XARF: Explanatory Argumentation Rule-Based Framework

Hugo Eduardo Sanches[1][a], Ayslan Trevizan Possebom[1][b] and Linnyer Beatrys Ruiz Aylon[2][c]
[1]*Manna Research Team, State University of Maringá, Brazil*
[2]*Federal Institute of Parana, Paranavaí, Brazil*

Keywords: Argumentation Framework, Explainable Artificial Intelligence, Machine Learning.

Abstract: This paper introduces the Explanatory Argumentation Rule-based Framework (XARF), a new approach in Explainable Artificial Intelligence (XAI) designed to provide clear and understandable explanations for machine learning predictions and classifications. By integrating a rule-based system with argumentation theory, XARF elucidates the reasoning behind machine learning outcomes, offering a transparent view into the otherwise opaque processes of these models. The core of XARF lies in its innovative utilization of the apriori algorithm for mining rules from datasets and using them to form the foundation of arguments. XARF further innovates by detailing a unique methodology for establishing attack relations between arguments, allowing for the construction of a robust argumentation structure. To validate the effectiveness and versatility of XARF, this study examines its application across seven distinct machine learning algorithms, utilizing two different datasets: a basic Boolean dataset for demonstrating fundamental concepts and methodologies of the framework, and the classic Iris dataset to illustrate its applicability to more complex scenarios. The results highlight the capability of XARF to generate transparent, rule-based explanations for a variety of machine learning models.

1 INTRODUCTION

Explainable Artificial Intelligence (XAI) aims to make AI systems more transparent, understandable, and trustworthy. As AI becomes integrated into daily life and critical decision-making, the need for interpretable systems increases. XAI addresses the gap between advanced AI capabilities and the human need for clarity, accountability, and ethical assurance (Gianini and Portier, 2024). An argumentation framework (AF) is a fundamental tool in artificial intelligence and computational logic, used to formalize and analyze argumentation processes. AF models, evaluates, and resolves conflicts between arguments, drawing on principles of classical and informal logic (Vassiliades and Patkos, 2021). In the context of XAI, the development of methods to interpret machine learning predictions is essential. This paper introduces the Explanatory Argumentation Rule-Based Framework (XARF), an innovative approach leveraging argumentation frameworks to explain machine learning decisions. XARF uses a rule-based mechanism rooted in explanatory argumentation to provide clear and understandable explanations. At the core of XARF is the use of rules derived from datasets via the apriori algorithm. These rules, which consist of premises and conclusions, form the backbone of the argumentation framework. XARF constructs arguments and defines attack relations among them, enabling structured reasoning that aligns with the extensions of the argumentation framework such as grounded and preferred semantics. The adaptability of XARF is demonstrated through its application to seven machine learning algorithms across two datasets. This highlights its effectiveness in generating transparent explanations for diverse models. By combining rule-based insights from the apriori algorithm with argumentation theory, XARF enhances transparency and understanding of machine learning models. This paper details the architecture, operational mechanisms, and empirical evidence of the ability of XARF to bridge the interpretability gap in machine learning, contributing a significant advancement to XAI. This work was conducted in the context of the Manna_Team. This article begins with a literature review, covering AF, XAI, and related work. It then details the methodology behind XARF, explaining its application to datasets and machine learning algorithms. The results section analyzes the effectiveness of the framework, followed

[a] https://orcid.org/0000-0002-4450-1865
[b] https://orcid.org/0000-0002-1347-5852
[c] https://orcid.org/0000-0002-4456-6829

Sanches, H. E., Possebom, A. T. and Aylon, L. B. R.
XARF: Explanatory Argumentation Rule-Based Framework.
DOI: 10.5220/0013193200003929
In *Proceedings of the 27th International Conference on Enterprise Information Systems (ICEIS 2025) - Volume 1*, pages 683-690
ISBN: 978-989-758-749-8; ISSN: 2184-4992

by a conclusion summarizing findings and potential future work.

