models or through the identification of the importance of some input attributes for the constructed model. The present work aims to contribute to this discussion, starting with the following research questions:

RQ1: How faithful an interpretable model can be to an opaque counterpart?

RQ2: Is there a performance difference between models produced by an interpretable algorithm and those produced by algorithms considered black-boxes?

RQ3: Is it possible to construct better explainable models based on the outputs produced by a black-box algorithm?

These questions will be answered in the specific context addressed in the present work in which a set of public datasets was chosen for the execution of the tests, a decision tree algorithm was adopted as an inherently interpretable algorithm, configured to produce trees of maximum height equal to three, and a set of Supervised Learning algorithms of different natures was selected to represent the black-box algorithms.

The rest of this article is organized as follows. Section 2 presents the main concepts used in this work, as well as the most relevant related work. Section 3 describes the materials and methods employed in the experiments. The results are presented in Section 4, whereas Section 5 contains our conclusions and directions for future work.

2 Related Work

In recent years, the creation of different regulations related to the protection of users' personal data and the right of explanation (such as GDPR (Council of the European Union, 2018) in Europe and LGPD (Brasil, 2018) in Brazil) has intensified the discussion about the use and impact of Artificial Intelligence algorithms in society. In particular, the right to explanation raised questions about the feasibility and legality of using some algorithms that affect people's daily lives.

In some areas, such as medicine, it has always been questioned how computer-generated diagnoses could be used. In the autonomous car development area (Nyholm and Smids, 2016), this subject is also widely discussed, including legal responsibility in the case of accidents caused by design or implementation problems. In addition to these specific cases, where human lives may be directly at risk due to the action of AI algorithms, there are several other activities that directly affect people's lives. For example, in the financial area, there are several companies that use algorithms to assist in the process of granting or not granting loans. There are also companies that use AI models in their human resources sectors. Moreover, different security and public services use different AI models, including facial recognition algorithms, as a part of their procedures (Angwin et al., 2022; Coelho and Burg, 2020; Francisco et al., 2020; Ramos, 2019).

In terms of machine learning models, two main characteristics are often used to indicate how interpretable a model is: (i) the intrinsic nature of how "knowledge" is represented in the model and (ii) the size of the model. For example, in a linear regression, the coefficients related to each attribute or feature are learned; in a decision tree, the decision nodes are learned; whereas in a neural network, the weights of the links are learned. The meaning of a coefficient value in a linear regression or a decision node in a decision tree is often considered easier to understand than the weights in a Deep Neural Network. On the other hand, the size of the model is also important. Understanding the importance and meaning of the coefficients in a linear regression that uses three attributes is considered easier than in a regression that uses one hundred of them, just as a decision tree of height three is more "interpretable" than one of height 20.

Models whose interpretation is not simple are usually considered opaque, even for those who understand the logic behind the construction of these models. For example, deep neural networks or models produced from large language models are often considered opaque. It is noteworthy that the opacity can be attributed directly based on the type of algorithm that produced the model or considering the attributes that were used. For example, if the attributes used to build the model were generated from a projection process (for example, using principal component analysis), the resulting model, even if produced by an algorithm considered inherently interpretable, will be considered opaque, because the features have no straightforward meaning for a human being.

The black-box concept has been used for decades in Software Engineering. Considering a system, a component, or a function as a black-box means treating this resource as if you don't know (or don't care about) its internal behavior. In the context of XAI, it is usual to observe the term black-box in two situations: to refer to opaque algorithms or models (often black-box and opaque are used interchangeably), and in the analysis of models, ignoring their internal functioning, considering only the outputs produced by the respective inputs (here, the term has the same meaning as used in Software Engineering). In the latter case, regardless of whether the model is interpretable or not or whether access to the code or model representing the model is available, it will be treated as a black-box.

The present work uses the term black-box in this second context. Regardless of the nature of the algorithm that produced the model, it will be considered a black-box if we do not have access to either the training data used for its production or the internal details of its model.

Explainable Artificial Intelligence, in turn, corresponds to an area that studies, in different ways, how AI can produce results that can be interpreted by human beings. Typically, this interpretation (or understanding) takes place in three main ways, presented and detailed as follows.

