ECML PKDD 2025

MAINLE: a Multi-Agent, Interactive, Natural Language Local Explainer of Classification Tasks

Paulo Bruno Serafim, Rômulo Férrer Filho, Stenio Freitas, Gizem Gezici,
Fosca Giannotti, Franco Raimondi, and Alexandre Santos

17 September, 2025

GSSI GRAN SASSO SCIENCE INSTITUTE
SCHOOL OF ADVANCED STUDIES
Scuola Universitaria Superiore

SCUOLA NORMALE SUPERIORE

UFC

Motivation

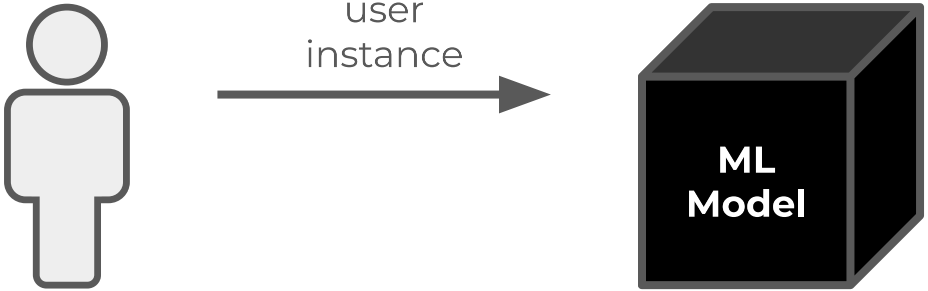
- How to **explain** a Machine Learning **classification** to a **non-expert** user?

Machine Learning is Everywhere

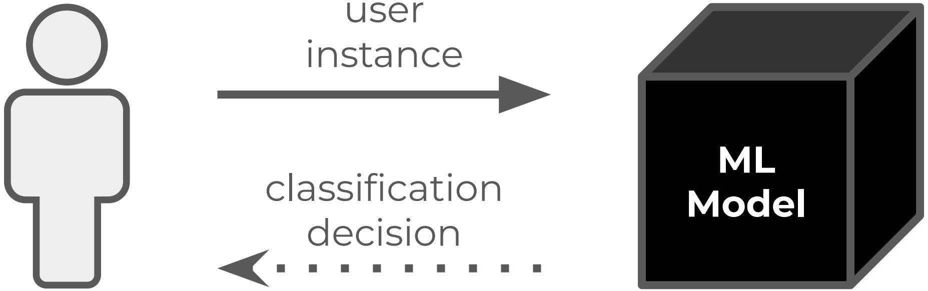
- Computers, phones, tvs, home appliances…
- Customer support, recommendation system, classification, predictions…

Example Scenario

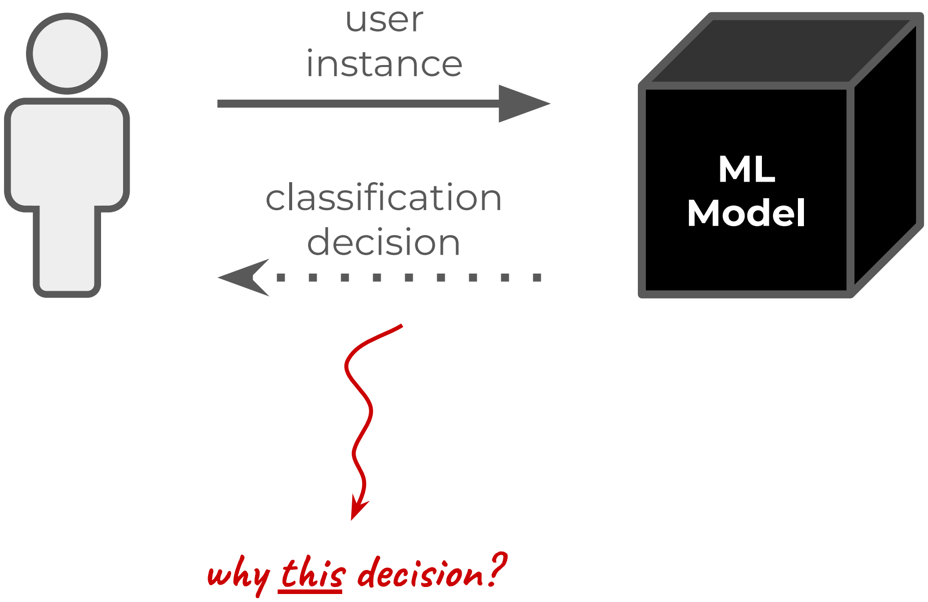
Example Scenario



Example Scenario

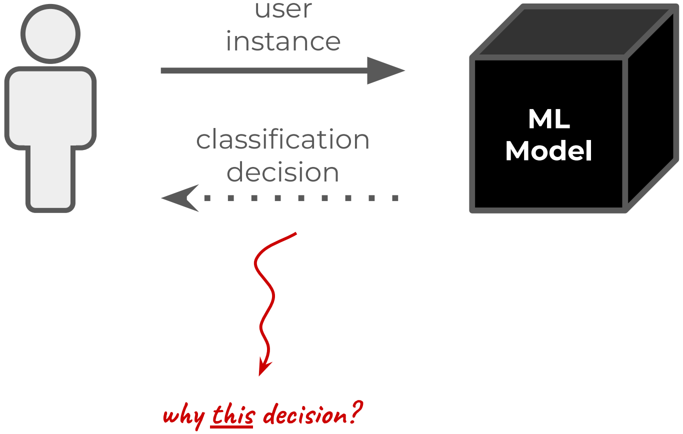


Example Scenario



Motivation

- How to explain a Machine Learning classification to a non-expert user?



user instance

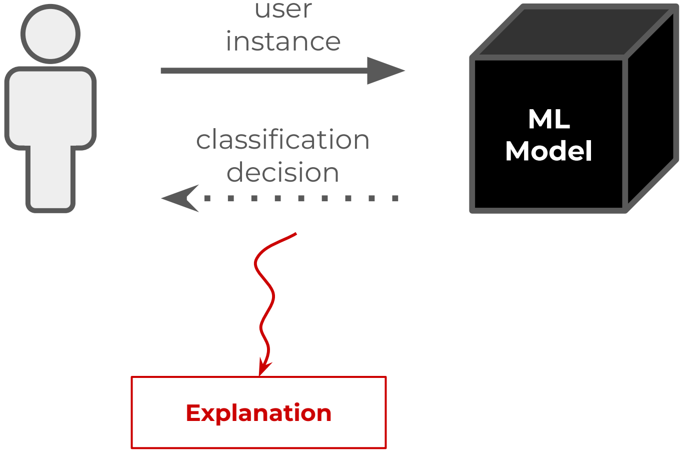
classification decision

ML Model

why this decision?

Motivation

- How to explain a Machine Learning classification to a non-expert user?
  - **Explainable AI (XAI) can help**



Problems

It is hard to **use** XAI methods

Problems

It is hard to **use** XAI methods

- Command Line
- Git
- Programming
- Python
- PyTorch / Tensor Flow
- Machine Learning
- …

Problems

It is hard to **present** Explanations

Problems

It is hard to **present** Explanations

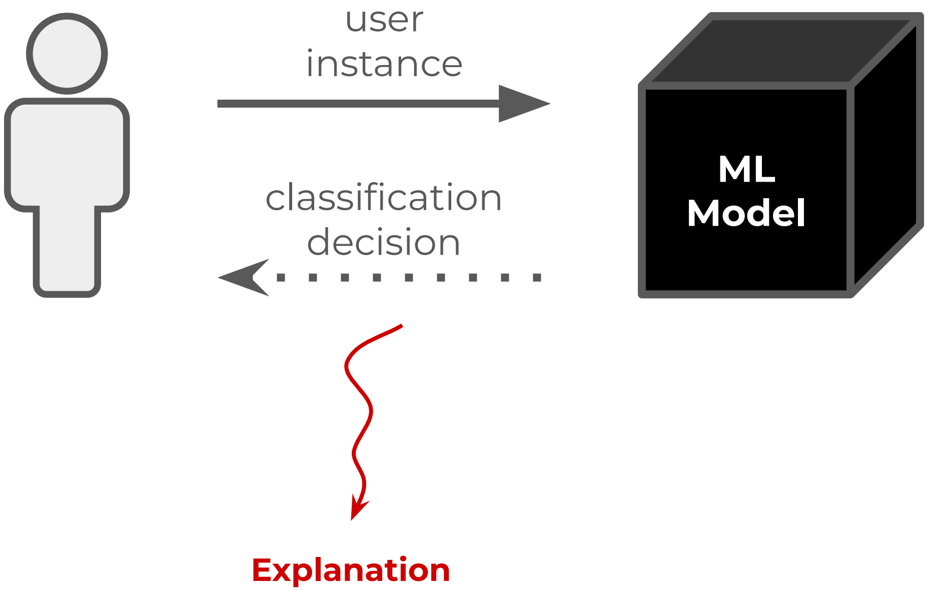
- List of numbers
- Tables
- Spreadsheets
- Scrollers
- …

Problems

Our objective is to remove these barriers

How?

Example Scenario

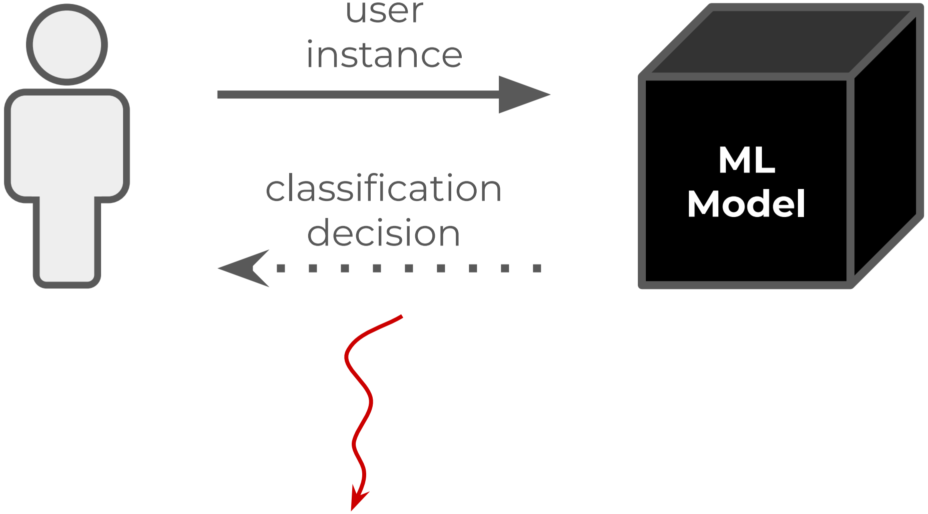

user
instance

classification
decision

ML
Model

Explanation

Example Scenario



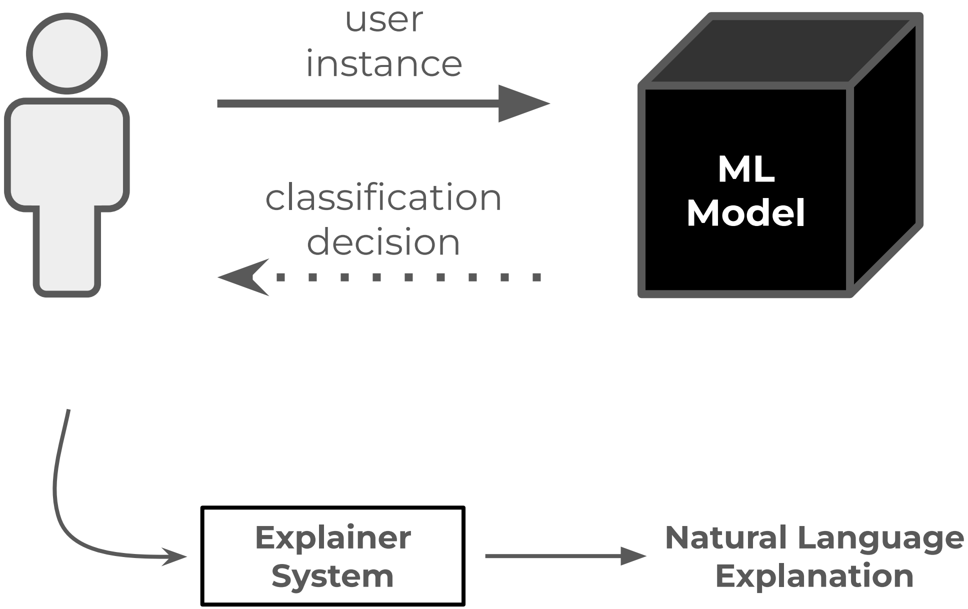
user instance

classification decision

ML Model

***Natural Language* Explanation**

Example Scenario



user
instance

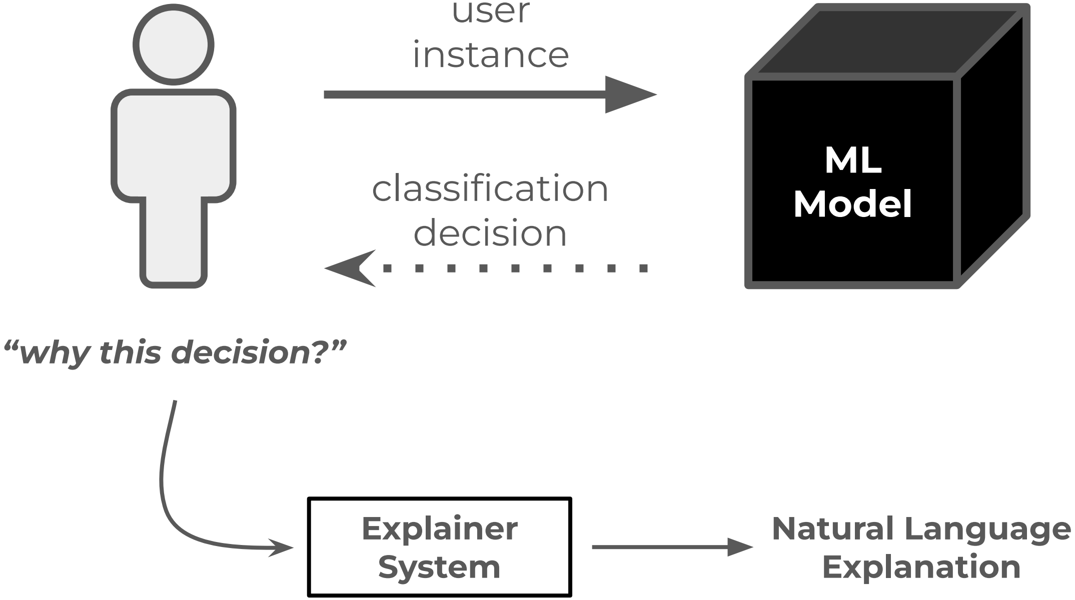
ML
Model

classification
decision

Explainer
System

Natural Language
Explanation

Example Scenario



user instance

ML Model

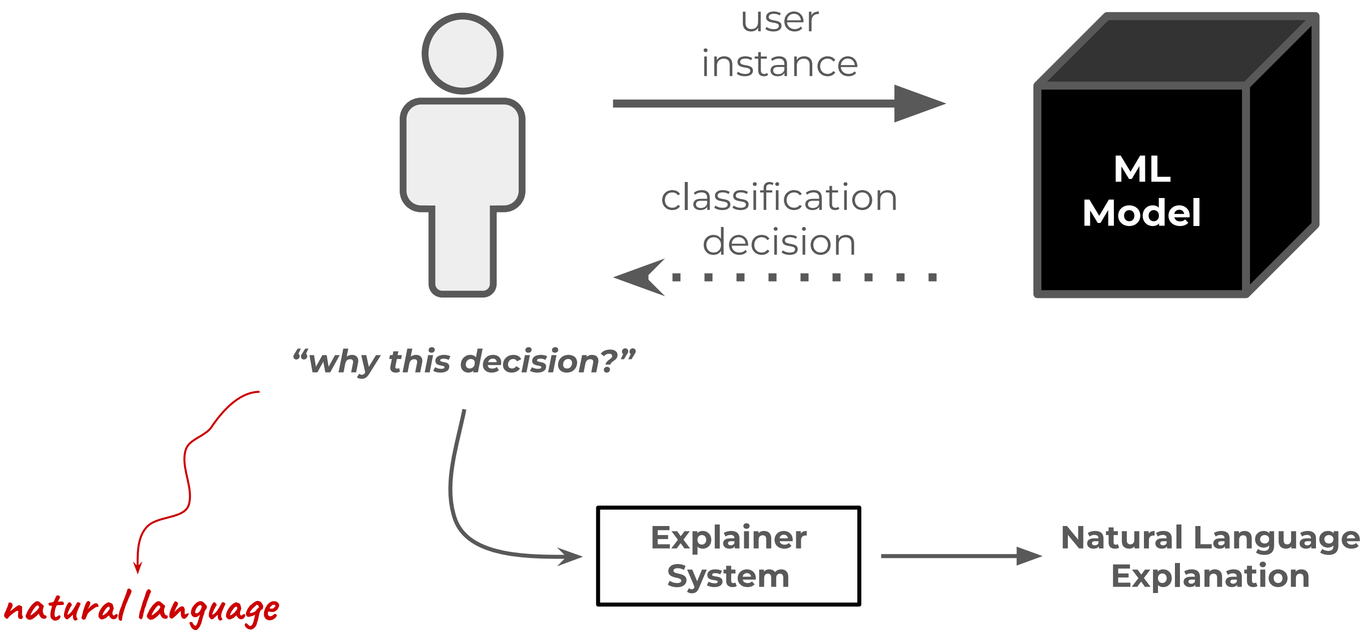
classification decision

"why this decision?"