2 BACKGROUND REVIEW

2.1 Argumentation

Argumentation is an interdisciplinary field that has attracted significant interest within artificial intelligence, particularly in reasoning and multi-agent systems, where agents evaluate arguments and reach conclusions collectively. This has spurred both theoretical and practical research within computer science (Lu, 2018). Therefore, the knowledge area known as computational argumentation was concisely developed and, furthermore, laid the groundwork for considering argumentation as a discipline within artificial intelligence, especially with the model proposed by Dung (Dung, 1995) defining what is known as Abstract Argumentation. Since Dung's model, various semantics have been developed to adapt the framework for different problems and applications. Argumentation semantics formally defines protocols that dictates all the rules for evaluating arguments (Jha and Toni, 2020). The extension-based semantics approach focuses on creating subsets (extensions) of arguments that can coexist (Amgoud and Ben-Naim, 2018).

2.2 XAI

Explainable Artificial Intelligence (XAI) emphasizes transparency and understandability in AI systems, addressing the "black box" nature of models to foster trust and accountability (Samek and Müller, 2021). Reviews by (Arrieta and Herrera, 2020) and (Linardatos and Kotsiantis, 2021) highlight the evolution of XAI, from simple methods to advanced techniques for deep learning models. XAI applications span sectors like healthcare, where it enhances diagnostic transparency, and finance, where it clarifies automated trading and risk assessment (Fan and Wang, 2021). For instance, XAI has improved trust in medical AI systems by explaining predictions related to patient outcomes (Tjoa and Guan, 2020).

2.3 Related Work

Argumentative systems have been applied to explain machine learning algorithms. For example, the DAX framework (Deep Argumentative Explanations) integrates computational argumentation with neural networks to enhance interpretability and explainability (Albini and Tintarev, 2020). Similarly, Quantitative Argumentation Frameworks (QAF) provide benefits for neural network explainability using a different methodology (Potyka, 2021). For Random Forest algorithms, Bipolar Argumentation Graphs (BAP) have been used to justify and explain decisions (Potyka and Toni, 2022). In another study (Achilleos and Pattichis, 2020), ML algorithms like Random Forest and Decision Trees were applied to brain MRI images to distinguish between normal controls and Alzheimer's Disease cases. Data-Empowered Argumentation (DEAr) offers an approach to generate explainable ML predictions by structuring arguments through data relationships (Cocarascu and Toni, 2020). Similarly, the EVAX framework (Everyday Argumentative Explanations) generates user-accessible explanations for AI decisions by mirroring natural human reasoning, demonstrating its versatility across four machine learning models (Van Lente and Sarkadi, 2022).

3 METHODOLOGY

This research was conducted using two distinct datasets, employing a consistent methodological approach with necessary adaptations tailored to each dataset. The first dataset, a Boolean dataset, comprises three attributes (Sun, Wind and Sore Knee), a binary class activity (which is 1 for surfing and 0 for fishing), and ten instances, serving as a foundation to elucidate the concepts and methodology integral to the XARF framework and can be found in Table I. The second dataset, the well known Iris dataset, is characterized by four numerical attributes and encompasses three distinct class labels representing different species of the Iris flower and 150 instances.

Table 1: Boolean dataset D. Source: (Dondio, 2021).

Sun	Wind	Sore Knee	Activity
1	1	1	0
1	1	0	1
1	0	0	0
1	0	0	0
0	1	1	0
0	0	1	0
0	1	0	1
0	0	0	0
0	1	1	1
0	1	0	1

3.1 Definitions of Dataset, Attributes, and Elements

Dataset (D)

Let D be a dataset consisting of tuples $(X_1, X_2, \ldots, X_n, CX)$, where each X_i represents an attribute of the data with i denoting the specific attribute index, and CX is the class attribute. D can

also be defined as a set of elements $(E_1, E_2, \ldots, E_n)$ where each E_i is an element.

Attribute (X)

Each X_i represents an attribute of the data where i denotes the specific attribute index.

Class Attribute (CX)

The class attribute CX indicates the outcome or class label for each tuple in D. It is a special type of attribute used to distinguish the categories or classes to which data instances belong. This is the target variable in the classification task. Class attributes can be represented as CX_j where j represents different classes. For example, in a dataset for iris plant classification, the class attribute could represent the type of iris plant, such as Setosa, Versicolour, or Virginica.