Development of Inherently Explainable Models: many of the first models used in Artificial Intelligence systems were considered inherently interpretable, such as models based on decision trees, linear regression, or models that use association rules. Over the decades, increasingly complex models have been developed, and the understanding of the reason or logic that leads these models to produce their outputs is increasingly nebulous (Lee et al., 2020; Butt and Iqbal, 2025). This branch of XAI aims at developing new algorithms that are inherently explainable or improving those that already exist, either by modifying some of their characteristics or refining the training process to produce better models.

Explanation of Opaque / Black-Box Models: due to the fact that several black-box models have reached the state-of-the-art for some specific families of problems, many researchers consider relevant to try to explain why these models produced their results. There are two main approaches to explaining such models (Vieira and Digiampietri, 2022). In the first one, an explainable model is built to mimic the behavior of the black-box model. To this end, during the explainable model's training, the output of the black-box model is used as the values of the target variable. In the second approach, the black-box model is explained on the basis of the estimated importance of each of its input features (Ribeiro et al., 2016). Typically, several outputs are produced by varying the value of different features, and the impact on the target attribute is analyzed considering these variations.

Presentation of Examples to the User: this branch of XAI assumes that human understanding, and in particular those who are not experts in AI, can benefit from the presentation of a set of examples. The rationale is that, given a set of classification examples (input data and corresponding outputs), a person can gain a better understanding of how the model works. In particular, this branch works with counterfactual examples (Byrne, 2019), *i.e.* examples where the output was different from the one of some specific input example. This type of approach is usually used not only to explain the rationale of a model but also to guide the user on what can be done to obtain a different output. For example, if a loan request was denied, it is possible, from counterfactual examples "near" to the user input data, to present what would be necessary for the request to be accepted.

The present work deals with two of these XAI aspects. Initially, this work evaluates the explanation of black-box models through inherently explainable models (Research Question 1). In this case, we use decision trees of height three. This work is also related to the production and use of inherently explainable models while investigating whether it is possible to build explainable models based on the results of black-box models (Research Question 3). The analysis of the produced results aims to answer Research Question 2, contributing to the discussion about the effective need for opaque models in problems of different natures.

Regarding evaluation, there are different ways of assessing the outcome of an explanation (Aggarwal et al., 2019; Hoffman et al., 2019; Papenmeier et al., 2022). One of the most robust is to check with a large set of users how adequate the explanations given about a model are. However, due to the complexity and cost of questioning a large number of users, there are also metrics that can be obtained automatically to assess the simplicity of a model and the quality of some explanation methods. Due to their simplicity, two of the most commonly used metrics are the model's size, which verifies how simple (*i.e.* how easy to understand) it is, and the fidelity of some models when used to explain another.

As an instance of an interpretable model that is simple to understand, in this work, we apply decision trees of height three. The quality of this model, when explaining its black-box counterpart, will be measured by its fidelity, *i.e.* the coincidence rate between the predictions made by the black-box and the interpretable models. Although this measure may be considered equivalent to accuracy, instead of comparing the output of the interpretable model with the target variable of the test set, this comparison is made with the output of the black-box model.

Finally, the area of Algorithmic Fairness, which is closely related to XAI, has also seen significant development in recent years. Although it lacks a precise and commonplace definition, Algorithmic Fairness typically refers to attempts to correct algorithmic bias in automated decision-making processes (Aggarwal et al., 2019; Mitchell et al., 2021).

In their work, Suresh and Guttag (Suresh and Guttag, 2021) identified potential sources of harm that can lead to seven biases in machine learning algorithms: historical bias, representation bias, measurement bias, aggregation bias, learning bias, evaluation bias, and deployment bias. While each of these biases has specific characteristics and is produced in different ways (for example, if there has historically been a prejudice in society that is eventually reproduced by the algorithm or the algorithm is being used incorrectly), XAI has tools to assist in the detection and potential remediation of these biases.