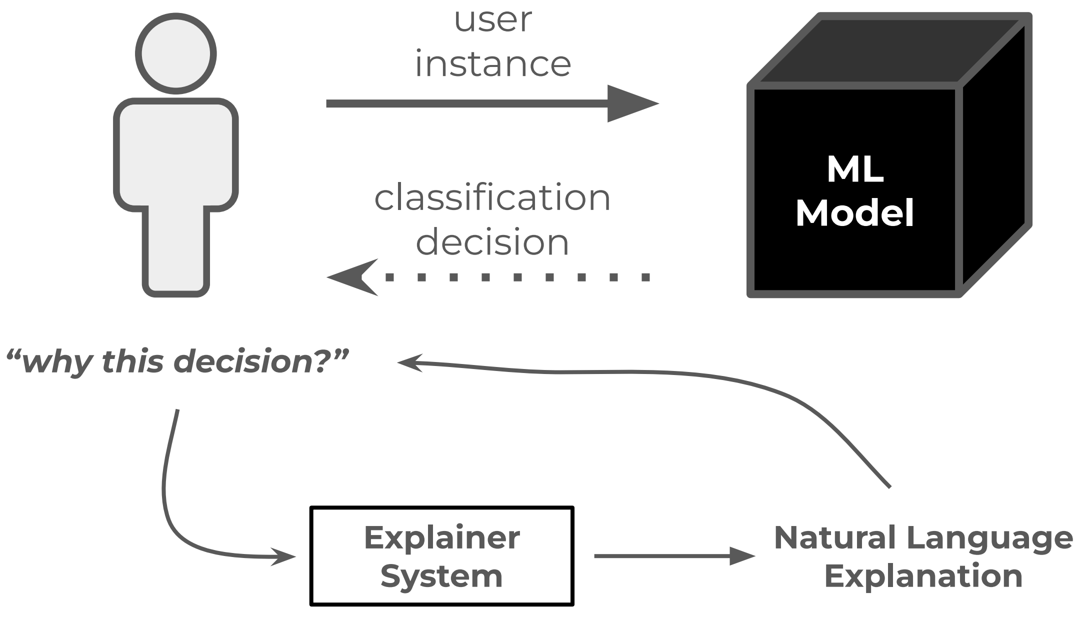
Explainer System

Natural Language Explanation

Example Scenario



Example Scenario



user instance

ML Model

classification decision

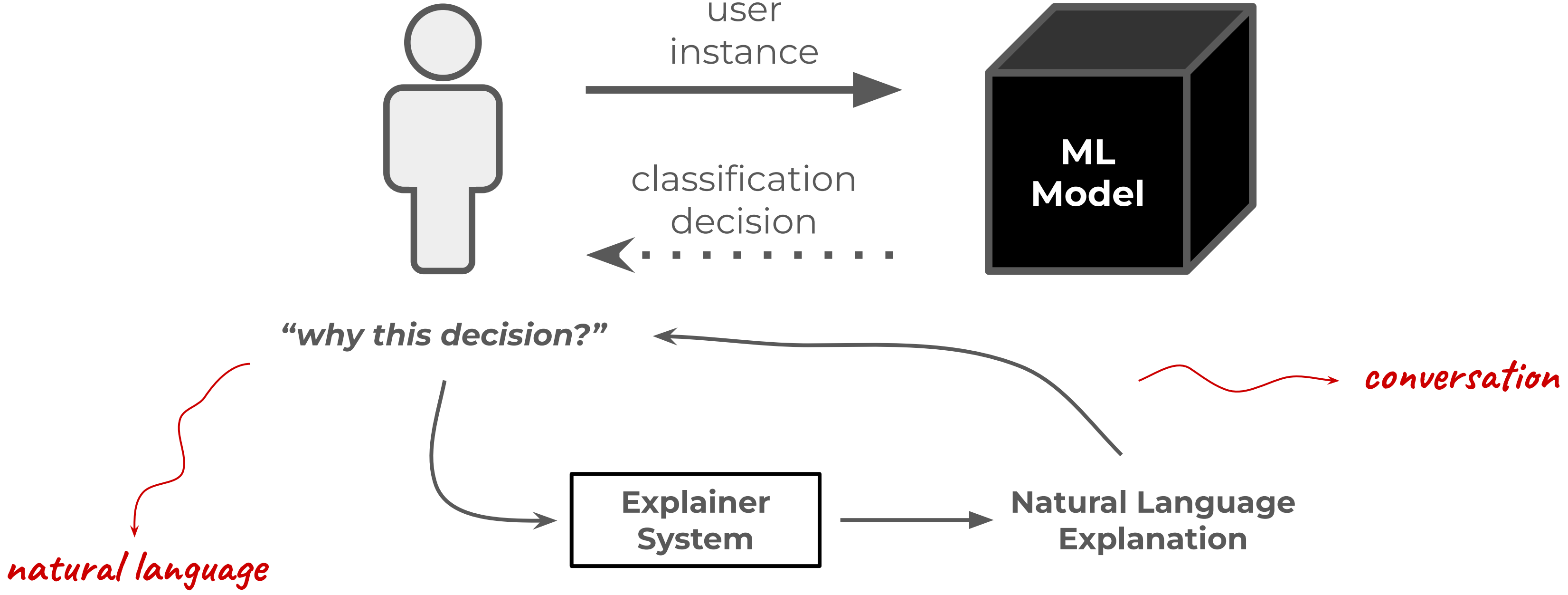
“why this decision?”

Explainer System

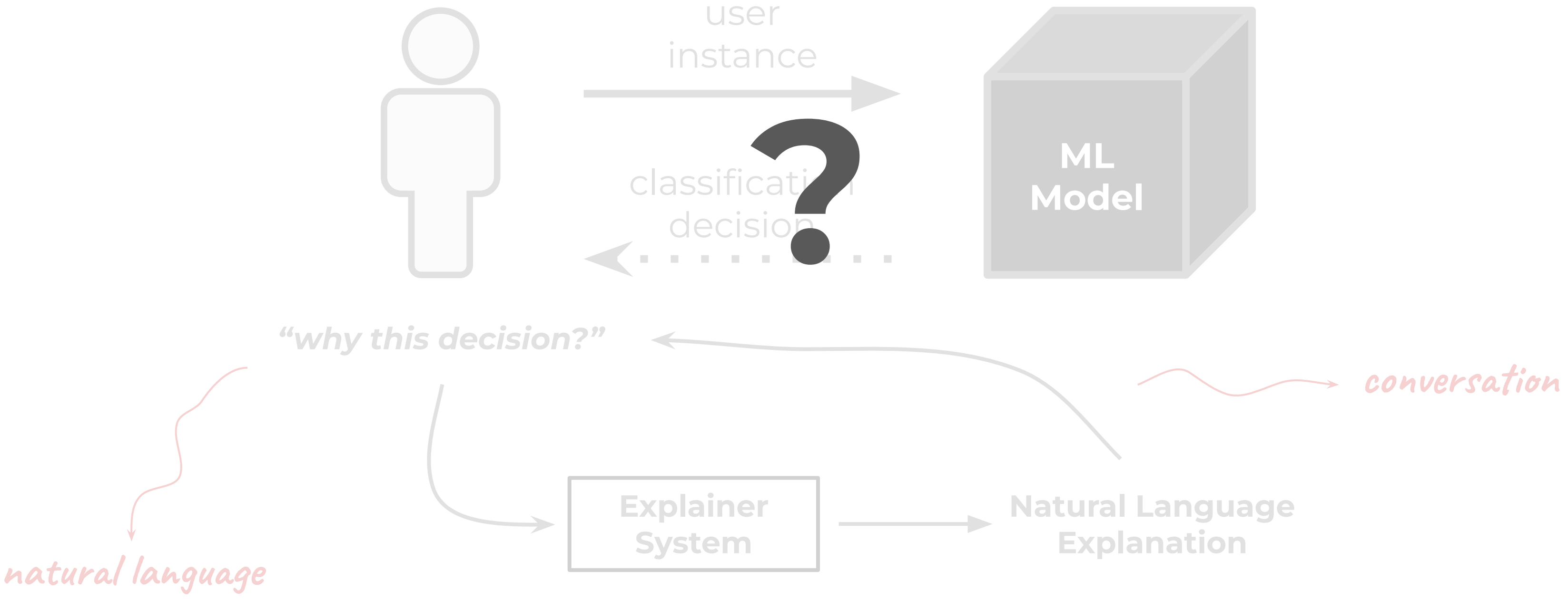
Natural Language Explanation

natural language

Example Scenario



Example Scenario



user instance

classification decision

ML Model

“why this decision?”

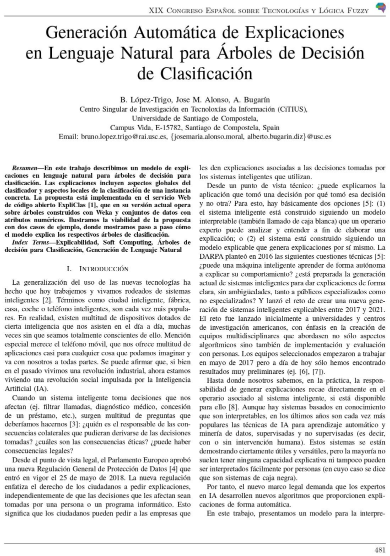
Explainer System

Natural Language Explanation

conversation

natural language

Related Work



Generación Automática de Explicaciones en Lenguaje Natural para Árboles de Decisión de Clasificación

B. López-Trigo, Jose M. Alonso, A. Bugarín
Centro Singular de Investigación en Tecnoloxías da Información (CiTIUS),
Universidade de Santiago de Compostela,
Campus Vida, E-15782, Santiago de Compostela, Spain
Email: bruno.lopez.trigo@rai.usc.es, {josemaria.alonso.moral, alberto.bugarin.diz}@usc.es

***Resumen*—En este trabajo describimos un modelo de explicaciones en lenguaje natural para árboles de decisión para clasificación. Las explicaciones incluyen aspectos globales del clasificador y aspectos locales de la clasificación de una instancia concreta. La propuesta está implementada en el servicio Web de código abierto ExpliClas [1], que en su versión actual opera sobre árboles construidos con Weka y conjuntos de datos con atributos numéricos. Ilustramos la viabilidad de la propuesta con dos casos de ejemplo, donde mostramos paso a paso cómo el modelo explica los respectivos árboles de clasificación.**

***Index Terms*—Explicabilidad, Soft Computing, Árboles de decisión para Clasificación, Generación de Lenguaje Natural**

I. INTRODUCCIÓN

La generalización del uso de las nuevas tecnologías ha hecho que hoy trabajemos y vivamos rodeados de sistemas inteligentes [2]. Términos como ciudad inteligente, fábrica, casa, coche o teléfono inteligentes, son cada vez más populares. En realidad, existen multitud de dispositivos dotados de cierta inteligencia que nos asisten en el día a día, muchas veces sin que seamos totalmente conscientes de ello. Mención especial merece el teléfono móvil, que nos ofrece multitud de aplicaciones casi para cualquier cosa que podamos imaginar y va con nosotros a todas partes. Se puede afirmar que, si bien en el pasado vivimos una revolución industrial, ahora estamos viviendo una revolución social impulsada por la Inteligencia Artificial (IA).

Cuando un sistema inteligente toma decisiones que nos afectan (ej. filtrar llamadas, diagnóstico médico, concesión de un préstamo, etc.), surgen multitud de preguntas que deberíamos hacernos [3]: ¿quién es el responsable de las consecuencias colaterales que pudieran derivarse de las decisiones tomadas? ¿cuáles son las consecuencias éticas? ¿puede haber consecuencias legales?

Desde el punto de vista legal, el Parlamento Europeo aprobó una nueva Regulación General de Protección de Datos [4] que entró en vigor el 25 de mayo de 2018. La nueva regulación enfatiza el derecho de los ciudadanos a pedir explicaciones, independientemente de que las decisiones que les afectan sean tomadas por una persona o un programa informático. Esto significa que los ciudadanos pueden pedir a las empresas que les den explicaciones asociadas a las decisiones tomadas por los sistemas inteligentes que utilizan.

Desde un punto de vista técnico: ¿puede explicarnos la aplicación que tomó una decisión por qué tomó esa decisión y no otra? Para esto, hay básicamente dos opciones [5]: (1) el sistema inteligente está construido siguiendo un modelo interpretable (también llamado de caja blanca) que un operario experto puede analizar y entender a fin de elaborar una explicación; o (2) el sistema está construido siguiendo un modelo explicable que genera explicaciones por sí mismo. La DARPA planteó en 2016 las siguientes cuestiones técnicas [5]: ¿puede una máquina inteligente aprender de forma autónoma a explicar su comportamiento? ¿está preparada la generación actual de sistemas inteligentes para dar explicaciones de forma clara, sin ambigüedades, tanto a públicos especializados como no especializados? Y lanzó el reto de crear una nueva generación de sistemas inteligentes explicables entre 2017 y 2021. El reto fue lanzado inicialmente a universidades y centros de investigación americanos, con énfasis en la creación de equipos multidisciplinares que abordasen no sólo aspectos algorítmicos sino también de implementación y evaluación con personas. Los equipos seleccionados empezaron a trabajar en mayo de 2017 pero a día de hoy sólo hemos encontrado resultados muy preliminares (ej. [6], [7]).

Hasta donde nosotros sabemos, en la práctica, la responsabilidad de generar explicaciones recae directamente en el operario asociado al sistema inteligente, si está disponible para ello [8]. Aunque hay sistemas basados en conocimiento que son interpretables, en los últimos años son cada vez más populares las técnicas de IA para aprendizaje automático y minería de datos, supervisadas y no supervisadas (es decir, con o sin intervención humana). Estos sistemas se están demostrando ciertamente útiles y versátiles, pero la mayoría no suelen tener ninguna capacidad explicativa ni tampoco pueden ser interpretados fácilmente por personas (en cuyo caso se dice que son sistemas de caja negra).

Por tanto, el nuevo marco legal demanda que los expertos en IA desarrollen nuevos algoritmos que proporcionen explicaciones de forma automática.