Element (E)

An element refers to any attribute, including class attributes. Thus, it can be represented as E_k, where k can refer to any attribute or class attribute index, and $E \in X \cup CX$. **Types of Attributes/Elements.**

- **Boolean Attributes/Elements.** A Boolean attribute X_i can have a true or false value, represented as X_i^1 (true) and X_i^2 (false).
- **Numerical/Categorical Attributes/Elements.** A numerical or categorical attribute X_i is discretized (divided into bins or encoded categories), for example, X_i^1, X_i^2, X_i^3, etc., representing different ranges of values or categories.

Preprocessing and Application of the Apriori Algorithm

Transformation for Boolean Attributes. Each attribute X_i is expanded into two columns representing the attribute X_i^1 (e.g., Sun) and its negation X_i^2 (no Sun), thus capturing both presence and absence. This dichotomy is crucial for constructing comprehensive association rules that account for both positive and negative associations.

Discretizing and Encoding for Numerical Attributes. Continuous attributes are discretized into bins. This step simplifies the data structure and enables the application of the Apriori algorithm, which typically operates on categorical data. For instance, a numerical attribute like petal length in the Iris dataset is divided into bins - such as $X_1^1 = [0-2.5)$, $X_1^2 = [2.5-5.0)$, $X_1^3 = [5.0-7.5)$, and so on - enhancing the detail and interpretability of the association rules derived.

Preprocessed Dataset (D'). This dataset, D', is formed by applying preprocessing steps necessary for the Apriori algorithm, including binning, encoding, and handling of Boolean attributes and elements as described previously.

3.2 Application of the Apriori Algorithm and Argumentation Framework

Apriori Algorithm

Following the preprocessing phase, the Apriori algorithm is applied to generate association rules from the dataset, thereby forming the basis for argument creation within the argumentation framework. The Apriori algorithm is a classic algorithm used to extract frequent itemsets from a dataset and derive association rules. It operates on a dataset D', producing a set of rules R, each of the form Ant $\Rightarrow$ Con where Ant (antecedent) and Con (consequent) are itemsets derived from D' and both are formed by subsets of elements $\text{Ant}, \text{Con} \subseteq E$.

Definition of Arguments

In the context of XARF, an argument arg is formed by premises and conclusions which are both structured sets of elements E derived from association rules where Premise (P) is the antecedent Ant of an association rule and Conclusion (C) is the consequent Con of an association rule. An argument arg is a pair (P,C) formed by a premise leading to a conclusion $\text{Arg} = (P,C)$. Like Ant and Con, $P,C \subseteq E$.

Definition of Argumentation Framework (AF)

An argumentation framework can be formally defined as a pair $\text{AF} = (\text{Ar}, \text{att})$, where:

- Ar is a set of arguments derived from the association rules.
- $\text{att} \subseteq \text{Ar} \times \text{Ar}$ is the attack relation among these arguments, defining which arguments attack others based on defined rules (e.g., conflicting premises and conclusions).

For the Boolean dataset, we provide a selection of randomly chosen examples of the arguments generated: arg1, Premise: 'Fishing', Conclusion: 'Sun' arg9, Premise: 'Sore Knee', Conclusion: 'Fishing' arg12, Premise: 'No Sun', 'Wind', Conclusion: 'Surf' arg16, Premise: 'Wind', 'Good Knee', Conclusion: 'Surf' For instance, arg12 implies that if there is no sun and there is wind, the activity is most likely to

be surfing. As arg = (P, C) argument 12 is mathematically described as:

$$\text{arg}_{12} = \left(P\{X_1^2 \text{ - No Sun}, X_2^1 \text{ - Wind}\}, C\{CX_1 \text{- Surf}\}\right).$$

Similarly, for the Iris dataset, we present a subset of randomly selected examples of the arguments produced: arg2, Premise: 'petal_length_bin_(4.0, 5.0]', Conclusion: 'species_versicolor' arg5, Premise: 'petal_width_bin_(0.1, 0.5]', 'petal_length_bin_(1.0, 2.0]', Conclusion: 'species_setosa'

