INTERNSHIP REPORT

M2 MIAGE INNOVATIVE INFORMATION SYSTEMS

Université Toulouse 1 Capitole (UT1 Capitole)

Presented and supported on *Date of defense (05/09/2024)* **by :**

Usma AKTAR

Abstract Theory of Explainability of Black-box Models

JURY

M. DAVID SIMONCINI	UT1 Capitole	Academic Supervisor
M. UMBERTO GRANDI	UT1 Capitole	Examiner
M. MARTIN COOPER	IRIT	Laboratory Supervisor

Research Team and Laboratory :

ADRIA, Institut de Recherche en Informatique de Toulouse

Supervisor(s) :

Prof. Martin COOPER and *Dr. Leila AMGOUD*

Acknowledgments

There are many individuals whose inspiration and contributions have been absolutely invaluable. Their tireless efforts, empathy, and unwavering support are the reasons I stand where I am today.

It is with immense pleasure that I express my deepest gratitude to my internship supervisors, Prof. Martin COOPER and Dr. Leila AMGOUD. From the very beginning until the end of my internship, you both supported me in every possible way. I truly appreciate the effort you put into providing me with a comfortable office space and all the necessary IT and technical resources. I am proud and deeply thankful for your steadfast support, without which I could not have successfully completed this important chapter of my life. Your guidance consistently steered me in the right direction whenever I veered off course.

I would also like to extend my heartfelt thanks to Prof. Chihab HANACHI, my academic supervisor, Prof. David SIMONCINI, and my examiner, Prof. Umberto GRANDI. Your positive demeanor and willingness to assist have been crucial in helping me achieve success in my master's program. I am truly fortunate to have had the opportunity to work with such kind and supportive individuals.

From the bottom of my heart, I want to express my deepest feelings of love, gratitude, and admiration for my father, who has been the greatest inspiration in my life. His wisdom and foresight have been the driving forces behind this beautiful day. I also extend my profound thanks to my mother, who has always been a source of emotional support throughout my life; my love for her is boundless and everlasting. Additionally, I am grateful to my sister and brother, whose pride in me and generous love have always given me the strength to persevere.

It is a genuine pleasure to express my deep gratitude to all of you.

Usma AKTAR

Contents

Table of abbreviations and acronyms

AI	***A**rtificial **I**ntelligence*
AXp	***A**bductive e**Xp**lanation*
Cv-dAXp	***C**ross **V**alidated **D**ata **A**bductive e**Xp**lanation*
dAXp	***D**ata **A**bductive e**Xp**lanation*
dwAXp	***D**ata **W**eak **A**bductive e**Xp**lanation*
XAI	*e**X**plainable **A**rtificial **I**ntelligence*

List of Figures

List of Tables

[1] ***C**ross **V**alidated **D**ata **A**bductive e**Xp**lanation*

Introduction

Contents

1.1 Context

The application of artificial intelligence (AI[1]) is advancing rapidly across all domains. However, the critical aspect is not merely the creation of a model, but its acceptance and adoption. This acceptance does not depend only on the model's performance accuracy. While many models can provide highly accurate answers, these answers often lack contextual relevance to the inputs and outputs. Therefore, a model cannot be considered sustainable for future use solely based on its ability to produce correct answers.

Understanding the inner workings of a model, particularly how it makes decisions, is paramount, especially when handling complex calculations. Deep neural networks are a prime example of such models, often referred to as "black boxes" due to their opaque decision-making processes. As AI technology evolves, researchers are increasingly focused on demystifying these black boxes, striving to understand the internal mechanisms that drive their outputs.

Our research is primarily dedicated to define how a model can be employed to meaningfully interpret the complex phenomena within it. This endeavor aims to bridge the gap between accurate performance and contextual understanding, ensuring that AI models are not only correct but also comprehensible and sustainable for future applications.