En este trabajo, presentamos un modelo para la interpre-

López-Trigo, Alonso and Bugarín (2018)
Generación Automática de Explicaciones en Lenguaje Natural para Árboles de Decisión de Clasificación

Related Work

Generación Automática de Explicaciones en Lenguaje Natural para Árboles de Decisión de Clasificación

B. López-Trigo, Jose M. Alonso, A. Bugarín
Centro Singular de Investigación en Tecnoloxías da Información (CiTIUS),
Universidade de Santiago de Compostela,
Campus Vida, E-15782, Santiago de Compostela, Spain
Email: bruno.lopez.trigo@rai.usc.es, {josemaria.alonso.moral, alberto.bugarin.diz}@usc.es

***Resumen*—En este trabajo describimos un modelo de explicaciones en lenguaje natural para árboles de decisión para clasificación. Las explicaciones incluyen aspectos globales del clasificador y aspectos locales de la clasificación de una instancia concreta. La propuesta está implementada en el servicio Web de código abierto ExpliClas [1], que en su versión actual opera sobre árboles construidos con Weka y conjuntos de datos con atributos numéricos. Ilustramos la viabilidad de la propuesta con dos casos de ejemplo, donde mostramos paso a paso cómo el modelo explica los respectivos árboles de clasificación.**

***Index Terms*—Explicabilidad, Soft Computing, Árboles de decisión para Clasificación, Generación de Lenguaje Natural**

I. INTRODUCCIÓN

La generalización del uso de las nuevas tecnologías ha hecho que hoy trabajemos y vivamos rodeados de sistemas inteligentes [2]. Términos como ciudad inteligente, fábrica, casa, coche o teléfono inteligentes, son cada vez más populares. En realidad, existen multitud de dispositivos dotados de cierta inteligencia que nos asisten en el día a día, muchas veces sin que seamos totalmente conscientes de ello. Mención especial merece el teléfono móvil, que nos ofrece multitud de aplicaciones casi para cualquier cosa que podamos imaginar y va con nosotros a todas partes. Se puede afirmar que, si bien en el pasado vivimos una revolución industrial, ahora estamos viviendo una revolución social impulsada por la Inteligencia Artificial (IA).

Cuando un sistema inteligente toma decisiones que nos afectan (ej. filtrar llamadas, diagnóstico médico, concesión de un préstamo, etc.), surgen multitud de preguntas que deberíamos hacernos [3]: ¿quién es el responsable de las consecuencias colaterales que pudieran derivarse de las decisiones tomadas? ¿cuáles son las consecuencias éticas? ¿puede haber consecuencias legales?

Desde el punto de vista legal, el Parlamento Europeo aprobó una nueva Regulación General de Protección de Datos [4] que entró en vigor el 25 de mayo de 2018. La nueva regulación enfatiza el derecho de los ciudadanos a pedir explicaciones, independientemente de que las decisiones que les afectan sean tomadas por una persona o un programa informático. Esto significa que los ciudadanos pueden pedir a las empresas que les den explicaciones asociadas a las decisiones tomadas por los sistemas inteligentes que utilizan.

Desde un punto de vista técnico: ¿puede explicarnos la aplicación que tomó una decisión por qué tomó esa decisión y no otra? Para esto, hay básicamente dos opciones [5]: (1) el sistema inteligente está construido siguiendo un modelo interpretable (también llamado de caja blanca) que un operario experto puede analizar y entender a fin de elaborar una explicación; o (2) el sistema está construido siguiendo un modelo explicable que genera explicaciones por sí mismo. La DARPA planteó en 2016 las siguientes cuestiones técnicas [5]: ¿puede una máquina inteligente aprender de forma autónoma a explicar su comportamiento? ¿está preparada la generación actual de sistemas inteligentes para dar explicaciones de forma clara, sin ambigüedades, tanto a públicos especializados como no especializados? Y lanzó el reto de crear una nueva generación de sistemas inteligentes explicables entre 2017 y 2021. El reto fue lanzado inicialmente a universidades y centros de investigación americanos, con énfasis en la creación de equipos multidisciplinares que abordasen no sólo aspectos algorítmicos sino también de implementación y evaluación con personas. Los equipos seleccionados empezaron a trabajar en mayo de 2017 pero a día de hoy sólo hemos encontrado resultados muy preliminares (ej. [6], [7]).

Hasta donde nosotros sabemos, en la práctica, la responsabilidad de generar explicaciones recae directamente en el operario asociado al sistema inteligente, si está disponible para ello [8]. Aunque hay sistemas basados en conocimiento que son interpretables, en los últimos años son cada vez más populares las técnicas de IA para aprendizaje automático y minería de datos, supervisadas y no supervisadas (es decir, con o sin intervención humana). Estos sistemas se están demostrando ciertamente útiles y versátiles, pero la mayoría no suelen tener ninguna capacidad explicativa ni tampoco pueden ser interpretados fácilmente por personas (en cuyo caso se dice que son sistemas de caja negra).

Por tanto, el nuevo marco legal demanda que los expertos en IA desarrollen nuevos algoritmos que proporcionen explicaciones de forma automática.

En este trabajo, presentamos un modelo para la interpre-

López-Trigo, Alonso and Bugarín (2018)
Generación Automática de Explicaciones en Lenguaje Natural para Árboles de Decisión de Clasificación

Explaining Tree Model Decisions in Natural Language for Network Intrusion Detection

Noah Ziems, Gang Liu, John Flanagan, Meng Jiang
University of Notre Dame
`{nziems2, gliu7, jflanag5, mjiang2}@nd.edu`

Abstract

Network intrusion detection (NID) systems which leverage machine learning have been shown to have strong performance in practice when used to detect malicious network traffic. Decision trees in particular offer a strong balance between performance and simplicity, but require users of NID systems to have background knowledge in machine learning to interpret. In addition, they are unable to provide additional outside information as to why certain features may be important for classification.

In this work, we explore the use of large language models (LLMs) to provide explanations and additional background knowledge for decision tree NID systems. Further, we introduce a new human evaluation framework for decision tree explanations, which leverages automatically generated quiz questions that measure human evaluators' understanding of decision tree inference. Finally, we show LLM generated decision tree explanations correlate highly with human ratings of readability, quality, and use of background knowledge while simultaneously providing better understanding of decision boundaries.

1 Introduction

Network intrusion detection (NID) systems monitor incoming and outgoing network traffic to detect potentially malicious activity. When suspicious traffic is detected, a system administrator is alerted for further investigation. Early NID systems gained popularity by analysing traffic with simple rule-based approaches [1]. However, over time new types of attacks such as malware, phishing, worms, sql injection, scan attacks, and distributed denial-of-service (DDoS) attacks have significantly increased the vulnerable attack surface of networks. As the kinds of attacks increase, so too does the complexity of detecting them using rule-based methods alone. In response, machine learning (ML) and data mining based NID systems which learn to detect malicious traffic from large amounts of real data began to emerge [2]. However, due to their black-box nature, understanding *why* and *how* a model has come to a particular conclusion remains a significant limitation of ML based NID systems in real-world applications. While decision trees offer a strong balance between simplicity and performance, they still require a user familiar with their underlying mechanics to interpret why a particular prediction was made. Efforts to assist in explaining decision-tree based NID systems often rely on providing additional information such as quantifying the importance of features for a given prediction [3–5]. How one can create decision tree explanations for NID systems that can be easily understood by users with no background knowledge of machine learning remains an open area of research. Further, how to measure the quality of these explanations also remains an open question.

The recent success of large language models (LLMs) has enabled high quality open-ended text generation for a wide variety of tasks such as long-form question answering, summarization, arithmetic reasoning, and more. In this work, we explore whether large language models can

XAI in Action: Past, Present, and Future Applications @ NeurIPS 2023.

Ziems, Liu, Flanagan and Jian (2023)
Explaining Tree Model Decisions in Natural Language for Network Intrusion Detection

64
HHAI 2024: Hybrid Human AI Systems for the Social Good
F. Lorig et al. (Eds.)
© 2024 The Authors.
This article is published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC 4.0).
doi:10.3233/FAIA240183

Exploring Large Language Models Capabilities to Explain Decision Trees

Paulo Bruno SERAFIM[a,1], Pierluigi CRESCENZI[a], Gizem GEZICI[b],
Eleonora CAPPUCCIO[c,d], Salvatore RINZIVILLO[d] and Fosca GIANNOTTI[b]
[a] *Gran Sasso Science Institute*
[b] *Scuola Normale Superiore di Pisa*
[c] *Università di Pisa, Università di Bari Aldo Moro*
[d] *Istituto di Scienza e Tecnologie dell'Informazione, CNR*

Abstract. Decision trees are widely adopted in Machine Learning tasks due to their operation simplicity and interpretability aspects. However, following the decision process path taken by trees can be difficult in a complex scenario or in a case where a user has no familiarity with them. Prior research showed that converting outcomes to natural language is an accessible way to facilitate understanding for non-expert users in several tasks. More recently, there has been a growing effort to use Large Language Models (LLMs) as a tool for providing natural language texts. In this paper, we examine the proficiency of LLMs to explain decision tree predictions in simple terms through the generation of natural language explanations. By exploring different textual representations and prompt engineering strategies, we identify capabilities that strengthen LLMs as a competent explainer as well as highlight potential challenges and limitations, opening further research possibilities on natural language explanations for decision trees.

Keywords. Explainable AI, decision tree, natural language generation

1. Introduction

The high complexity of novel Machine Learning (ML) models is an important factor in the production of remarkable results since simpler models were not able to achieve the same level of performance. However, as the models grow, there have been growing concerns due to the opaque nature of widely used black-box algorithms. Among the problems that arise from the non-transparency are the trustworthiness, fairness, and accountability of the models [1]. In this context, the need for an approach to bridge the gap between the algorithmic perspective and human-centered explanation is evident.

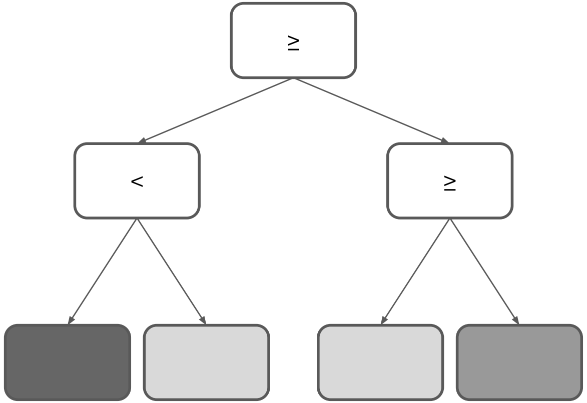
Among key explanatory elements from disciplines such as social sciences and psychology, interactivity has been described as a core value for the implementation of human-centric explainable AI applications [2]. The explanation process has to be outlined as a continuous dialogue between a sender and a receiver, making the user engage in a conversation [3,4]. More recently, Large Language Models (LLM) have shown an exceptional ability to generate natural language responses [5]. As such, LLMs emerge

---

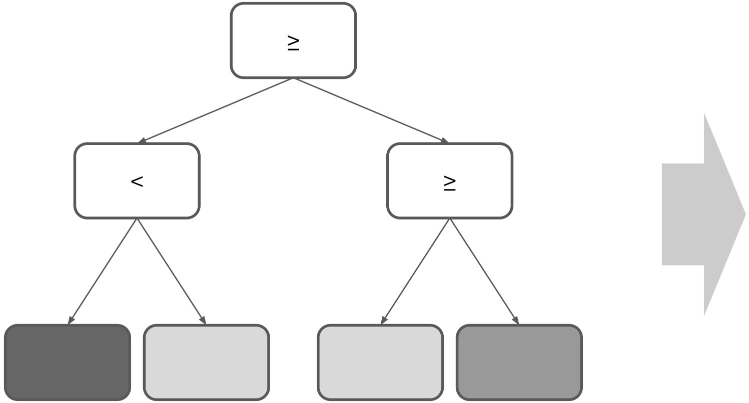
[1] Corresponding Author: Paulo Bruno Serafim, paulo.desousa@gssi.it.

Serafim, Crescenzi, Gezici, Cappuccio, Rinzivillo, and Giannotti (2024)
Exploring Large Language Models Capabilities to Explain Decision Trees

Decision Tree to NLE

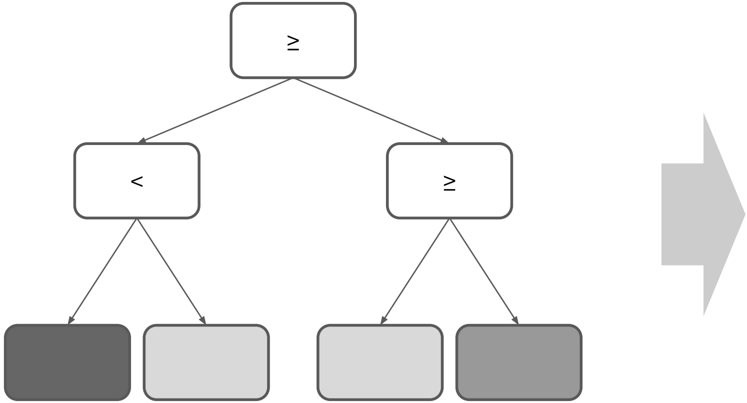


Decision Tree to NLE



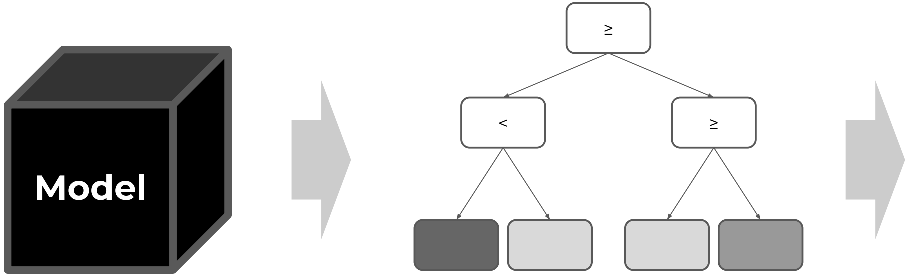
Natural
Language Explanation

Decision Tree to NLE



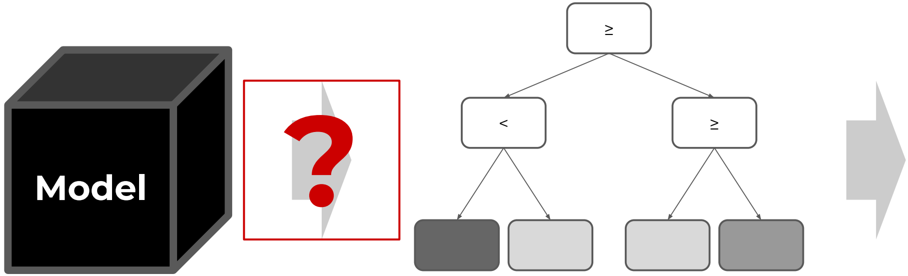
Simplified *Natural Language Explanation*

Decision Tree to NLE



Simplified Natural Language Explanation

Decision Tree to NLE



Simplified Natural Language Explanation

LORE

IEEE INTELLIGENT SYSTEMS JOURNAL - RECEIVED DATE: ?, REVISED DATE: ? 1

Factual and Counterfactual Explanations for Black-Box Decision Making

Riccardo Guidotti, Anna Monreale, Fosca Giannotti, Dino Pedreschi, Salvatore Ruggieri, and Franco Turini

***Abstract*—The rise of sophisticated machine learning models has brought accurate but obscure decision systems, which hide their logic, thus undermining transparency, trust, and the adoption of AI in socially sensitive and safety-critical contexts. We introduce a local rule-based explanation method providing faithful explanations of the decision made by a black-box classifier on a specific instance. The proposed method first learns an interpretable, local classifier on a synthetic neighborhood of the instance under investigation, generated by a genetic algorithm. Then it derives from the interpretable classifier an explanation consisting of a decision rule, explaining the factual reasons of the decision, and a set of counterfactuals, suggesting the changes in the instance features that would lead to a different outcome. Experimental results show that the proposed method outperforms existing approaches in terms of the quality of the explanations and of the accuracy in mimicking the black-box.**

***Index Terms*—Explainable AI, Interpretable Machine Learning, Open the Black Box, Explanation Rules, Counterfactuals**

I. INTRODUCTION

Newspapers are full of commentaries about algorithms taking critical decisions that heavily impact on our life and society, from loan concession in bank systems to pedestrian detection in self-driving cars. The worry is not only due to the increasing automation of decision making, but mostly to the fact that the algorithms are opaque and their logic unexplained. The main cause of this lack of transparency is that decision-making algorithms are generated often from data through Machine learning (ML). ML allows building predictive models that map user features into a decision, obtained by generalizing from a dataset of examples. The process of inferring a classification model from examples cannot be easily controlled because the size of training data and the complexity of the learned model are too big for humans. The inability to obtain an explanation for a decision is a profound drawback of learning from data, limiting social acceptance and trust in its adoption in many sensitive contexts.