Support and Minimum Support Value. The support of a rule, and the minimum support of a rule, which are critical in the Apriori algorithm, are defined as:

$$\text{Support}(R) = \text{Probability}(Ant \cap Con)$$

$$\text{Support}(R) \geq \text{min\_support}$$

where Probability denotes the probability of occurrence within the dataset, and min_support is a threshold value determining which rules are significant enough to consider. For example, a minimum support of 0.75 means only the rules found in at least 75% of the data instances are considered. This introduces a significant trade-off: a higher minimum support value yields fewer arguments, leading to more generalized yet precise explanations, whereas a lower value enhances detail at the cost of increased computational complexity. Several minimum support values were tested, however, for the purposes of the research documented in this paper, minimum support values of 0.2 and 0.3 were selected for the Boolean and Iris datasets, respectively. This selection was to optimize the balance between the granularity of the explanations generated and the computational efficiency of the Explanatory Argumentation Rule-based Framework (XARF). Other values can be used without changing the concepts and the main results presented in the paper.

3.3 Establishing Attack Relations Within the Argumentation Framework

For the arguments derived from association rules to integrate into an argumentation framework, it is imperative to define attack relations among them. As previously mentioned, both premises and conclusions can comprise multiple elements. However, for clarity, the exemplification of attack rules will consider premises and conclusions consisting of a singular element each.

1- Mutual Attack (Based on Opposite Conclusions Within the Same Attribute). For all arg$i$ = $((Xi^a),(X_k^b))$ and arg$j$ = $((Xi^a),(X_k^c)) \in$ Ar, where a, b, c, i, j, k are indices representing values and $b \neq c$, $i \neq k$, we have att$(\text{arg}_i, \text{arg}_j)$ and att$(\text{arg}_j, \text{arg}_i)$. This relation highlights a mutual attack when two arguments share the same premise attribute values for X_i but have conclusions with different values for the attribute X_k, such as X_k^b and X_k^c, where these values are directly opposite (e.g., true and false for Boolean attributes). For instance, if arg1 = $(X_1^1; X_2^1)$ and arg2 = $(X_1^1; X_2^2)$ there is a mutual attack between arg1 and arg2, but notice that, if arg1 = $(X_1^1; X_2^1)$ and arg3 = $(X_1^1; X_3^1)$ there are no attacks between these arguments.

2- Mutual Attack (Based on Opposite Premises within the Same Attribute). For all arg$i$ = $((Xk^b),(X_j^a))$ and arg$j$ = $((Xk^c),(X_j^a)) \in$ Ar, where $b \neq c$, we have att$(\text{arg}_i, \text{arg}_j)$ and att$(\text{arg}_j, \text{arg}_i)$. This specifies a mutual attack when two arguments share the same conclusion but have premises that include different values of the same attribute X_k, such as X_k^b and X_k^c. For instance, if arg1 = $(X_1^1; X_2^1)$ and arg2 = $(X_1^2; X_2^1)$ there is a mutual attack between arg1 and arg2, but notice that, if arg1 = $(X_1^1; X_2^1)$ and arg3 = $(X_3^1; X_2^1)$ there are no attacks between these arguments.

3- Single Direction Attack (Based on Conclusion-Premise Attribute Disagreement). For all arg$i$ = $((Xk^a),(X_j^b))$ and arg$j$ = $((Xj^c),(X_k^a)) \in$ Ar, where $b \neq c$, we have att$(\text{arg}_i, \text{arg}_j)$. This details a directed attack where the conclusion of arg_i containing X_j^b directly conflicts with the premise of arg_j containing X_j^c, and both arguments share another attribute X_k^a that ties them together. For instance, if arg1 = $(X_1^1; X_2^1)$ and arg2 = $(X_2^2; X_1^1)$ there is a single direction attack between arg1 and arg2.

4- Mutual Attack (Based on Different Class Attributes). For all arg$i$ = $((Xk^a),(CX_m))$ and arg$j$ = $((Xk^a),(CX_n)) \in$ Ar, where $m \neq n$, we have att$(\text{arg}_i, \text{arg}_j)$ and att$(\text{arg}_j, \text{arg}_i)$. This rule accounts for a mutual attack based on differing class attributes CX_m and CX_n where $m \neq n$, occurring despite sharing the same attribute Xk^a in their premises or conclusions. For instance, if arg1 = $(X_1^1; CX_1)$ and arg2 = $(X_2^2; CX_2)$ there is a mutual attack between arg1 and arg2. In constructing the argumentation framework (AF), four attack rules are defined as the foundation of the methodology. Rules 1 and 3 correspond to classical rebuttal and undercut attack types. Attack rule 2 enhances explanation consistency, reduces argument redundancy, and lowers computational complexity by simplifying the system architecture through a higher number of attacks. Attack rule 4, combined with the mapping procedure, ensures model class consistency by preventing the coexistence of class-conflicting ar-