1.2 Mission

The mission of this M2 internship is to advance the field of Explainable Artificial Intelligence (XAI[2]) by developing an abstract theoretical framework for explaining the behavior of black-box AI models. This research aims to bridge the gap between diverse AI models and their explainers by defining

[1] ***A**rtificial **I**ntelligence*
[2] *e**X**plainable **A**rtificial **I**ntelligence*

a universal formal setting where AI models are treated as abstract entities operating as black-box functions. Therefore key objectives include:

(a) **Defining a Formal Setting for Abstract Models:** Establishing a formal context where AI models are treated as abstract entities. These entities are based on black-box functions, meaning they can be queried to produce outputs from given inputs without revealing their internal workings.

(b) **Defining the Notion of an Explainer:** Developing a formal definition for explainer's methods that provide explanations for the behavior of these black-box models. This includes identifying and specifying the properties that an effective explainer should possess, drawing from existing literature on explainability in AI.

This internship is part of a broader project aimed at creating an abstract theory of explainability that can be applied across various types of AI models, which may manipulate different objects and perform different tasks. The ultimate goal is to enhance trust in automated decision-making systems by providing understandable and transparent explanations for their decisions, regardless of the specific nature of the models being used.

1.3 Achievements

In this internship, we successfully developed a robust theoretical framework for Explainable Artificial Intelligence (XAI), focusing on the explainability of black-box AI models. This included study on an *explainer*, the Cross Validated Data Abductive Explanation (Cv-dAXp), aimed at improving the transparency and interpretability of machine learning models. My work played a key role in narrowing the gap between accurate performance and contextual understanding in AI systems, contributing to greater trust and ethical implementation across various real-world applications.

1.4 Organization of the report

The remainder of this report is organized as follows: Chapter 2 provides an overview of the laboratory, the Institut de Recherche en Informatique de Toulouse (IRIT); Chapter 3 outlines the goals of the internship; Chapter 4 discusses various solution approaches; Chapter 5 presents and evaluates the proposed solution; Chapter 6 focuses on the innovative aspects; Chapter 7 details the outcomes; and Chapter 8 concludes the report, along with suggestions for future work.

Lab Overview: Institut de Recherche en Informatique de Toulouse (IRIT)

Contents

2.1 Introduction

The Institut de Recherche en Informatique de Toulouse (IRIT) is one of the largest computer science research laboratories in France, known for its excellence in both theoretical and applied research across various domains of computer science. According to [Ins24], IRIT focuses on a wide array of disciplines, ranging from artificial intelligence and data science to formal methods and human-computer interaction. Established in 1990, IRIT is a joint research unit (UMR 5505) affiliated with the Centre National de la Recherche Scientifique (CNRS), Université Toulouse III - Paul Sabatier, Université Toulouse I - Capitole, and Institut National Polytechnique de Toulouse. The lab's mission is to push the boundaries of knowledge in computer science and to apply these advancements to solve real-world problems. IRIT's vision is to serve as a bridge between theoretical research and practical applications, ensuring that the insights gained from fundamental studies translate into technologies that have real-world impact. The laboratory's commitment to fostering interdisciplinary collaboration and addressing societal challenges through scientific excellence is reflected in its diverse research teams and projects.

2.2 Research Areas

IRIT's research is organized into several interdisciplinary teams, each focusing on specific domains of computer science. The lab's primary research areas include the design and construction of systems, digital modeling of the real world, concepts for cognition and interaction, autonomous systems

adaptive to their environment, and moving from raw data to intelligible information. The lab fosters a collaborative environment where theoretical advancements are often coupled with practical applications, making it a hub for innovation in computer science.

2.3 ADRIA team

The ADRIA (Argumentation, Décision, Raisonnement, Incertitude et Apprentissage) team, under which my M2 internship was conducted, one of IRIT's core research groups. ADRIA plays a pivotal role in IRIT's overarching mission by focusing on the development of advanced methods for reasoning, decision-making, and learning, particularly in contexts where information is incomplete, uncertain, or contradictory. The team's research is characterized by a deep integration of formal theoretical approaches with practical computational models, striving to overcome the limitations of classical logic in handling complex real-world problems.

ADRIA's key activities include pioneering research in non-monotonic logics, imprecise probabilities, and possibility theory, all aimed at enhancing the capacity of artificial intelligence systems to make informed decisions under uncertainty. The team is also dedicated to exploring novel methods for information revision, fusion of conflicting data from multiple sources, and abductive reasoning, which is critical for diagnosing plausible causes of observed phenomena. Additionally, ADRIA is actively involved in the formalization of argumentation, a crucial aspect of AI that supports explainability, negotiation dialogues, and the management of contradictions.