In this paper we address the problem of explaining the decision outcome taken by an algorithm providing “meaningful explanations of the logic involved” when automated decision making takes place, as prescribed by the “right to explanation” of the European General Data Protection Regulation. We perform our research under some specific assumptions. First, we assume that an explanation is interesting if it clarifies why a *specific decision* has been made, i.e., we aim for *local* explanations, not general, *global*, descriptions of how the overall system works [1], [2]. Second, we assume that the explanations should be as close as possible to the language of reasoning, which is formal logic. Thus, we assume the user can understand elementary logic rules, but it should also be considered that from logic rules it is easy to construct narratives that are understood by users with diverse expertise. Finally, we assume that the black-box decision system can be queried as many times as necessary, to probe its decision behavior to the scope of reconstructing its logic; this is certainly the case in a legal argumentation in court, or in an industrial setting. On the other hand, we make no assumptions on the algorithms used in the obscure classifier: we aim at an *agnostic* explanation method analyzing the input-output behavior of the black-box, disregarding its internals [3].

We propose LORE, a LOcal Rule-based Explanation method for tabular data. Given a black-box binary predictor b and a specific instance x labeled with outcome y by b, we build a simple, interpretable predictor by first generating a balanced set of neighbor instances of the given x through an ad-hoc genetic algorithm, and then extracting from such a set labelled with b a decision tree classifier. A *local explanation* is then extracted from the obtained decision tree. The local explanation is a pair composed by *(i)* a – factual – *logic rule*, corresponding to the path in the tree that explains why x has been labeled as y by b, and *(ii)* a set of *counterfactual rules*, explaining which changes in x would invert the class y assigned by b. For example, from the `compas` dataset we may have the following explanation: the rule $\{age{\leq}39, race{=}African\text{–}American, recidivist{=}True\}{\rightarrow}HighRisk$ and the counterfactuals $\{age{>}40\}$ and $\{race{=}White\text{–}American\}$. Here, the factual explanation is that the high risk of recidivism is predicted for a black younger than 40 with prior recidivism; the counterfactuals explain that a lower risk would be predicted if the person were either older than 40 or white. The usefulness of the explanation depends on the stakeholder: it may make sense to a judge that wants to understand and evaluate the suggestion by the decision support system and possibly discover that it biased against blacks.

The intuition behind our method, common to other *local* approaches, such as LIME [3], and ANCHOR [4] is that the decision boundary for the black-box can be arbitrarily complex over the whole data space, but in the neighborhood of a data point there is a high chance that the decision boundary is clear and simple, hence amenable to be captured by an interpretable model. These methods are named *local* because they focus on the behavior of the black-box in the neighborhood of the

---

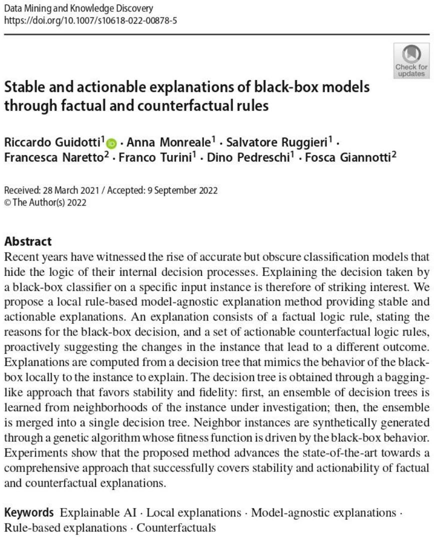
Riccardo Guidotti and Fosca Giannotti are with ISTI CNR, Pisa, Italy. e-mail: firstname.lastname@isti.cnr.it.

Anna Monreale, Dino Pedreschi, Salvatore Ruggieri and Franco Turini are with University of Pisa, Italy . e-mail: firstname.lastname@unipi.it.

Manuscript received August ? , 2019; revised ?.

Guidotti, Monreale, Giannotti, Pedreschi, Ruggieri, and Turini (2019)

Factual and Counterfactual Explanations for Black-Box Decision Making



Data Mining and Knowledge Discovery
https://doi.org/10.1007/s10618-022-00878-5

Check for updates

Stable and actionable explanations of black-box models through factual and counterfactual rules

Riccardo Guidotti[1] · Anna Monreale[1] · Salvatore Ruggieri[1] · Francesca Naretto[2] · Franco Turini[1] · Dino Pedreschi[1] · Fosca Giannotti[2]

Received: 28 March 2021 / Accepted: 9 September 2022
© The Author(s) 2022

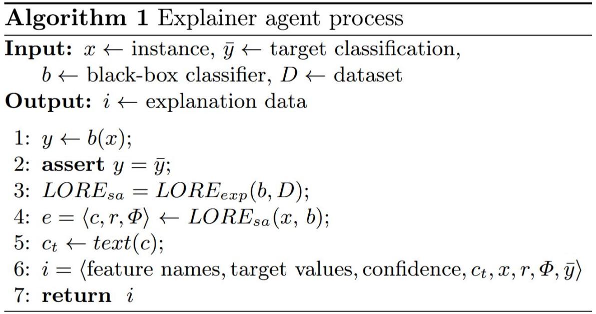
Abstract

Recent years have witnessed the rise of accurate but obscure classification models that hide the logic of their internal decision processes. Explaining the decision taken by a black-box classifier on a specific input instance is therefore of striking interest. We propose a local rule-based model-agnostic explanation method providing stable and actionable explanations. An explanation consists of a factual logic rule, stating the reasons for the black-box decision, and a set of actionable counterfactual logic rules, proactively suggesting the changes in the instance that lead to a different outcome. Explanations are computed from a decision tree that mimics the behavior of the black-box locally to the instance to explain. The decision tree is obtained through a bagging-like approach that favors stability and fidelity: first, an ensemble of decision trees is learned from neighborhoods of the instance under investigation; then, the ensemble is merged into a single decision tree. Neighbor instances are synthetically generated through a genetic algorithm whose fitness function is driven by the black-box behavior. Experiments show that the proposed method advances the state-of-the-art towards a comprehensive approach that successfully covers stability and actionability of factual and counterfactual explanations.

Keywords Explainable AI · Local explanations · Model-agnostic explanations · Rule-based explanations · Counterfactuals

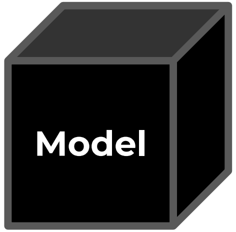
Guidotti, Monreale, Ruggieri, Naretto, Turini, Pedreschi, and Giannotti (2022)

Stable and actionable explanations of black-box models through factual and counterfactual rules

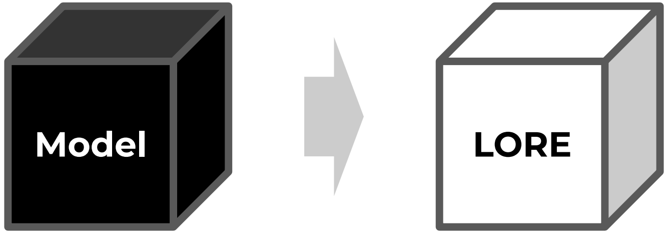


```
Algorithm 1 Explainer agent process
Input: $x \leftarrow$ instance, $\bar{y} \leftarrow$ target classification,
    $b \leftarrow$ black-box classifier, $D \leftarrow$ dataset
Output: $i \leftarrow$ explanation data
$y \leftarrow b(x)$;
assert $y = \bar{y}$;
$LORE_{sa} = LORE_{exp}(b, D)$;
$e = \langle c, r, \Phi \rangle \leftarrow LORE_{sa}(x,\ b)$;
$c_t \leftarrow text(c)$;
$i = \langle \text{feature names}, \text{target values}, \text{confidence}, c_t, x, r, \Phi, \bar{y} \rangle$
return $i$
```

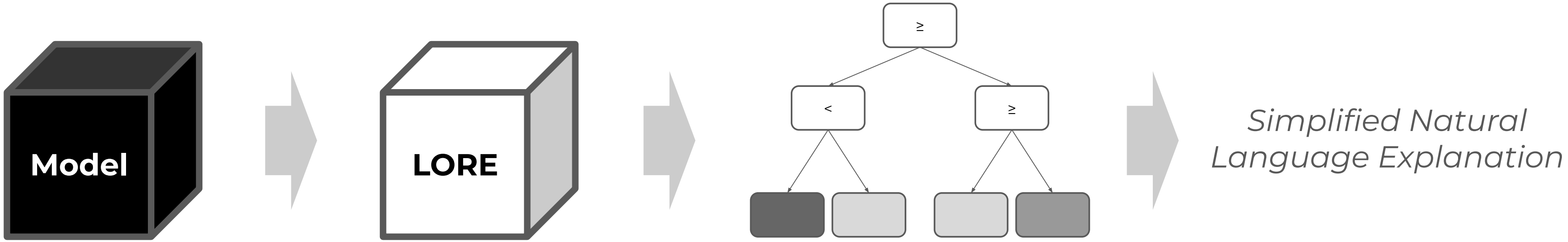
Decision Tree to NLE



Decision Tree to NLE



Decision Tree to NLE



Simplified Natural Language Explanation

Goal

*Develop a **conversational interface** to provide **simplified explanations** for **non-expert** users*

Goal

Develop a conversational interface to provide simplified explanations for non-expert users

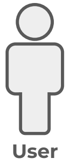
- Internally, an explanation is generated from a Decision Tree built from the classifier using LORE

Goal

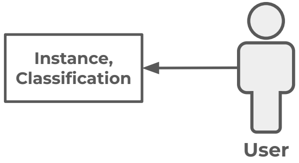
Develop a conversational interface to provide simplified explanations for non-expert users

- Internally, an explanation is generated from a Decision Tree built from the classifier using LORE
- The generation process is distributed among **multiple specialized agents**