Understanding an AI model, whether through an interpretable model or various explanations of opaque models, allows for an evaluation beyond traditional performance metrics. When conducted by an expert committed to minimizing biases, this process enables the model to be revised or improved before implementation, potentially making it fairer.

3 Materials and Methods

Eight public datasets commonly used for the construction and validation of artificial intelligence models were arbitrarily selected from the UCI[1] and Kaggle[2] repositories. All datasets underwent a similar pre-processing step, in which categorical data was converted to numeric data, missing values were replaced by default values, and problems with more than two classes were reduced to two classes. There were datasets with four classes following a gradation (for example, very bad, bad, good, and very good), in which case the classes were mapped to two (bad and good) in order to treat only binary problems in this

[1]UC Irvine Machine Learning Repository: https://archive.ics.uci.edu

[2]Kaggle: https://www.kaggle.com/

Table 1: Datasets description

Name	Number of Lines	Percentage of elements in the Majority Class
Breast Cancer	286	70.3%
Congressional Voting	435	61.4%
Diabetes	768	65.1%
German Credit	1,000	70.0%
Heart Failure	918	55.3%
Online News Popularity	39,644	50.7%
Online Shoppers Purchasing	12,330	84.5%
Student Performance	649	50.5%

work. Table 1 summarizes the characteristics of the data sets used after this pre-processing step.

Although we are aware that it would be possible to apply other more sophisticated pre-processing tasks, we decided to keep this step as simple as possible, producing adequate data for the next stages of the process, since this was not part of the main objectives of this work and would not help answer the research questions.

Seven algorithms were selected to create the models. This selection was made so as to use models that are constructed under different principles. Table 2 shows the algorithms used and their parameters. Notably, Decision Trees were used in two different contexts: as an interpretable model with a height limited to three and as a black-box model with default values for all its parameters, except the maximum number of interactions for Logistic Regression and Multilayer Perceptron, we increase this number in order to guarantee the convergence of these models. SVM also had two implementations, both used as a black-box: one using a linear kernel and the other using a polynomial kernel. In this work, we use the implementations of these algorithms presented in the Scikit-Learn Python library[3].

Table 2: Algorithms and parameters

Algorithm	Parameters	Use in this work
Decision Tree	*max_depth* = 3	Interpretable
Decision Tree	(*default*)	Black-Box
Random Forest	(*default*)	Black-Box
SVM	(*default*)	Black-Box
SVM	*kernel* = '*poly*'	Black-Box
Logistic Regression	*max_iter* = 5000	Black-Box
Multilayer Perceptron	*max_iter* = 5000	Black-Box
Gaussian Naive Bayes	(*default*)	Black-Box
KNN	(*default*)	Black-Box

After pre-processing, each dataset was randomly split into three subsets, through stratified random sampling from a uniform distribution. As a result, 40% of the data built the first training set and other 40% the second, with the remaining 20% of the data being regarded as a test set, as illustrated in Fig. 1. This division is justified by the need for different training sets for the black-box models and for the interpretable models that intend to explain those black-boxes. In particular, when a company uses a black-box system, it is common not to have access to the training data of that system. This division of sets makes it possible to simulate this situation.

Figure 1: Data set stratified split strategy.

After splitting up the data, we took a four-stage approach to this research, as illustrated in Fig. 2. It is worth stressing, at this point, that this process was repeated for each combination of black-box model and dataset. In the first step (Fig. 2A), Training Set 1 is used to train both the interpretable model (Interpretable Model 1) and the black-box model. The purpose of Interpretable Model 1 is not to explain the black-box but instead to be used in the comparative analysis of the models in order to answer Research Question 2.

In the second stage (Fig. 2B), the trained black-box model is applied to Training Set 2, making its predictions for these data ($prediction_2$). Next, at the third stage (Fig. 2C), two additional interpretable models (Interpretable Model 2 and 3) are trained in Training Set 2. One of them has as its target variable the values from the training set, while the other takes the output produced by the black-box model ($prediction_2$). Finally, at the fourth step (Fig. 2D), all four models make their predictions in the Test set. These predictions are, in turn, used to evaluate the performance of the models.