2.4 My Internship Context

My internship is focused on the "Abstract Theory of Explainability of Black-box Models," an area of growing importance in the AI research community. This internship is a part of ADRIA's broader research agenda on explainable AI (XAI), which seeks to create methods for making complex AI models more understandable and transparent to users. The work involves both theoretical foundations and practical approaches to explaining black-box models, contributing to the team's goal of improving the interpretability and trustworthiness of AI systems. So, my research aimed to develop a formal framework for improving the interpretability of complex, data-driven models, addressing the growing demand for transparency and trust in AI systems.

Being part of the ADRIA team not only allowed me to contribute to cutting-edge research but also enabled me to gain valuable insights into the broader challenges and methodologies that define the field of AI explainability. The collaborative and intellectually stimulating atmosphere of IRIT, coupled with ADRIA's focus on high-impact research, made this internship a pivotal experience in my academic and professional development.

2.5 Facilities and Collaborations

IRIT is equipped with state-of-the-art research facilities and computational resources, enabling cutting-edge research in various domains. The lab maintains strong collaborations with academic institutions, industrial partners, and research organizations at both national and international levels. These collaborations ensure that IRIT's research is both impactful and relevant, addressing current and future challenges in computer science.

2.6 Role within the Organization

During my internship at the IRIT lab in Toulouse, I had the opportunity to work on a significant project focused on developing an abstract theory of explainability for black-box models. Here are the key aspects of my contributions and experiences:

(a) **Research and Literature Review:**

- Conducted extensive research to define formal settings for explainability.
- Reviewed current literature to identify gaps in existing theories.

(b) **Conceptualization of Explainer:**

- Developed a precise, formal understanding of what constitutes an explainer in the context of black-box models.
- Exchanged with my supervisor and a senior researcher to draft and refine these definitions.

(c) **Collaboration and Teamwork:**

- Worked alongside the experienced researcher of the ADRIA team, participating in regular meetings and presenting my work to the team to discuss work progress and resolve any issues.
- Gained insights into the importance of interdisciplinary approaches in research.

(d) **Dynamic Research Environment:**

- Gained valuable experience working in a dynamic research environment, which enhanced my practical research skills.
- Contributed to cutting-edge theoretical developments in artificial intelligence.

Overall, my internship at the IRIT lab not only deepened my knowledge of artificial intelligence and explainability but also provided me with essential experience in theoretical and applied research within this field.

2.7 Relevance to Field

IRIT plays an important role in the academic field of Computer Science and Innovative Information Systems. The lab's work covers many areas of computer science, like artificial intelligence, data science, formal methods, and human-computer interaction. By developing new algorithms, methods, and models, IRIT helps push forward academic research and expands our understanding of complex problems in these fields.

In the area of Innovative Information Systems, IRIT's research provides the foundation for creating systems that can handle and analyze large amounts of data in challenging environments. The lab's work on topics like reasoning with uncertain information and combining data from different sources helps build more intelligent and reliable information systems. These advancements are especially valuable to academics who are focused on making information systems more effective and dependable.

IRIT also takes an interdisciplinary approach to research, which increases its impact on the academic field. By combining knowledge from areas like cognitive psychology and risk analysis, IRIT ensures that its work is not only technically strong but also considers the needs of people who use these systems. This approach helps bridge the gap between theory and real-world applications, making IRIT's research highly relevant to the study of Computer Science and Innovative Information Systems.

2.8 Conclusion

In conclusion, the Institut de Recherche en Informatique de Toulouse (IRIT) stands as a cornerstone of computer science research in France, exemplifying excellence through its extensive interdisciplinary collaborations and cutting-edge innovations. The laboratory's diverse research areas, from artificial intelligence to human-computer interaction, reflect its commitment to advancing both theoretical knowledge and practical applications. My internship within the ADRIA team at IRIT provided invaluable experience, allowing me to contribute to the emerging field of explainable AI (XAI). Through its rigorous research, IRIT continues to play a pivotal role in shaping the future of computer science, fostering advancements that bridge the gap between complex theories and their real-world applications.