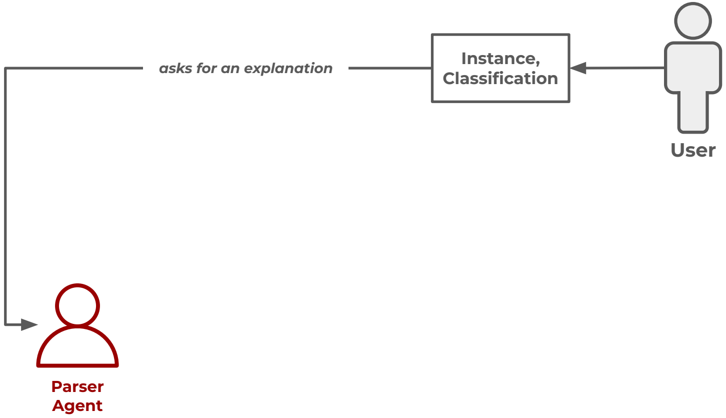
MAINLE’s Simplified Architecture



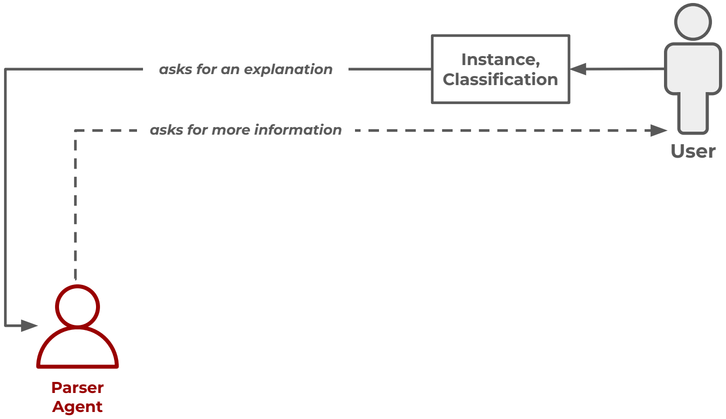
MAINLE's Simplified Architecture



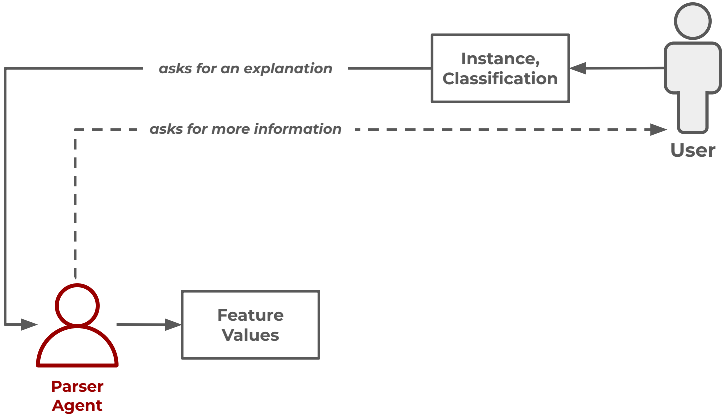
MAINLE's Simplified Architecture



MAINLE’s Simplified Architecture



MAINLE’s Simplified Architecture



asks for an explanation

Instance, Classification

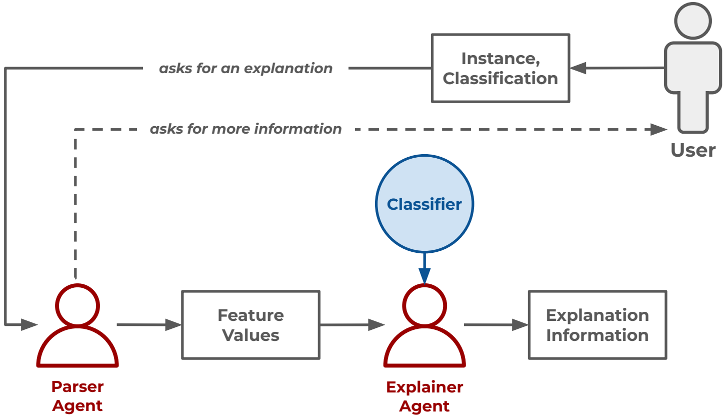
asks for more information

User

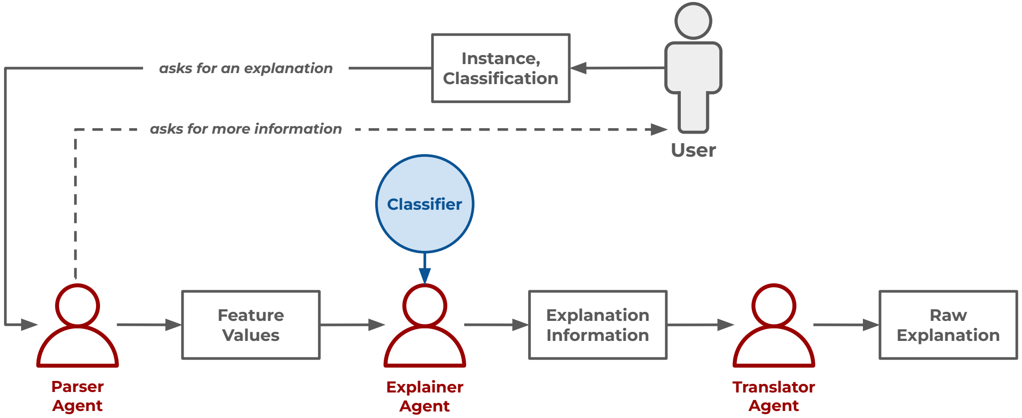
Parser Agent

Feature Values

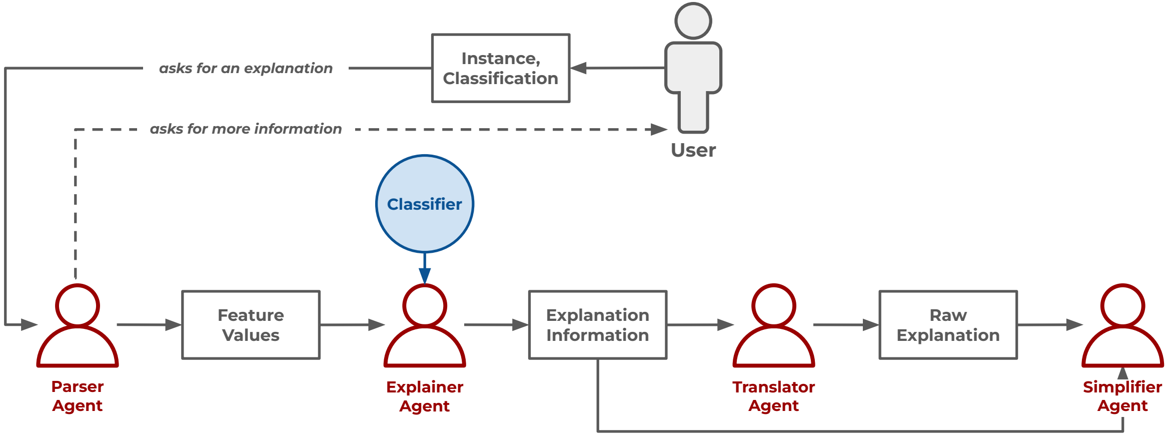
MAINLE’s Simplified Architecture



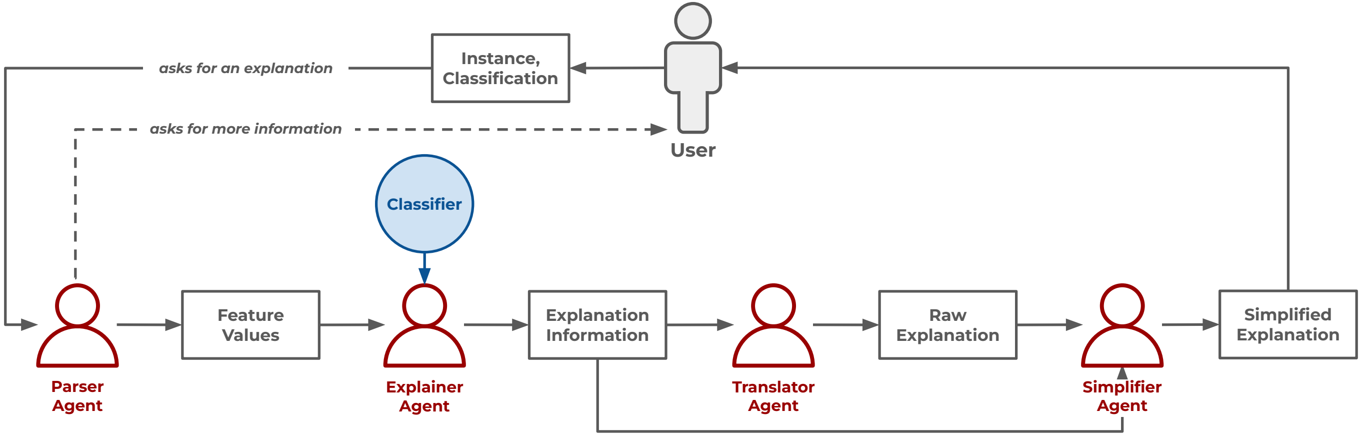
MAINLE's Simplified Architecture



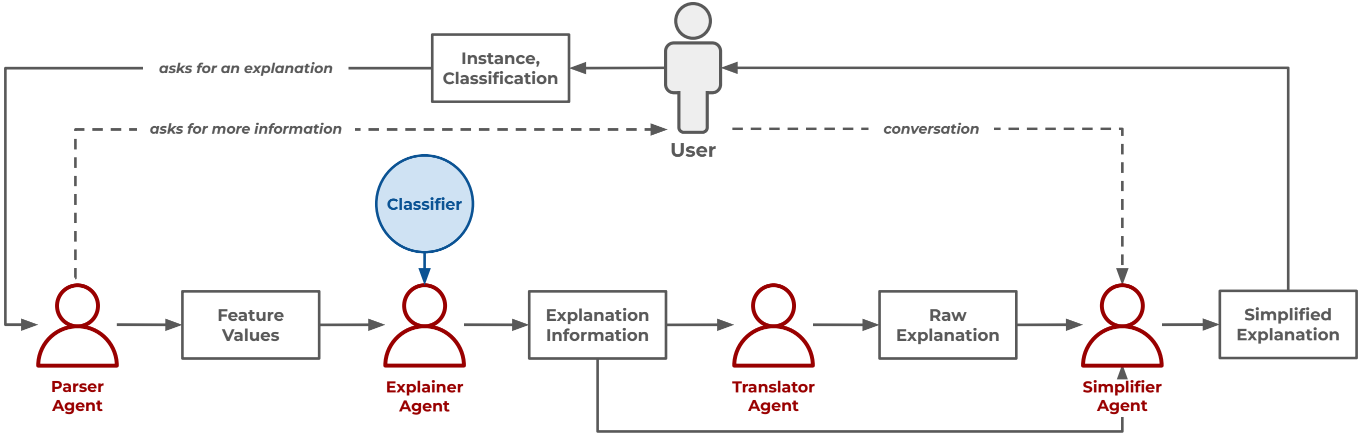
MAINLE's Simplified Architecture



MAINLE's Simplified Architecture

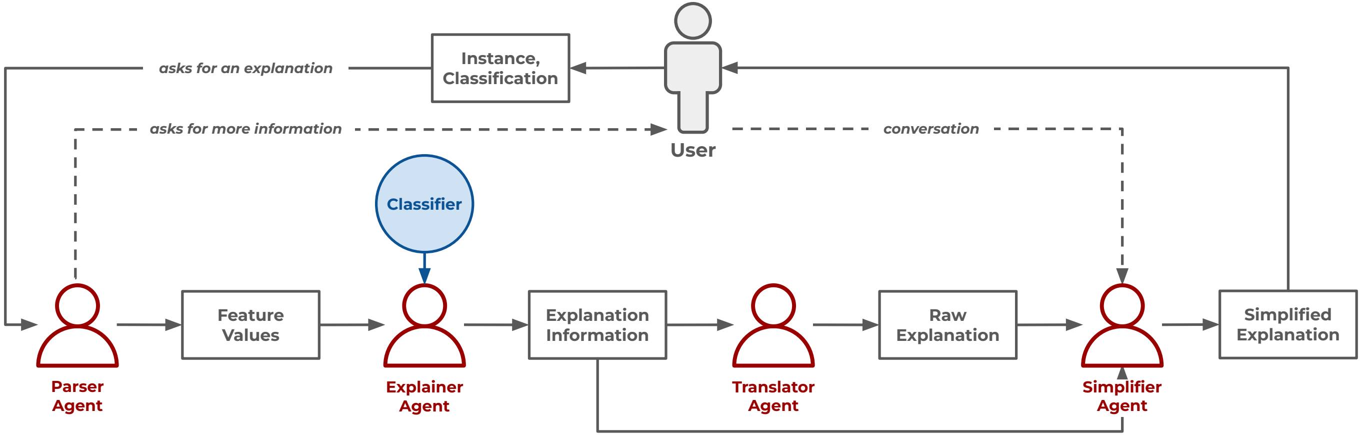


MAINLE’s Simplified Architecture

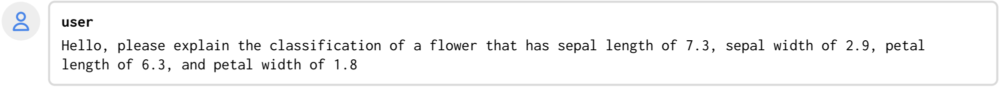


MAINLE

a Multi-Agent architecture, interactive, that provides Natural Language Local Explanations of classification outcomes



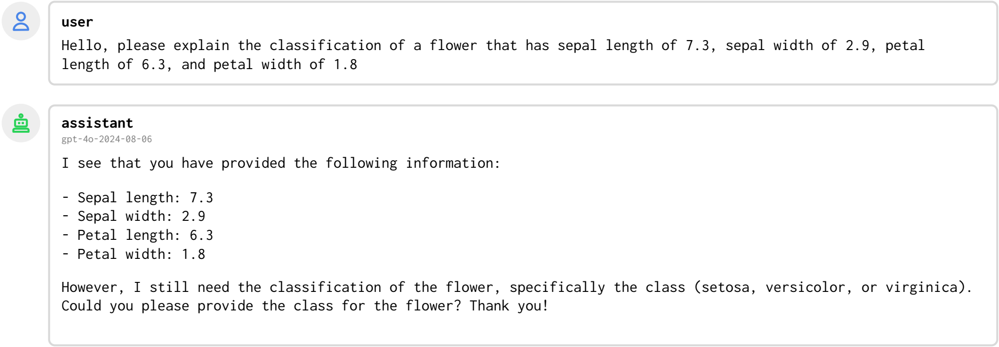
Example



user

Hello, please explain the classification of a flower that has sepal length of 7.3, sepal width of 2.9, petal length of 6.3, and petal width of 1.8

Example



user

Hello, please explain the classification of a flower that has sepal length of 7.3, sepal width of 2.9, petal length of 6.3, and petal width of 1.8

assistant

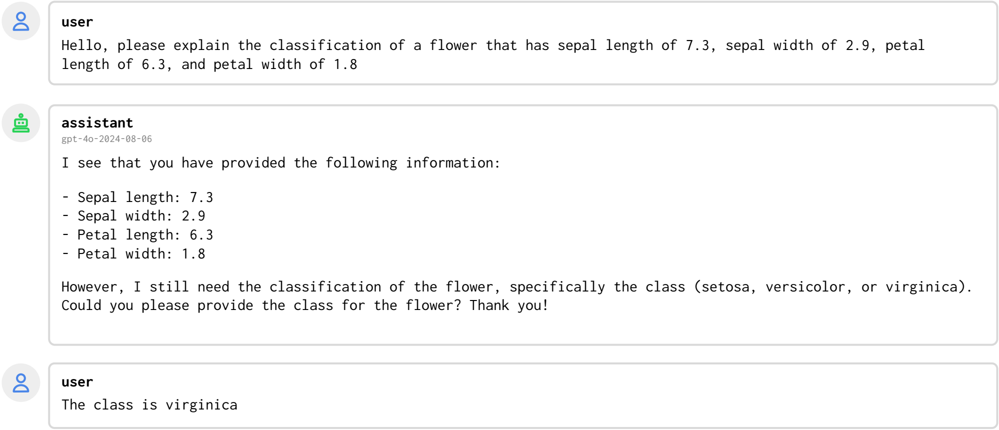
gpt-4o-2024-08-06

I see that you have provided the following information:

- Sepal length: 7.3
- Sepal width: 2.9
- Petal length: 6.3
- Petal width: 1.8

However, I still need the classification of the flower, specifically the class (setosa, versicolor, or virginica). Could you please provide the class for the flower? Thank you!

Example



user

Hello, please explain the classification of a flower that has sepal length of 7.3, sepal width of 2.9, petal length of 6.3, and petal width of 1.8

assistant

gpt-4o-2024-08-06

I see that you have provided the following information:

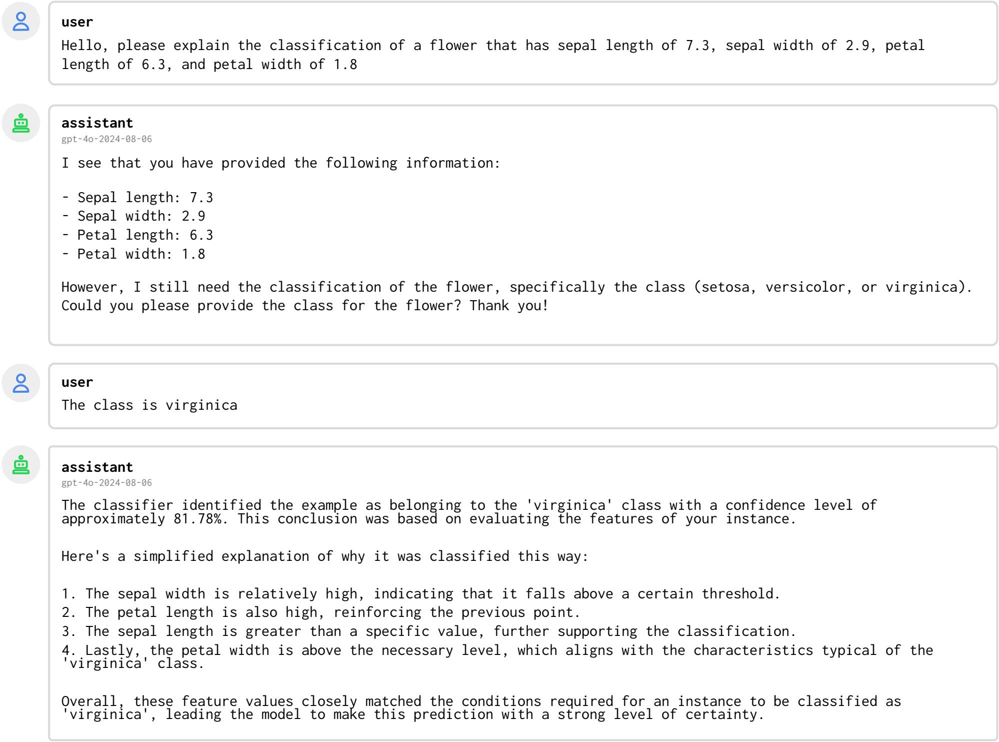
- Sepal length: 7.3
- Sepal width: 2.9
- Petal length: 6.3
- Petal width: 1.8

However, I still need the classification of the flower, specifically the class (setosa, versicolor, or virginica). Could you please provide the class for the flower? Thank you!

user

The class is virginica

Example



user

Hello, please explain the classification of a flower that has sepal length of 7.3, sepal width of 2.9, petal length of 6.3, and petal width of 1.8

assistant

gpt-4o-2024-08-06

I see that you have provided the following information:

- Sepal length: 7.3
- Sepal width: 2.9
- Petal length: 6.3
- Petal width: 1.8

However, I still need the classification of the flower, specifically the class (setosa, versicolor, or virginica). Could you please provide the class for the flower? Thank you!

user

The class is virginica

assistant

gpt-4o-2024-08-06

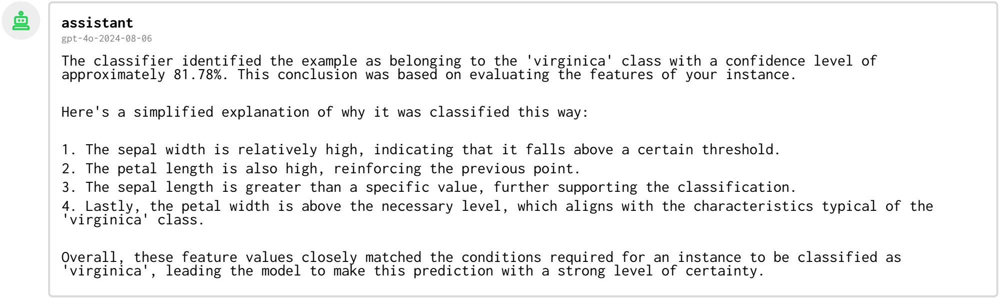
The classifier identified the example as belonging to the 'virginica' class with a confidence level of approximately 81.78%. This conclusion was based on evaluating the features of your instance.

Here's a simplified explanation of why it was classified this way:

1. The sepal width is relatively high, indicating that it falls above a certain threshold.
2. The petal length is also high, reinforcing the previous point.
3. The sepal length is greater than a specific value, further supporting the classification.
4. Lastly, the petal width is above the necessary level, which aligns with the characteristics typical of the 'virginica' class.

Overall, these feature values closely matched the conditions required for an instance to be classified as 'virginica', leading the model to make this prediction with a strong level of certainty.

Example



assistant

gpt-4o-2024-08-06

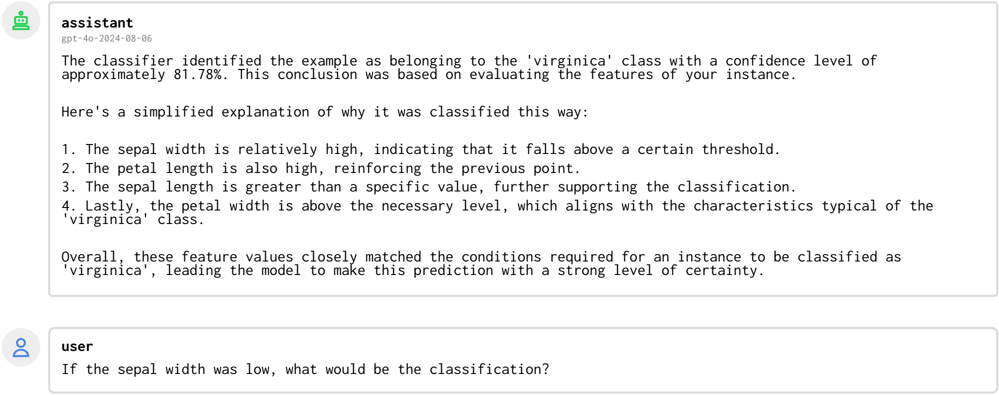
The classifier identified the example as belonging to the 'virginica' class with a confidence level of approximately 81.78%. This conclusion was based on evaluating the features of your instance.