In the present work, all models were evaluated using accuracy and macro-F1 score. The explanatory capacity of Interpretable Model 3, in relation to the black-box models, was assessed through the fidelity measure in order to answer Research Question 1. Research Question 3, in turn, will be answered based on the evaluation of the outputs of Interpretable Model 2 ($prediction_{I2}$) when compared to the values of the Test Set's target variable. All the materials used or developed in this project were made publicly available[4].

[3]Scikit-Learn Version 1.3: https://scikit-learn.org/stable/

[4]https://drive.google.com/drive/folders/1gL7OGYkWgOEjAbSiB4J8hOBE8UTj966r

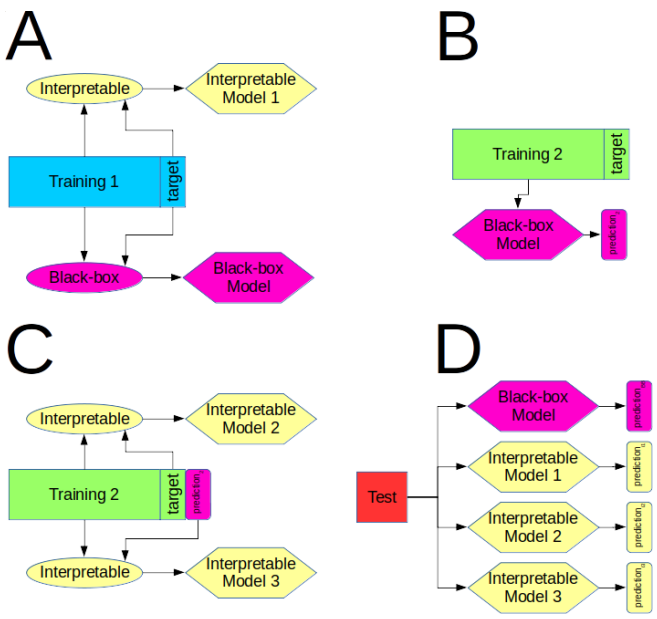

Figure 2: The four stages of this research.

Table 3: Accuracy for Interpretable Model 1 and each of the Black-Box Models

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

4 Results and Discussion

We start our discussion by presenting the performance of the models trained at Training Set 1 (Section 4.1). Next, in Section 4.2, we analyze our results regarding fidelity. Finally, in Section 4.3, we describe our findings related to the improvement of the interpretable model training based on the outputs of the black-box models.

4.1 Model Performance

The model accuracy achieved when training at Training Set 1 can be seen in Table 3, with their corresponding F1-scores being shown in Table 4. To ease visualization, in these tables, cells with a more intense green background correspond to higher values. The lowest values are indicated by cells in more intense red.

It can be observed that for the selected models and datasets, Random Forest was the model that delivered the best accuracy in seven of the eight datasets (tying with the other four models at the Congressional Voting dataset), losing its position to SVM with a polynomial kernel only in the Student Performance set, where it scored third (around 4% worse than the SVM). In terms of accuracy, the performance of the interpretable model was, on average[5], 93.2% of the performance of the best model for each of the datasets (ranging from 82.3% for the Student Performance dataset to 100%, *i.e.* equal to the best result obtained for the Congressional Voting dataset).

Considering the macro-F1 score, the model with the best performance was again the Random Forest Algorithm, obtaining the best result for six of the eight data sets. The Gaussian Naive Bayes model reached the best F1 score for the German Credit and Student Performance sets, tying, in the latter case, with the SVM with a polynomial kernel.

On average, the F1 score of the interpretable model corresponded to 90.5% of the values obtained by the best-performing models (ranging from 82.7% for the Student Performance set to 100% for the Congressional Voting set). Based on these results, it can be said that the interpretable model used in this work performed worse (on average) than the best model for each of the datasets. As such, RQ2 was answered confirming there is a difference in performance between them.

[5]All averages in this work correspond to the arithmetic mean.