guments. In parallel with the development of association rules and argumentation frameworks, machine learning (ML) predictive models are utilized. Consistent with XARF's objective of employing a black-box methodology, any ML model capable of class prediction can generate explanations within XARF. This study employs seven widely used ML algorithms: decision trees, random forests, k-nearest neighbors (KNN), neural networks, logistic regression, support vector machines (SVM), and naive Bayes. Upon the establishment of the AF, the next step is to apply extensions, the selection of which is contingent upon the dataset context. Extensions can vary from being skeptical, which may yield fewer but more precise explanations, to naive, potentially resulting in a greater quantity of arguments and explanations, albeit with potentially less precision. In this research, the completed, preferred, and grounded extensions were evaluated. The grounded extension yielded negligible valid outcomes and was consequently considered unsuitable for these datasets. On the other hand, the completed extensions were rejected due to their propensity for generating redundant extensions. Thus, for the purposes of this paper, the preferred extension was selected.

3.4 Mapping

The mapping procedure constitutes a critical component of our methodology, systematically associating each argument of every extension with the current query (the input data for the prediction model) and the predicted class. This procedure assigns a score to every argument and, by extension, to every extension, thereby identifying the combination that provides the most cogent explanation.

Definitions

- **Query** (Q)**:** the current input data for which a prediction is made.
- **Predicted Class** (C_p)**:** the class predicted by the machine learning model for the query Q.

Scoring of Arguments

Each argument arg is evaluated based on its alignment with the query Q and the predicted class C_p. The score of an argument arg can be denoted as $\text{score}(\text{arg})$.

- **1- Complete Match with Query**

  $$\text{score}(\text{arg}) += 1 \quad \text{if } P \subseteq Q \text{ and } C \subseteq Q$$

  This rule awards a point if all elements of both premises (P) and conclusions (C) of the argument match elements in the query.

- **2- Extra Elements in Premises/Conclusions**

  $$\text{score}(\text{arg}) += |\text{elements}(\text{arg}) - 1| \quad \text{if rule 1 is 1}$$

  This rule awards additional points for each element beyond the first in the premises or conclusions, conditional on rule 1 being positive (1).

- **3- Alignment with Predicted Class**

  $$\text{score}(\text{arg}) += 1 \quad \text{if } C_p \in C$$

  This rule gives an extra point if the predicted class is part of the argument's conclusion.

- **4- Disagreement with Predicted Class**

  $\text{score}(\text{arg}) = 0$ if exists a C_x in P or C and $C_x \neq C_p$

  The score is set to zero if any class attribute in the premises or conclusions conflicts with the predicted class.

Extension Scoring and Selection

The score of an extension ext, which is a set of arguments, is the sum of the scores of its arguments:

$$\text{score}(\text{ext}) = \sum_{\text{arg} \in \text{ext}} \text{score}(\text{arg})$$

The goal is to identify the extension ext^* that maximizes $\text{score}(\text{ext})$ while providing the most relevant and understandable explanation:

$$\text{ext}^* = \arg\max_{\text{ext}} \text{score}(\text{ext})$$

Following the application of the mapping procedure to AF extensions, with thorough consideration of the ML predicted class, the system provides the explanation behind the ML's class prediction relative to the query. A comprehensive summary of the methodology is depicted in the enclosed chart found in (Figure 1), offering a view of the procedural framework and its implementation.

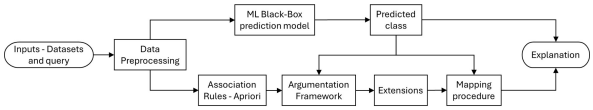

Figure 1: Methodology of XARF.

4 RESULTS AND ANALYSIS

This section presents the explanations generated by the XARF framework based on various input parameters. The focus is not on evaluating machine learning (ML) models' performance metrics such as accuracy or precision but on assessing the quality and relevance of explanations produced by XARF.