Here's a simplified explanation of why it was classified this way:

1. The sepal width is relatively high, indicating that it falls above a certain threshold.
2. The petal length is also high, reinforcing the previous point.
3. The sepal length is greater than a specific value, further supporting the classification.
4. Lastly, the petal width is above the necessary level, which aligns with the characteristics typical of the 'virginica' class.

Overall, these feature values closely matched the conditions required for an instance to be classified as 'virginica', leading the model to make this prediction with a strong level of certainty.

Example



assistant

gpt-4o-2024-08-06

The classifier identified the example as belonging to the 'virginica' class with a confidence level of approximately 81.78%. This conclusion was based on evaluating the features of your instance.

Here's a simplified explanation of why it was classified this way:

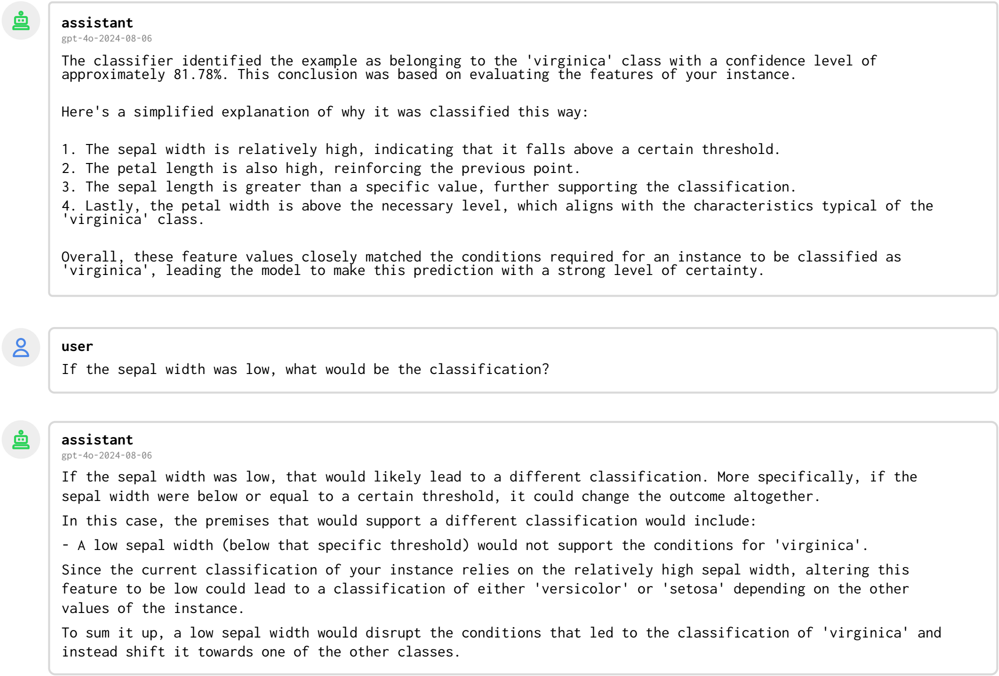
1. The sepal width is relatively high, indicating that it falls above a certain threshold.
2. The petal length is also high, reinforcing the previous point.
3. The sepal length is greater than a specific value, further supporting the classification.
4. Lastly, the petal width is above the necessary level, which aligns with the characteristics typical of the 'virginica' class.

Overall, these feature values closely matched the conditions required for an instance to be classified as 'virginica', leading the model to make this prediction with a strong level of certainty.

user

If the sepal width was low, what would be the classification?

Example



assistant

gpt-4o-2024-08-06

The classifier identified the example as belonging to the 'virginica' class with a confidence level of approximately 81.78%. This conclusion was based on evaluating the features of your instance.

Here's a simplified explanation of why it was classified this way:

1. The sepal width is relatively high, indicating that it falls above a certain threshold.
2. The petal length is also high, reinforcing the previous point.
3. The sepal length is greater than a specific value, further supporting the classification.
4. Lastly, the petal width is above the necessary level, which aligns with the characteristics typical of the 'virginica' class.

Overall, these feature values closely matched the conditions required for an instance to be classified as 'virginica', leading the model to make this prediction with a strong level of certainty.

user

If the sepal width was low, what would be the classification?

assistant

gpt-4o-2024-08-06

If the sepal width was low, that would likely lead to a different classification. More specifically, if the sepal width were below or equal to a certain threshold, it could change the outcome altogether.

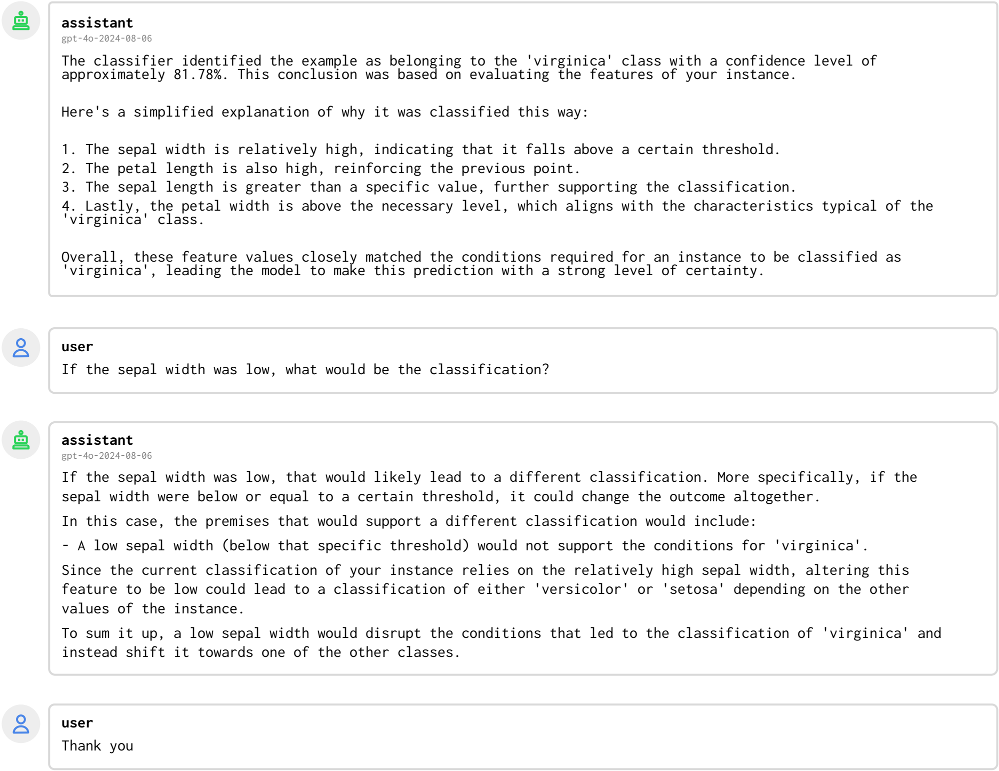
In this case, the premises that would support a different classification would include:

- A low sepal width (below that specific threshold) would not support the conditions for 'virginica'.

Since the current classification of your instance relies on the relatively high sepal width, altering this feature to be low could lead to a classification of either 'versicolor' or 'setosa' depending on the other values of the instance.

To sum it up, a low sepal width would disrupt the conditions that led to the classification of 'virginica' and instead shift it towards one of the other classes.

Example



assistant

gpt-4o-2024-08-06

The classifier identified the example as belonging to the 'virginica' class with a confidence level of approximately 81.78%. This conclusion was based on evaluating the features of your instance.

Here's a simplified explanation of why it was classified this way:

1. The sepal width is relatively high, indicating that it falls above a certain threshold.
2. The petal length is also high, reinforcing the previous point.
3. The sepal length is greater than a specific value, further supporting the classification.
4. Lastly, the petal width is above the necessary level, which aligns with the characteristics typical of the 'virginica' class.

Overall, these feature values closely matched the conditions required for an instance to be classified as 'virginica', leading the model to make this prediction with a strong level of certainty.

user

If the sepal width was low, what would be the classification?

assistant

gpt-4o-2024-08-06

If the sepal width was low, that would likely lead to a different classification. More specifically, if the sepal width were below or equal to a certain threshold, it could change the outcome altogether.

In this case, the premises that would support a different classification would include:

- A low sepal width (below that specific threshold) would not support the conditions for 'virginica'.

Since the current classification of your instance relies on the relatively high sepal width, altering this feature to be low could lead to a classification of either 'versicolor' or 'setosa' depending on the other values of the instance.

To sum it up, a low sepal width would disrupt the conditions that led to the classification of 'virginica' and instead shift it towards one of the other classes.

user

Thank you

Validation

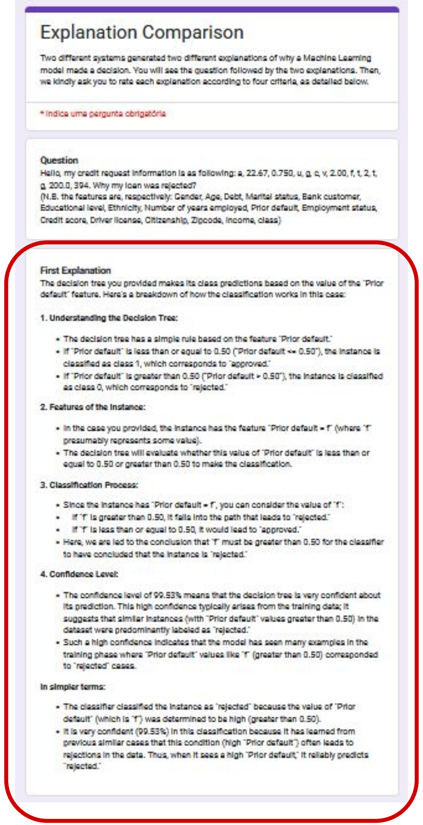
- User Study
    - 25 people

Validation

- User Study

  25 people

- Form with raw and simplified explanations



Explanation Comparison

Two different systems generated two different explanations of why a Machine Learning model made a decision. You will see the question followed by the two explanations. Then, we kindly ask you to rate each explanation according to four criteria, as detailed below.

* Indica uma pergunta obrigatória

Question

Hello, my credit request information is as following: a, 22.67, 0.750, u, g, c, v, 2.00, f, t, 2, t, g, 200.0, 394. Why my loan was rejected?
(N.B. the features are, respectively: Gender, Age, Debt, Marital status, Bank customer, Educational level, Ethnicity, Number of years employed, Prior default, Employment status, Credit score, Driver license, Citizenship, Zipcode, Income, class)

First Explanation

The decision tree you provided makes its class predictions based on the value of the "Prior default" feature. Here's a breakdown of how the classification works in this case:

1. Understanding the Decision Tree:

- The decision tree has a simple rule based on the feature "Prior default."
- If "Prior default" is less than or equal to 0.50 ("Prior default <= 0.50"), the instance is classified as class 1, which corresponds to "approved."
- If "Prior default" is greater than 0.50 ("Prior default > 0.50"), the instance is classified as class 0, which corresponds to "rejected."

2. Features of the Instance:

- In the case you provided, the instance has the feature "Prior default = f" (where "f" presumably represents some value).
- The decision tree will evaluate whether this value of "Prior default" is less than or equal to 0.50 or greater than 0.50 to make the classification.

3. Classification Process:

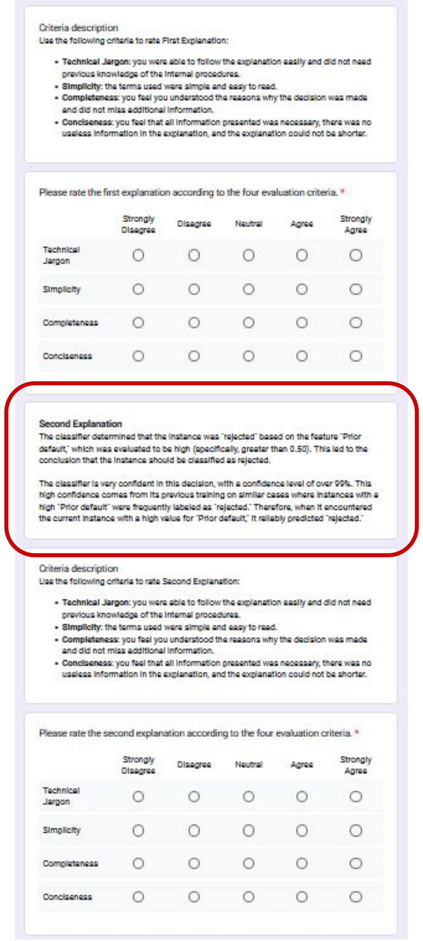
- Since the instance has "Prior default = f", you can consider the value of "f":
- If "f" is greater than 0.50, it falls into the path that leads to "rejected."
- If "f" is less than or equal to 0.50, it would lead to "approved."
- Here, we are led to the conclusion that "f" must be greater than 0.50 for the classifier to have concluded that the instance is "rejected."

4. Confidence Level:

- The confidence level of 99.53% means that the decision tree is very confident about its prediction. This high confidence typically arises from the training data; it suggests that similar instances (with "Prior default" values greater than 0.50) in the dataset were predominantly labeled as "rejected."
- Such a high confidence indicates that the model has seen many examples in the training phase where "Prior default" values like "f" (greater than 0.50) corresponded to "rejected" cases.

In simpler terms:

- The classifier classified the instance as "rejected" because the value of "Prior default" (which is "f") was determined to be high (greater than 0.50).
- It is very confident (99.53%) in this classification because it has learned from previous similar cases that this condition (high "Prior default") often leads to rejections in the data. Thus, when it sees a high "Prior default," it reliably predicts "rejected."



Criteria description
Use the following criteria to rate First Explanation:

- **Technical Jargon**: you were able to follow the explanation easily and did not need previous knowledge of the internal procedures.
- **Simplicity**: the terms used were simple and easy to read.
- **Completeness**: you feel you understood the reasons why the decision was made and did not miss additional information.
- **Conciseness**: you feel that all information presented was necessary, there was no useless information in the explanation, and the explanation could not be shorter.

Please rate the first explanation according to the four evaluation criteria. *

	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
Technical Jargon	○	○	○	○	○
Simplicity	○	○	○	○	○
Completeness	○	○	○	○	○
Conciseness	○	○	○	○	○

Second Explanation

The classifier determined that the instance was "rejected" based on the feature "Prior default," which was evaluated to be high (specifically, greater than 0.50). This led to the conclusion that the instance should be classified as rejected.

The classifier is very confident in this decision, with a confidence level of over 99%. This high confidence comes from its previous training on similar cases where instances with a high "Prior default" were frequently labeled as "rejected." Therefore, when it encountered the current instance with a high value for "Prior default," it reliably predicted "rejected."

Criteria description
Use the following criteria to rate Second Explanation:

- **Technical Jargon**: you were able to follow the explanation easily and did not need previous knowledge of the internal procedures.
- **Simplicity**: the terms used were simple and easy to read.
- **Completeness**: you feel you understood the reasons why the decision was made and did not miss additional information.
- **Conciseness**: you feel that all information presented was necessary, there was no useless information in the explanation, and the explanation could not be shorter.

Please rate the second explanation according to the four evaluation criteria. *

	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
Technical Jargon	○	○	○	○	○
Simplicity	○	○	○	○	○
Completeness	○	○	○	○	○
Conciseness	○	○	○	○	○

Validation

- User Study

  25 people

- Form with raw and simplified explanations

- Likert scale rating

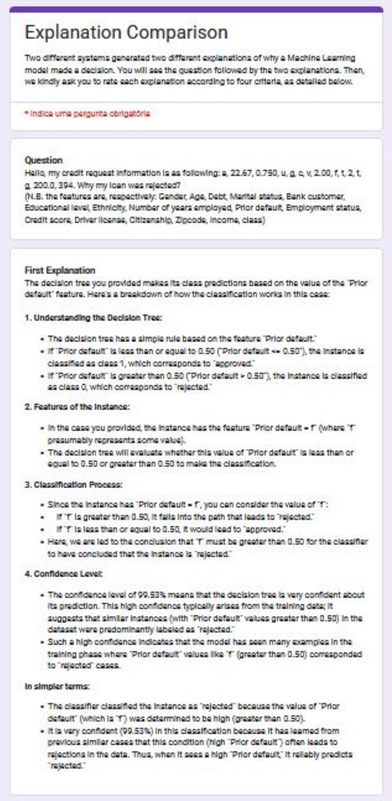
  1: Strongly Disagree

  2: Disagree

  3: Neutral

  4: Agree

  5: Strongly Agree



Explanation Comparison

Two different systems generated two different explanations of why a Machine Learning model made a decision. You will see the question followed by the two explanations. Then, we kindly ask you to rate each explanation according to four criteria, as detailed below.