Table 4: F1 score results from Interpretable Model 1 and each of the Black-Box Models

Dataset	Interpretable Model 1	Decision Tree	Random Forest	SVM (linear)	SVM (polynomial)	Logistic Regression	Multilayer Perceptron	Gaussian Naive Bayes	KNN
Breast Cancer	0,610	0,610	0,670	0,410	0,530	0,660	0,410	0,640	0,590
Congressional Voting	0,950	0,950	0,950	0,950	0,930	0,930	0,940	0,950	0,900
Diabetes	0,680	0,700	0,770	0,720	0,720	0,740	0,700	0,740	0,690
German Credit	0,570	0,650	0,610	0,500	0,470	0,570	0,500	0,640	0,540
Heart Failure	0,840	0,790	0,900	0,700	0,720	0,890	0,860	0,890	0,630
Online News Popularity	0,620	0,570	0,660	0,530	0,530	0,600	0,350	0,410	0,560
Online Shoppers Purchasing	0,690	0,750	0,790	0,460	0,460	0,700	0,690	0,690	0,610
Student Performance	0,620	0,610	0,720	0,720	0,750	0,710	0,680	0,750	0,710

Table 5: Fidelity results of Interpretable Model 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

4.2 Fidelity of the interpretable model

In this section, we evaluate the ability of Interpretable Model 3 to mimic the results produced by the black-box models, calculated through the fidelity measure. Table 5 presents the fidelity of the results of Interpretable Model 3 concerning the results of the black-boxes models.

As it turns out, the overall average fidelity was 88.9%. When considering each black-box model, it is observed that, on average, the highest fidelity of Interpretable Model 3 occurred with SVM using a linear kernel (94.6%). On the other hand, the worst average performance was 77.1%, for the Decision Tree with no height limit. This last result is rather interesting as it indicates that a smaller tree (height three) could not mimic well the results of a tree that had no height limit.

Here, it is worth noting that the decision tree model treated as a black-box had, on average, achieved the worst value for the F1 score and one of the worst accuracy results, both values lower than the results of the Interpretable Model 3 (see Section 4.1). This indicates that the model was unable to create rules that generalized the learned knowledge. Not limiting the tree height may have resulted in an overly specific model, i.e., overfitting, which the interpretable model could not replicate.

In addition to the average fidelity (88.9%), as well as the lowest and highest average per model (respectively 77.1% and 94.6%), we considered it appropriate to compare these values with the equivalent values produced by Interpretable Model 1. It is important to remember that Interpretable Model 1 was not built to mimic the black-box models, but rather to solve the same problem that the black-box models were solving. On average, the intersection of the results of the interpretable model with the black-box models (which is equivalent to fidelity) was 79.2%, ranging from 70.5% for the model based on Decision Trees to 86.4% for the SVM with a linear kernel.

Thus, on average, Interpretable Model 3 managed to satisfactorily mimic the black-box models, significantly better than the corresponding model, which had not been trained to this end (Interpretable Model 1), thereby answering RQ1. It is noteworthy that, as presented in the related literature (*e.g.* Rudin and Radin (2019)), even high values of fidelity may not signify that the same logic "learned" by the black-box model was also "learned" by the interpretable model, and this is one of the main limitations of this specific type of models' explanation.

4.3 Improvements to the interpretable model

As explained in Section 3, the interpretable models were built using the output of the black-box models (what was called Interpretable Model 3). The fidelity of these models was presented in the previous section. In addition to fidelity, this work hypothesized that interpretable models built from the output of black-box models could outperform interpretable models trained using the target variable of the training set. This hypothesis was based on the fact that black-box models correspond to simplifications of the world and, therefore, it might be easier for the interpretable model to learn from this simplification instead of directly learning from the original data.