Examples are provided across different scenarios, classes, and datasets. Consider the query with input parameters [Sun: 1, Wind: 1, Sore Knee: 0], or, for a more detailed representation [Sun: 1, No Sun: 0, Wind: 1, No Wind: 0, Sore Knee: 0, Good Knee: 1]. For this query, all seven algorithms unanimously predicted class 1, representing the Surf activity. Given the size of the database, only two extensions were identified as preferred extensions. Among these, one extension had a uniform score of 1 for its arguments, whereas the other achieved a score of 10 for all its arguments. The standout argument was Argument 16, which is arg16, Premise: 'Wind', 'Good Knee', Conclusion: 'Surf', achieving a score of 3. This score was attributed to its perfect alignment with the query (+1), the presence of two elements in the premise (+1), and its conclusion matching the predicted class (Surf). This argument logically correlates with the database, indicating that windy conditions and the absence of knee soreness, rather than sunny weather, influenced the ML algorithm's prediction favoring Surf as the activity. Had the ML classification model predicted Fishing for the same query, a completely different set of arguments and extensions would have emerged, such as Argument 1, with a premise of 'Fishing' leading to a conclusion of 'Sun', thereby attributing the sunny condition as a decisive factor in predicting Fishing, according to XARF's explanation. Examining a query that elicited split predictions from the ML classifiers [Sun: 0, Wind: 1, Sore Knee: 1], or more succinctly: [Sun: 0, No Sun: 1, Wind: 1, No Wind: 0, Sore Knee: 1, Good Knee: 0]. The majority of classifiers (5 out of 7) favored class 0 (Fishing), while Random Forest and Naive Bayes opted for class 1 (Surf). For models predicting Fishing, the most robust extension featured a score of 5, with the higher score argument, arg9, Premise: 'Sore Knee', Conclusion: 'Fishing' with a score 2, indicating that a sore knee is a deterrence from surfing. Additional arguments in this extension included correlations between the absence of sun and wind, and a sore knee, further supporting the Fishing prediction. Conversely, for models predicting Surfing, the leading extension scored 13, highlighted by arg12, Premise: 'No Sun', 'Wind', Conclusion: 'Surf'. This implies that the lack of sunshine combined with windy conditions were considered significant by the classifiers for a Surf prediction. These explanations, coherent with both the database content and the predicted classes, underscore the capability of XARF to generate plausible explanations, even when classifiers diverge in their predictions for the same query. Relating to the experiments in the Iris dataset, upon the implementation of the Apriori algorithm and the formulation of attack relations, the argumentation framework (AF) for the Iris dataset was meticulously constructed. Analogously to the Boolean dataset, we herein exhibit examples of explanations generated by XARF across different scenarios. Consider a query with the following characteristics: Sepal length (cm): 5.4, Sepal width (cm): 3.7, Petal length (cm): 1.5, Petal width (cm): 0.2. For this dataset, all seven ML predictors accurately classified the query as class 0, corresponding to the Setosa species. Among the extensions evaluated, one with a score of 12 was selected, prominently featuring Argument 5 which is arg5, Premise: 'petal_width_bin_(0.1, 0.5]', 'petal_length_bin_(1.0, 2.0]', Conclusion: 'species_setosa'. This argument, which aligns with two premises and the conclusion being the predicted class, received a score of 3. It underscores the importance of both petal length and width in determining the Setosa classification. Another query is examined: Sepal length (cm): 7.0, Sepal width (cm): 3.2, Petal length (cm): 4.7, Petal width (cm): 1.4. Here, all algorithms concurred on class 1, Versicolor. The chosen extension scored 4, highlighting two arguments, Argument 2 and Argument 3, as equally significant: arg2, Premise: 'petal_length_bin_(4.0, 5.0]', Conclusion: 'species_versicolor' arg3, Premise: 'petal_width_bin_(1.0, 1.5]', Conclusion: 'species_versicolor' Both arguments are consistent with the query and show that petal sizes were the most important attributes for the decision of ML to assign class Versicolor. A contentious example involves the query: Sepal length (cm): 5.9, Sepal width (cm): 3.2, Petal length (cm): 4.8, Petal width (cm): 1.8. Here, a split in predictions occurred: Naive Bayes, Logistic Regression, KNN, and Neural Networks opted for class 2, while Decision Tree, Random Forest, and SVM selected class 1. For predictions of class 1, XARF highlighted Argument 2, which is arg2, Premise: 'petal_length_bin_(4.0, 5.0]', Conclusion: 'species_versicolor', as the sole explanatory factor with a score of 2. Conversely, for class 2 predictions, the framework found no supporting arguments, resulting in an extension score of zero. Although no explanation was discovered in this particular instance, the outcome aligns with the dataset and the predicted class. This underscores the integrity of the framework, as it avoids creating explanations in the absence of sufficient evidence. The challenge presented by the query, which divided the "opinion" of the ML algorithms, highlights its complexity and the difficulty in generating explanations for such cases. Nevertheless, there is a requirement for a broader spectrum of arguments and, conse-