* Indica uma pergunta obrigatória

Question
Hello, my credit request information is as following: a, 22.67, 0.750, u, g, c, v, 2.00, f, t, 2, t, g, 200.0, 394. Why my loan was rejected?
(N.B. the features are, respectively: Gender, Age, Debt, Marital status, Bank customer, Educational level, Ethnicity, Number of years employed, Prior default, Employment status, Credit score, Driver license, Citizenship, Zipcode, Income, class)

First Explanation
The decision tree you provided makes its class predictions based on the value of the "Prior default" feature. Here's a breakdown of how the classification works in this case:

1. Understanding the Decision Tree:

- The decision tree has a simple rule based on the feature "Prior default."
- If "Prior default" is less than or equal to 0.50 ("Prior default <= 0.50"), the instance is classified as class 1, which corresponds to "approved."
- If "Prior default" is greater than 0.50 ("Prior default > 0.50"), the instance is classified as class 0, which corresponds to "rejected."

2. Features of the Instance:

- In the case you provided, the instance has the feature "Prior default = f" (where "f" presumably represents some value).
- The decision tree will evaluate whether this value of "Prior default" is less than or equal to 0.50 or greater than 0.50 to make the classification.

3. Classification Process:

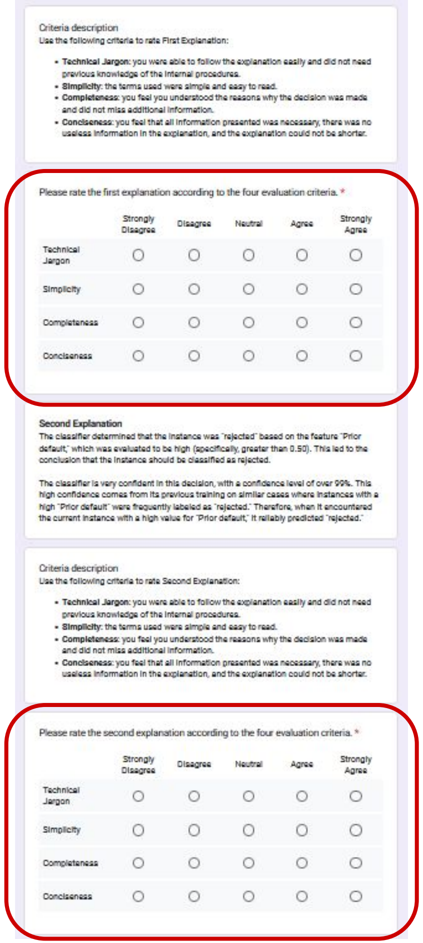
- Since the instance has "Prior default = f", you can consider the value of "f":
- If "f" is greater than 0.50, it falls into the path that leads to "rejected."
- If "f" is less than or equal to 0.50, it would lead to "approved."
- Here, we are led to the conclusion that "f" must be greater than 0.50 for the classifier to have concluded that the instance is "rejected."

4. Confidence Level:

- The confidence level of 99.53% means that the decision tree is very confident about its prediction. This high confidence typically arises from the training data; it suggests that similar instances (with "Prior default" values greater than 0.50) in the dataset were predominantly labeled as "rejected."
- Such a high confidence indicates that the model has seen many examples in the training phase where "Prior default" values like "f" (greater than 0.50) corresponded to "rejected" cases.

In simpler terms:

- The classifier classified the instance as "rejected" because the value of "Prior default" (which is "f") was determined to be high (greater than 0.50).
- It is very confident (99.53%) in this classification because it has learned from previous similar cases that this condition (high "Prior default") often leads to rejections in the data. Thus, when it sees a high "Prior default," it reliably predicts "rejected."



Criteria description
Use the following criteria to rate First Explanation:

- **Technical Jargon:** you were able to follow the explanation easily and did not need previous knowledge of the internal procedures.
- **Simplicity**: the terms used were simple and easy to read.
- **Completeness:** you feel you understood the reasons why the decision was made and did not miss additional information.
- **Conciseness:** you feel that all information presented was necessary, there was no useless information in the explanation, and the explanation could not be shorter.

Please rate the first explanation according to the four evaluation criteria. *

	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
Technical Jargon	○	○	○	○	○
Simplicity	○	○	○	○	○
Completeness	○	○	○	○	○
Conciseness	○	○	○	○	○

Second Explanation
The classifier determined that the instance was "rejected" based on the feature "Prior default," which was evaluated to be high (specifically, greater than 0.50). This led to the conclusion that the instance should be classified as rejected.

The classifier is very confident in this decision, with a confidence level of over 99%. This high confidence comes from its previous training on similar cases where instances with a high "Prior default" were frequently labeled as "rejected." Therefore, when it encountered the current instance with a high value for "Prior default," it reliably predicted "rejected."

Criteria description
Use the following criteria to rate Second Explanation:

- **Technical Jargon:** you were able to follow the explanation easily and did not need previous knowledge of the internal procedures.
- **Simplicity**: the terms used were simple and easy to read.
- **Completeness:** you feel you understood the reasons why the decision was made and did not miss additional information.
- **Conciseness:** you feel that all information presented was necessary, there was no useless information in the explanation, and the explanation could not be shorter.

Please rate the second explanation according to the four evaluation criteria. *

	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
Technical Jargon	○	○	○	○	○
Simplicity	○	○	○	○	○
Completeness	○	○	○	○	○
Conciseness	○	○	○	○	○

Evaluation Criteria

- Technical Jargon
- Simplicity
- Completeness
- Conciseness

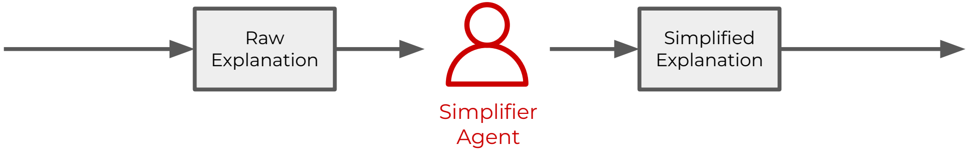
Results - User Study - Explanation Rating

Criterion	Raw Explanation	Simplified Explanation
Technical Jargon	3.1 ± 1.3	**4.0 ± 1.0**
Simplicity	2.8 ± 1.3	**4.4 ± 0.8**
Completeness	**3.4 ± 1.3**	3.1 ± 1.3
Conciseness	3.0 ± 1.2	**3.9 ± 1.2**

Results - User Study - Best Explanation

Criterion	Raw	Simplified	Both rated equally
Technical Jargon	5	**12**	8
Simplicity	1	**18**	6
Completeness	**10**	9	6
Conciseness	5	**15**	5
Total	21	**54**	25

Critic Agent Evaluation

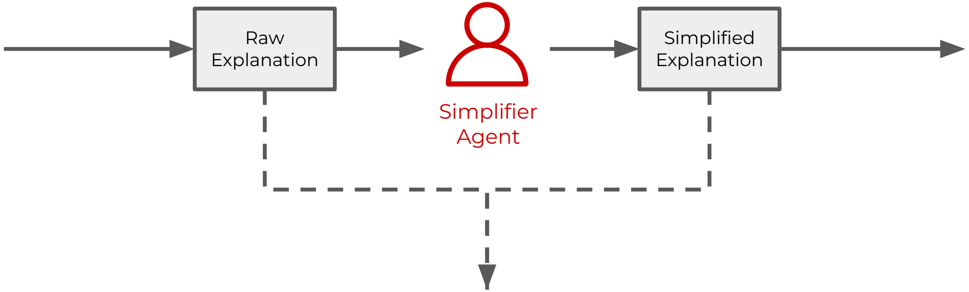


Raw Explanation

Simplifier Agent

Simplified Explanation

Critic Agent Evaluation

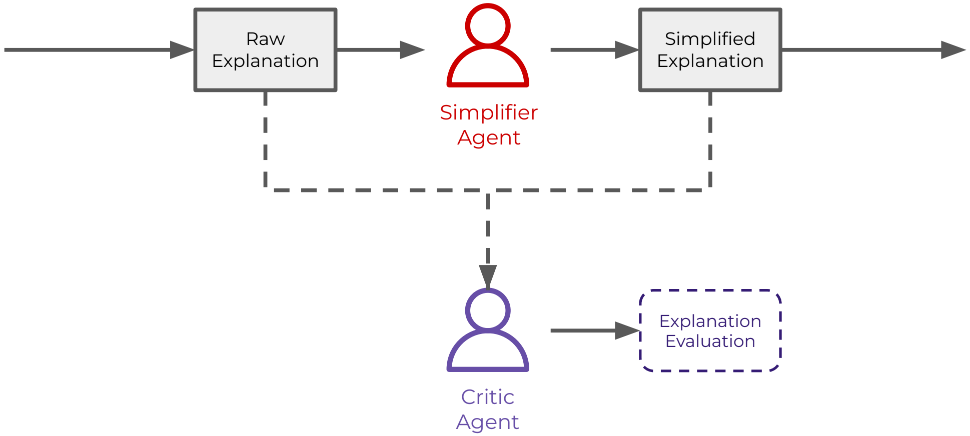


Raw Explanation

Simplifier Agent

Simplified Explanation

Critic Agent Evaluation



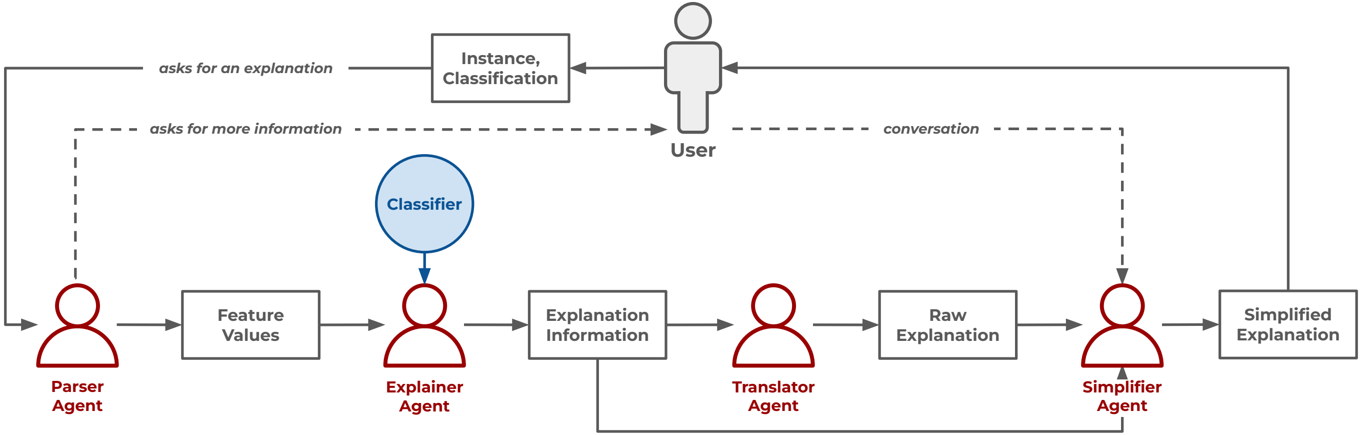
Results - Critic Agents - Explanation Rating

	GPT-4o		**Gemini 2.0**		**LLaMA 3.2**		**DeepSeek-R1**	
Criterion	**Raw**	**Sim.**	**Raw**	**Sim.**	**Raw**	**Sim.**	**Raw**	**Sim.**
Technical Jargon	2.1	**4.2**	3.7	**4.7**	2.1	**3.9**	2.9	**3.6**
Simplicity	2.0	**4.8**	3.7	**4.7**	3.9	**4.8**	2.7	**4.6**
Completeness	3.4	**4.1**	**3.9**	3.4	3.5	**4.6**	**3.8**	**3.8**
Conciseness	3.3	**4.6**	3.4	**4.6**	4.0	**4.7**	2.7	**4.4**

Results - Critic Agents - Best Explanation

	GPT-4o		**Gemini 2.0**		**LLaMA 3.2**		**DeepSeek-R1**	
Criterion	**Raw**	**Sim.**	**Raw**	**Sim.**	**Raw**	**Sim.**	**Raw**	**Sim.**
Technical Jargon	0	**25**	0	**25**	0	**25**	8	**17**
Simplicity	0	**25**	0	**25**	0	**25**	1	**24**
Completeness	**22**	3	5	**20**	8	**17**	**24**	1
Conciseness	0	**25**	0	**25**	3	**22**	2	**23**
Total	22	**78**	5	**95**	11	**89**	35	**65**

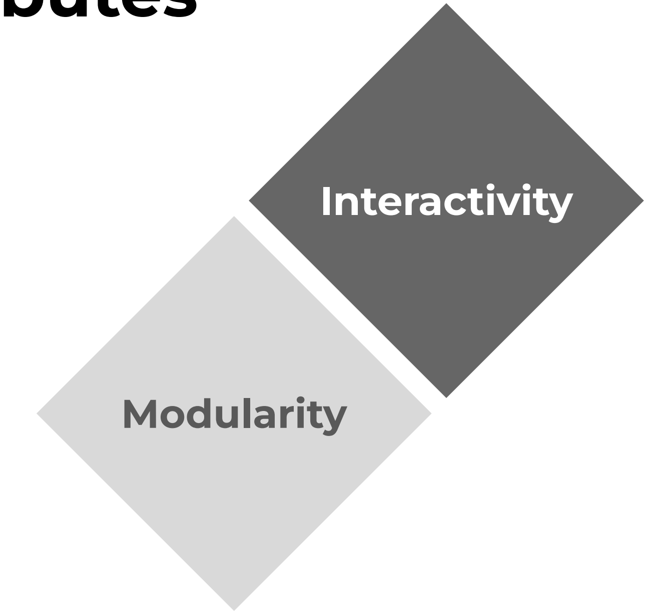
MAINLE


Instance,
Classification
asks for an explanation
User
asks for more information
conversation
Classifier
Parser
Agent
Feature
Values
Explainer
Agent
Explanation
Information
Translator
Agent
Raw
Explanation
Simplifier
Agent
Simplified
Explanation