To carry out this analysis, the performance of Interpretable Models 2 and 3 were compared (both were built based on Training Set 2). In this case, while Model 2 took its target variable from the training set, Interpretable Model 3 used the output of each black-box model as the target. Table 6 and Table 7 present the results for accuracy

Table 6: Accuracy comparison of Interpretable Models 2 and 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

Table 7: F1 score comparison of Interpretable Models 2 and 3

Dataset	Interpretable Model 2	Decision Tree	Random Forest	SVM (linear)	SVM (polynomial)	Logistic Regression	Multilayer Perceptron	Gaussian Naive Bayes	KNN
Breast Cancer	0,470	0,490	0,720	0,410	0,410	0,730	0,410	0,600	0,560
Congressional Voting	0,940	0,950	0,950	0,940	0,950	0,950	0,930	0,950	0,900
Diabetes	0,650	0,670	0,720	0,680	0,690	0,700	0,680	0,720	0,670
German Credit	0,610	0,620	0,570	0,480	0,470	0,610	0,530	0,590	0,530
Heart Failure	0,850	0,850	0,850	0,680	0,720	0,890	0,830	0,860	0,640
Online News Popularity	0,630	0,620	0,620	0,530	0,510	0,580	0,340	0,400	0,560
Online Shoppers Purchasing	0,800	0,730	0,780	0,460	0,460	0,710	0,690	0,700	0,590
Student Performance	0,630	0,610	0,590	0,640	0,610	0,600	0,640	0,580	0,580

and F1 score. In these tables, the first column indicates the dataset, the second shows the results of Interpretable Model 2, and the remaining columns contain the results for Interpretable Model 3, according to the output of each black-box model.

As shown in Table 6, Interpretable Model 2 obtained only three of the best accuracy results, showing that models trained with the output of the black-box models presented better results for five of the data sets. The interpretable model built using the output of the Random Forest model stands out, being superior (although sometimes slightly) to Interpretable Model 2 in five of the eight datasets.

Regarding F1 score, Interpretable Model 2 achieved the best performance only in two datasets (Online News Popularity and Online Shoppers Purchasing). The interpretable models built using the output of the Logistic Regression Model performed equal to or better than Interpretable Model 2 on five of the eight datasets. The same occurred with the models that used the output of the model based on Decision Trees as input with no height limit.

These results help to provide an initial answer to RQ3, indicating that training interpretable models based on the values of the target variable of the training set does not always deliver the best performances when compared to models trained using the output of other black-box models. Additional investigations are necessary, however, as well as the calculation of different statistical measures. Our results nevertheless indicate that this is an interesting venue for investigation.

5 Conclusion

Machine learning has revolutionized how knowledge can be discovered from data. Since the emergence of Artificial Intelligence, different models have been created, allowing not only automated problem-solving but also a better understanding of the data.

Although the desire to understand the processes that involve the decisions made by models created by AI is not something new and is fundamental in some areas, such as health for example, some of the newer techniques have an inherent complexity that makes it virtually impossible to fully understand the process that leads to each of their outputs. Thus, the development of strategies to try to explain the decisions made by black-box algorithms has become fundamental.

In this work, we compared the performance of models produced by an interpretable algorithm to models taken as black-boxes produced by algorithms of different kinds. In addition to comparing performance between algorithms, fidelity (the ability of the interpretable model to mimic part of the behavior of black-box models) was also measured. Finally, it was verified whether it is possible to use black-box models to assist in the interpretable models training.

All in all, the expected difference in performance between the black-box models and the interpretable model was verified, which, for many problems, serves as a justification for the use of black-box models. A satisfactory ability of the interpretable model to mimic the behavior of the black-box model was also observed. Fidelity, on average, was higher than the accuracy of the interpretable model solving the problem at hand.

As limitations of the present work, we emphasize that the results and conclusions are solely based on

the materials and methods of use, i.e., eight datasets, an algorithm used to build the interpretable models, and eight models treated as black-boxes built by seven different algorithms. As directions for future work, we intend to expand the study, including a larger set of datasets, different interpretable algorithms, and more black-box algorithms.

References

Aggarwal, A., Lohia, P., Nagar, S., Dey, K. and Saha, D. (2019). Black box fairness testing of machine learning models, *Proceedings of the 2019 27th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, ESEC/FSE 2019, Association for Computing Machinery, New York, NY, USA, p. 625–635. https://doi.org/10.1145/3338906.3338937.

Angwin, J., Larson, J., Mattu, S. and Kirchner, L. (2022). *Machine Bias*, Auerbach Publications, New York, NY, p. 11. https://doi.org/10.1201/9781003278290.