quently, a wider range of possible explanations for complex queries. This can be addressed by adjusting the minimum support threshold within the Apriori algorithm. As mentioned earlier, opting for a lower minimum support value facilitates the generation of additional explanations, but this adjustment increases computational complexity and the risk of overfitting explanations to the data.

4.1 Quantitative Analysis and Limitations

This section demonstrates the selected granularity and provides a quantitative analysis of the results by calculating the fidelity metric for both datasets, each evaluated at two distinct minimum support thresholds. Fidelity measures the accuracy with which the explanation approximates the prediction of the underlying black box model. A high fidelity explanation should faithfully mimic the behavior of the black box model (Lakkaraju and Leskovec, 2019). It is quantified as the percentage of instances where both the Explainable AI Framework (XARF) and the black box model assign the same output class. The formula for calculating fidelity is as follows:

$$\text{Fidelity} = \left(\frac{\text{Number of Correct Predictions}}{\text{Total Predictions}}\right) \times 100$$

A notable feature of XARF, as delineated by its third rule, is its assurance of producing explanations that always align with the class predicted by the machine learning (ML) model. This characteristic distinguishes XARF from other explainable AI (XAI) models and impacts how fidelity is calculated, particularly in cases of missing explanations. In practical terms, the framework is designed to avoid incorrect explanations entirely, potentially achieving a 100% fidelity score across all ML models if the minimum support is optimally configured and the instance complexity is manageable. Conversely, when the minimum support threshold is set sufficiently high, XARF may fail to generate any explanations. This is particularly likely in scenarios where ML models deliver conflicting predictions, thereby complicating the instance. For our fidelity assessments, minimum support values of 0.3 and 0.4 were tested for the first dataset, and 0.19 and 0.2 for the second dataset. These values were chosen to delineate the boundary between achieving 100% fidelity and observing a decline. As observed in Table 2, a fidelity of 100% is achievable provided there are sufficient arguments to support a specific explanation. Nonetheless, the system's fidelity may diminish when the arguments are insufficient to substantiate the ML prediction. Moreover, as different ML models find different predictions, the fidelity could vary. In such cases, while XARF may provide an explanation for one predicted class, it may fail to do so for another.

Table 2: Fidelity.

	Database 1		Database 2	
	min_support = 0.3	min_support = 0.4	min_support = 0.19	min_support = 0.2
Naive Bayes	100%	100%	100%	66,67%
Decision Tree	100%	90%	100%	66,67%
Random Forest	100%	90%	100%	66,67%
Logistic Regression	100%	90%	100%	65,33%
KNN	100%	90%	100%	66,00%
SVM	100%	90%	100%	66,67%
Neural Networks	100%	90%	100%	65,33%