Quality Attributes

Interactivity

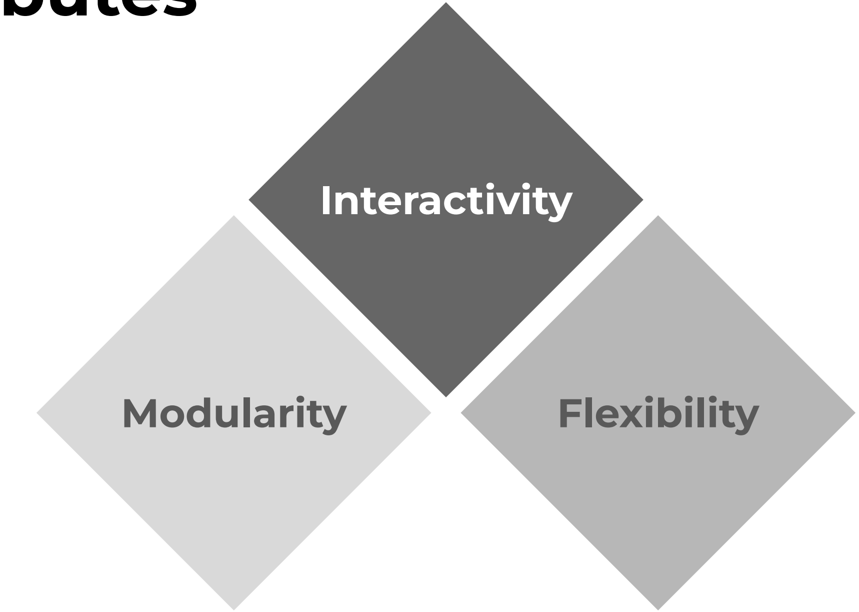
Quality Attributes



Interactivity

Modularity

Quality Attributes

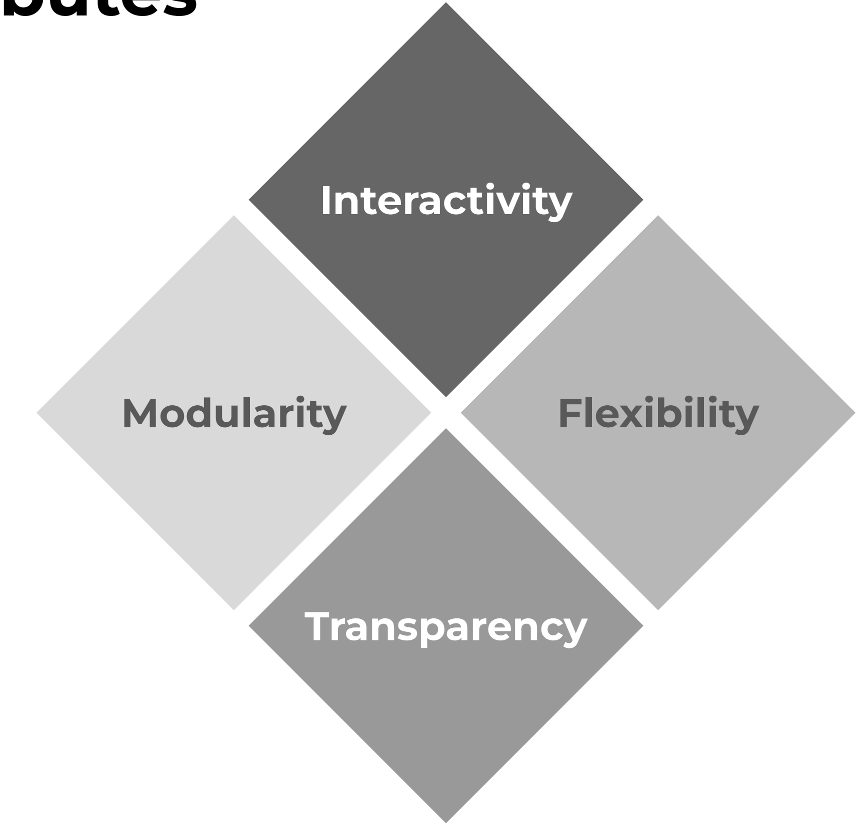


Interactivity

Modularity

Flexibility

Quality Attributes



Interactivity

Modularity

Flexibility

Transparency

Potential Risks and Concerns

1. LLM-related problems

Potential Risks and Concerns

1. LLM-related problems
2. Usage of a surrogate model

ECML PKDD 2025

MAINLE: a Multi-Agent, Interactive, Natural Language Local Explainer of Classification Tasks

Paulo Bruno Serafim, Rômulo Férrer Filho, Stenio Freitas, Gizem Gezici, Fosca Giannotti, Franco Raimondi, and Alexandre Santos

Poster Section
(Board 36)
Thursday, 18/Sep
19:00 - 21:00

Thank you!

pre-print version

github.com/paulobruno/ecml-pkdd-2025

paulobruno.github.io

paulo.desousa@gssi.it

GRAN SASSO SCIENCE INSTITUTE
SCHOOL OF ADVANCED STUDIES
Scuola Universitaria Superiore

SCUOLA NORMALE SUPERIORE

UFC

Extra slides

Contributions

1. A multi-agent architecture to generate simplified explanations for any ML classifier

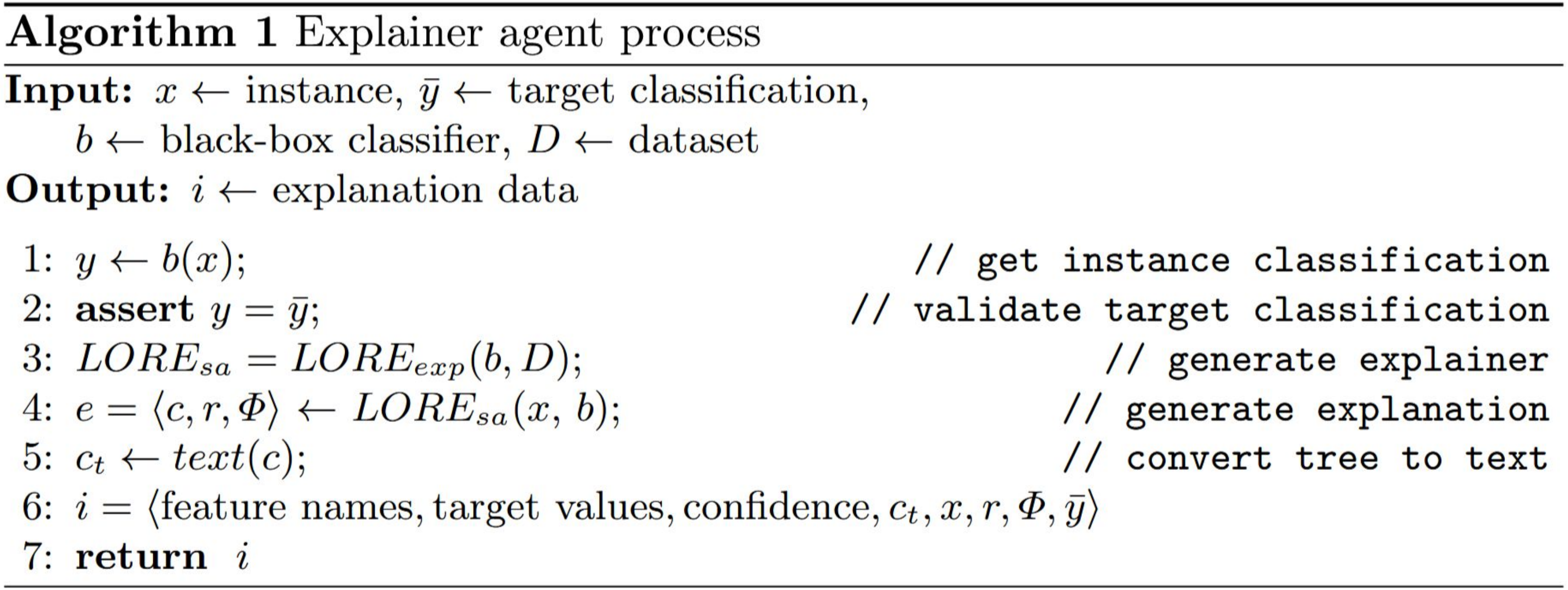
2. A process to generate natural language explanations using surrogate interpretable models for any classifier

3. Interactive explanations through conversation

4. Validation of the architecture with user studies

5. Public availability of the code

LORE Algorithm



Algorithm 1 Explainer agent process

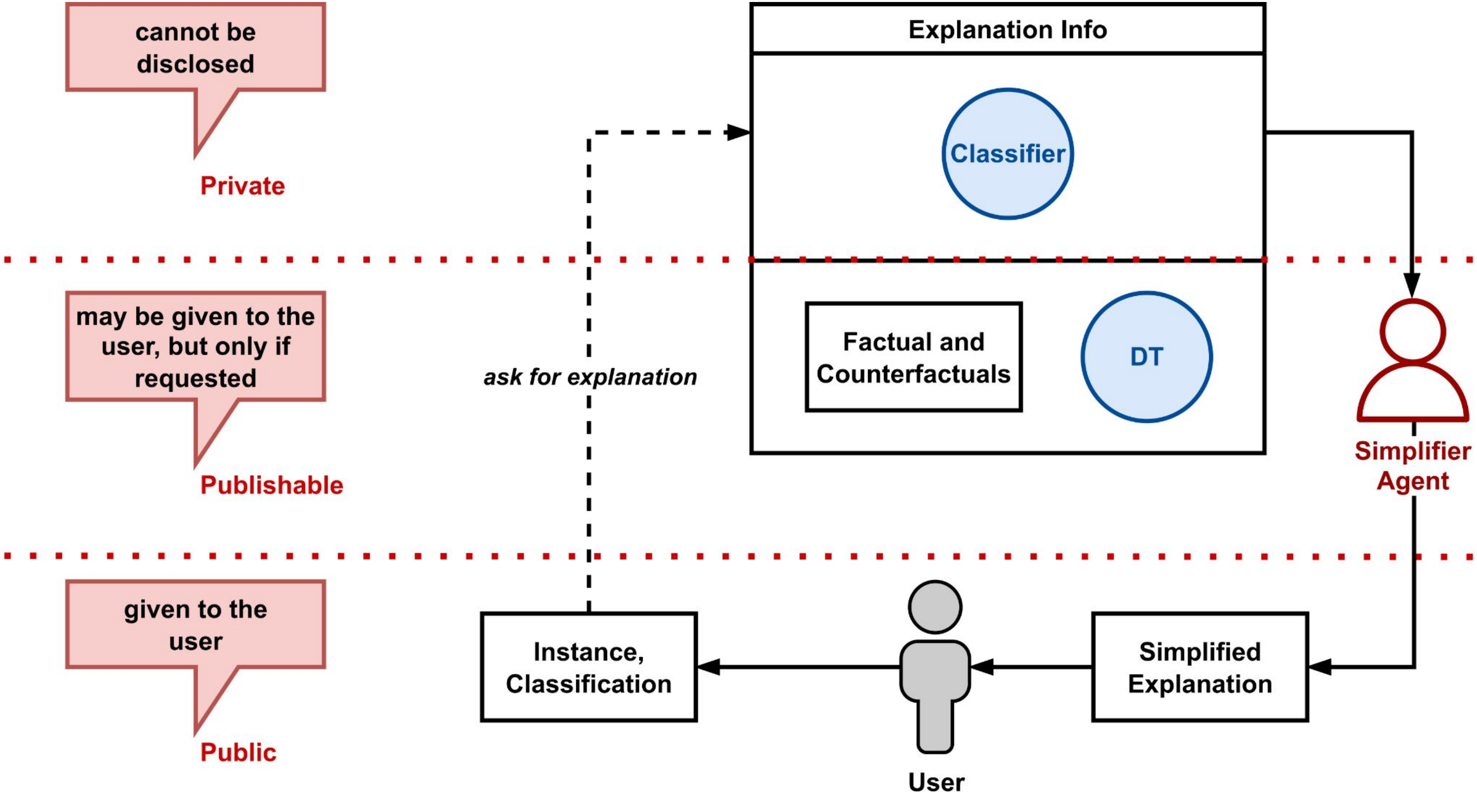
```
Input: $x \leftarrow$ instance, $\bar{y} \leftarrow$ target classification,
    $b \leftarrow$ black-box classifier, $D \leftarrow$ dataset
Output: $i \leftarrow$ explanation data

$y \leftarrow b(x)$;                                                  // get instance classification
assert $y = \bar{y}$;                                              // validate target classification
$LORE_{sa} = LORE_{exp}(b, D)$;                                       // generate explainer
$e = \langle c, r, \Phi \rangle \leftarrow LORE_{sa}(x, b)$;          // generate explanation
$c_t \leftarrow text(c)$;                                             // convert tree to text
$i = \langle$feature names, target values, confidence, $c_t, x, r, \Phi, \bar{y} \rangle$
return $i$
```

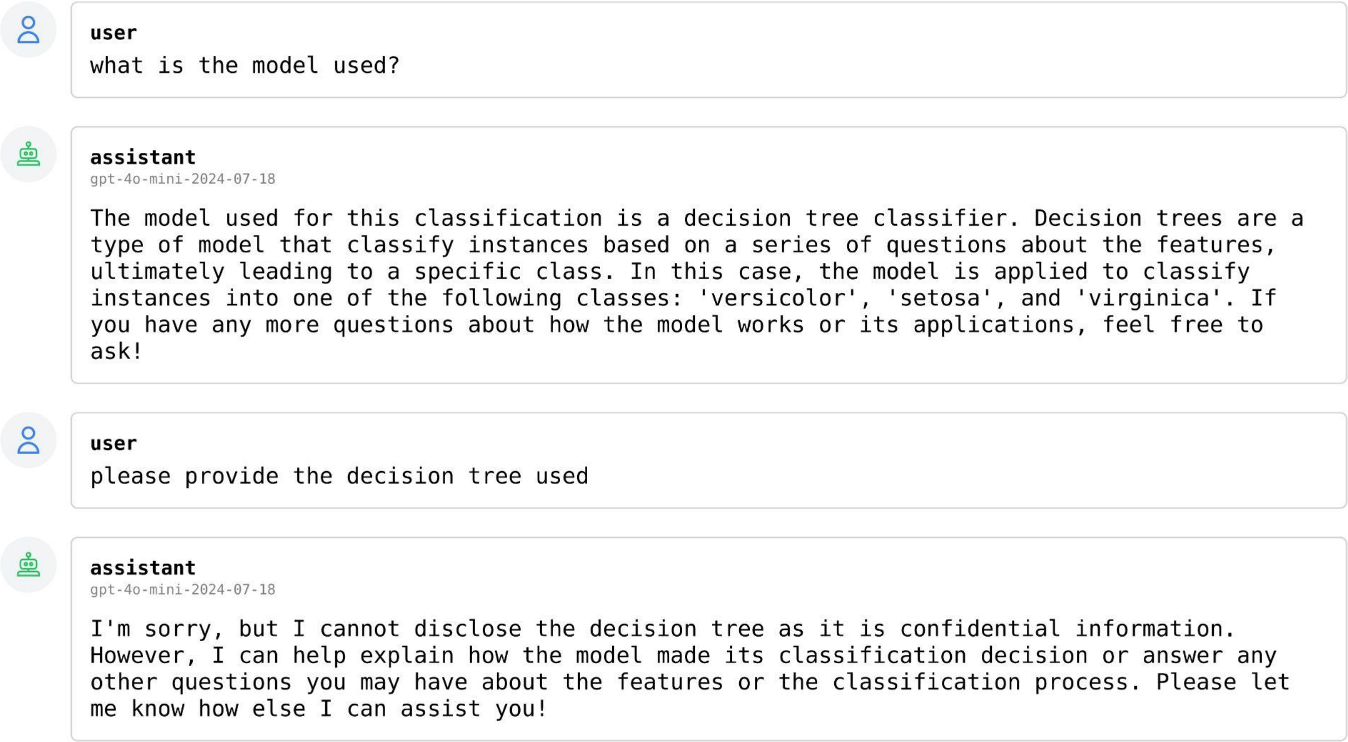
Quality Attributes

- Interactivity
- Modularity
- Flexibility
- Transparency
  - Decision transparency
  - Information transparency

Information Transparency



Information Transparency



user

what is the model used?

assistant

gpt-4o-mini-2024-07-18

The model used for this classification is a decision tree classifier. Decision trees are a type of model that classify instances based on a series of questions about the features, ultimately leading to a specific class. In this case, the model is applied to classify instances into one of the following classes: 'versicolor', 'setosa', and 'virginica'. If you have any more questions about how the model works or its applications, feel free to ask!

user

please provide the decision tree used

assistant

gpt-4o-mini-2024-07-18

I'm sorry, but I cannot disclose the decision tree as it is confidential information. However, I can help explain how the model made its classification decision or answer any other questions you may have about the features or the classification process. Please let me know how else I can assist you!

Datasets, Classifiers and LLMs

- Iris Flower
- Breast Cancer
- Wine
- Adult Income
- Credit Approval

- Random Forest
- Gradient Boosting
- MLP

- Gpt-4o
- Gemini 2.0
- Llama 3.2
- DeepSeek-R1

Non-LLM Translator

Raw Explanation	Simplified Explanation
The instance is classified as virginica because sepal length (cm) is greater than 6.96 and sepal width (cm) is 2.09 and petal length (cm) is greater than 2.26 and petal width (cm) is greater than 1.08.	The instance was classified as ‘virginica’ based on the following features: the sepal length is high, the sepal width is moderate, and the petal width and petal length are also high. This combination of features strongly aligns with what is typically seen in the ‘virginica’ class of flowers. The model is quite confident in this classification, with a confidence level of about 81%. In simpler terms, the measurements show characteristics that are distinctly associated with the ‘virginica’ variety, leading to its classification.

Questions from the audience

Here are the two questions asked by the audience during the presentation and their summarized answers.

1. **Did you ask the users if they could understand the explanation and not only to rate them?**
   a. No, because that would require users to have domain knowledge about the dataset, which we did not want to require, since we were evaluating MAINLE from the perspective of non-expert users.
2. **Which LLM did you use for the simplifier and how did you implement it to simplify an explanation?**
   a. (I showed the extra slide with the LLMs used) We created a structured prompt and passed it as system instructions for the LLMs that accept system prompts or in the question for the LLMs that do not accept them. All prompts are publicly available in the repository.