Brasil (2018). Lei Geral de Proteção de Dados Pessoais. http://www.planalto.gov.br/ccivil_03/_Ato2015-2018/2018/Lei/L13709.htm.

Butt, T. and Iqbal, M. (2025). Explainable ai: Applications, challenges, current solutions and future research directions, *12th International Conference on Information Technology (ICIT)*, p. 108 – 113. https://doi.org/10.1145/3718391.3718408.

Byrne, R. M. J. (2019). Counterfactuals in explainable artificial intelligence (xai): Evidence from human reasoning, *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, International Joint Conferences on Artificial Intelligence Organization, Macao, China, pp. 6276–6282. https://doi.org/10.24963/ijcai.2019/876.

Coelho, J. and Burg, T. (2020). Uso de inteligencia artificial pelo poder publico.

Council of the European Union (2018). General Data Protection Regulation. https://gdpr-info.eu/.

Francisco, P. A., Hurel, L. M. and Rielli, M. M. (2020). Regulacao do reconhecimento facial no setor publico. https://igarape.org.br/regulacao-do-reconhecimento-facial-no-setor-publico/.

Hoffman, R. R., Mueller, S. T., Klein, G. and Litman, J. (2019). Metrics for explainable ai: Challenges and prospects.

Lee, S. J., Pandey, H. S. and Garcia, G.-G. P. (2020). *Designing Interpretable Machine Learning Models using Mixed Integer Programming*, Springer International Publishing, Cham, pp. 1–8. https://doi.org/10.1007/978-3-030-54621-2_867-1.

Lipton, Z. C. (2018). The mythos of model interpretability, *Commun. ACM* **61**(10): 36–43. https://doi.org/10.1145/3233231.

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences, *Artificial Intelligence* **267**: 1–38. https://www.sciencedirect.com/science/article/pii/S0004370218305988.

Mitchell, S., Potash, E., Barocas, S., D'Amour, A. and Lum, K. (2021). Algorithmic fairness: Choices, assumptions, and definitions, *Annual Review of Statistics and Its Application* **8**(1): 141–163. https://doi.org/10.1146/annurev-statistics-042720-125902.

Mohseni, S., Zarei, N. and Ragan, E. D. (2021). A multidisciplinary survey and framework for design and evaluation of explainable ai systems, *ACM Trans. Interact. Intell. Syst.* **11**(3–4). https://doi.org/10.1145/3387166.

Nyholm, S. and Smids, J. (2016). The ethics of accident-algorithms for self-driving cars: an applied trolley problem?, *Ethical Theory and Moral Practice Article* **19**: 1275–1289. https://doi.org/10.1007/s10677-016-9745-2.

Papenmeier, A., Kern, D., Englebienne, G. and Seifert, C. (2022). It's complicated: The relationship between user trust, model accuracy and explanations in ai, *ACM Trans. Comput.-Hum. Interact.* **29**(4). https://doi.org/10.1145/3495013.

Ramos, S. (2019). Retratos da violencia - cinco meses de monitoramento, analises e descobertas, *Rede de Observatorios da Seguranca/CESeC* **1**(1): 1–72. https://cesecseguranca.com.br/textodownload/retratos-da-violencia-cinco-meses-de-monitoramento-analises-e-descobertas/.

Ribeiro, M., Singh, S. and Guestrin, C. (2016). "why should I trust you?": Explaining the predictions of any classifier, *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, Association for Computational Linguistics, San Diego, California, pp. 97–101.
URL: *https://aclanthology.org/N16-3020*

Rudin, C. and Radin, J. (2019). 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): 10. https://hdsr.mitpress.mit.edu/pub/f9kuryi8.

Suresh, H. and Guttag, J. (2021). A framework for understanding sources of harm throughout the machine learning life cycle, *Equity and Access in Algorithms, Mechanisms, and Optimization*, EAAMO '21, Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3465416.3483305.

Vieira, C. P. and Digiampietri, L. A. (2022). Machine learning post-hoc interpretability: A systematic mapping study, *XVIII Brazilian Symposium on Information Systems*, SBSI, Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3535511.3535512.