5 CONCLUSION

This paper introduced the Explanatory Argumentation Rule-based Framework (XARF), a novel approach in Explainable Artificial Intelligence (XAI) aimed at explaining the decision-making processes of machine learning models. Combining argumentation theory and machine learning, XARF uses a rule-based methodology to generate clear and understandable explanations for predictions made by various ML algorithms. By applying the Apriori algorithm and defining attack relations within an argumentation framework, XARF demonstrated its ability to interpret ML predictions across datasets, including the Iris dataset. XARF extracts rules from datasets using the Apriori algorithm, transforming these rules into arguments with premises and conclusions. Its innovative method for defining attack relations enhances the framework, allowing the application of standard argumentation extensions like grounded and preferred semantics. These features improve the interpretability of ML models, offering stakeholders valuable insights into AI decision-making. Empirical evaluations using a basic Boolean dataset and the Iris dataset demonstrated the adaptability of XARF to different data complexities and machine learning paradigms. XARF constructs logical arguments that align with ML predictions, even in cases where traditional explanatory models face challenges. These results highlight XARF as a valuable tool in the XAI field, addressing the critical need for transparency and accountability in AI systems. Future research could refine the rule-mining process, expand compatibility with additional ML algorithms, and explore applications in more diverse and complex datasets. A specific enhancement could involve assigning varying weights to rules based on application requirements, replacing the current uniform +1 weighting. In conclusion, XARF represents a significant advancement in XAI, offering a robust tool for uncovering the rationale behind machine learning decisions. As it evolves,

XARF holds potential for creating more transparent, trustworthy, and user-focused AI systems.

ACKNOWLEDGEMENTS

Thanks to @manna_team, the Araucária Foundation to Support Scientific and Technological Development of the State of Paraná (FA) and the National Council for Scientific and Technological Development (CNPq) - Brazil process 421548/2022-3 for the support.

REFERENCES

Achilleos, K. and Pattichis, C. (2020). Extracting explainable assessments of alzheimer's disease via machine learning on brain mri imaging data. In *Proceedings of the 2020 IEEE 20th International Conference on Bioinformatics and Bioengineering (BIBE)*, pages 1036–1041. IEEE.

Albini, E. and Tintarev, N. (2020). Deep argumentative explanations. *arXiv preprint arXiv:2012.05766*.

Amgoud, L. and Ben-Naim, J. (2018). Weighted bipolar argumentation graphs: Axioms and semantics. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI)*, pages 5194–5198.

Arrieta, A. and Herrera, F. (2020). Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information Fusion*, 58:82–115.

Cocarascu, Oana, S. and Toni, F. (2020). Data-empowered argumentation for dialectically explainable predictions. In *Proceedings of the ECAI 2020*, pages 2449–2456. IOS Press.

Dondio, P. (2021). Towards argumentative decision graphs: Learning argumentation graphs from data. In *Proceedings of the AI³@AI IA*.

Dung, P. M. (1995). On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artificial Intelligence*, 77(2):321–357.

Fan, Chao, X. and Wang, F.-Y. (2021). On interpretability of artificial neural networks: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, 32(11):4793–4813.

Gianini, G. and Portier, P.-E. (2024). *Advances in Explainable Artificial Intelligence*. Springer.

Jha, Rakesh, B. F. and Toni, F. (2020). Formal verification of debates in argumentation theory. In *Proceedings of the 35th Annual ACM Symposium on Applied Computing*, pages 940–947.

Lakkaraju, Himabindu, K. and Leskovec, J. (2019). Faithful and customizable explanations of black box models. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 131–138.

Linardatos, Pantelis, P. V. and Kotsiantis, S. (2021). Explainable ai: A review of machine learning interpretability methods. *Entropy*, 23(1):18.

Lu, Hong, e. a. (2018). Brain intelligence: Go beyond artificial intelligence. *Mobile Networks and Applications*, 23:368–375.

Potyka, N. (2021). Interpreting neural networks as quantitative argumentation frameworks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 6463–6470.

Potyka, Nils, Y. and Toni, F. (2022). Explaining random forests using bipolar argumentation and markov networks. *arXiv preprint arXiv:2211.11699*.

Samek, Wojciech, M. and Müller, K.-R. (2021). *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*. Springer Nature.

Tjoa, E. and Guan, C. (2020). A survey on explainable artificial intelligence (xai): Toward medical xai. *IEEE Transactions on Neural Networks and Learning Systems*, 32(11):4793–4813.

Van Lente, Jacob, B.-A. and Sarkadi, S. (2022). Everyday argumentative explanations for classification. *Argumentation & Machine Learning*, 3208:14–26.

Vassiliades, Alexander, B. and Patkos, T. (2021). Argumentation and explainable artificial intelligence: A survey. *The Knowledge Engineering Review*, 36:e5.