Internship Goals

Contents

3.1 Goal

The primary goal of this M2 internship is to establish a comprehensive theoretical foundation for the explainability of black-box AI models. This involves advancing the transparency and interpretability of machine learning models by studying a particular explainer proposed by my supervisors. This explainer offers meaningful insights into complex phenomena, helping to bridge the gap between accurate performance and contextual understanding. The key goals are to:

(a) **Understanding how a model makes decisions through black-box functions:** Investigate the internal workings of AI models, often referred to as "black-box" functions, to decipher how they arrive at specific decisions. This involves analyzing the model's processes and the factors influencing its outputs.

(b) **Defining meaningful interpretation of complex phenomena within models:** Develop methods to translate the intricate behaviors and interactions within models into comprehensible explanations. This aims to make the decision-making process of models more accessible and understandable to users, stakeholders, and domain experts.

(c) **Bridge the gap between accurate performance and contextual understanding by analyzing a novel explainer**: Studied on an explainer that not only maintains the model's high performance but also provides contextual insights into its operations. This explainer will serve as a tool to elucidate how models achieve their results, ensuring that users can trust and effectively utilize the model's predictions.

(d) **Studying the concept of an explainer and the formal properties it would satisfy or not:** Establish a clear explanations of what constitutes an explainer, outlining the formal properties it should satisfy. This includes criteria for validity, reliability, comprehensibility, and applicability, as well as identifying any limitations or constraints the explainer may have.

This goal statement outlines a comprehensive approach to improving the interpretability of machine learning models, balancing the need for both accuracy and understanding.

3.2 Problem Significance

The problems addressed by this project are crucial both in practical and theoretical contexts, given the increasing reliance on automated models in various domains, such as finance, healthcare, and autonomous systems. These models, often treated as "black boxes," produce decisions and predictions without providing transparent reasoning. This lack of transparency raises significant concerns about trust, accountability, and ethical implications, making it imperative to develop methodologies for explaining their behavior.

3.2.1 Practical Importance

(a) **Enhancing Trust and Adoption:** As AI models are deployed in critical decision-making processes, it is vital that end-users, including domain experts and non-experts, trust the decisions made by these models. The absence of clear, understandable explanations can lead to skepticism and resistance to adopting AI systems. By addressing the explainability of black-box models, this project directly contributes to enhancing user trust and confidence in AI, facilitating broader acceptance and integration of these technologies into real-world applications.

(b) **Compliance with Regulations:** In many industries, regulations and ethical guidelines require that automated decisions be explainable. For instance, the General Data Protection Regulation (GDPR) in Europe mandates that individuals have the right to obtain an explanation for decisions made by automated systems [Kam21]. By providing a formal framework for explainability, this project helps ensure that AI models comply with such legal and ethical standards, thereby avoiding potential legal risks and penalties.

(c) **Improving Decision-Making:** In practical settings, stakeholders often need to understand the reasoning behind AI-driven decisions to make informed choices. For example, in healthcare, a clinician may need to know why an AI model predicts a certain diagnosis or treatment. Explainability allows for the validation and verification of model outputs, reducing the likelihood of errors and improving overall decision quality.

3.2.2 Theoretical Importance

(a) **Abstract Framework for Explainability:** The internship aims to study on a general theory of explainability that can be applied across various AI models, regardless of their specific nature or the objects they manipulate. This theoretical framework addresses the fragmentation in current XAI approaches by providing a unified perspective on what it means to explain a model. Such a framework is critical for advancing the field of AI by offering a common ground for comparing and evaluating different explainability methods.

(b) **Foundational Understanding of Black-Box Models:** The project also seeks to formalize the notion of black-box models as abstract entities defined by functions with unknown internal workings. This theoretical exploration is significant because it deepens our understanding of the fundamental properties that any explainer must satisfy to be considered valid. By rigorously exploring these properties, the project contributes to the foundational knowledge necessary for building robust and reliable explainers.

(c) **Advancement of Explainability Research:** The project's focus on abstracting the concept of explainability across diverse AI models supports the advancement of research in XAI. By offering a formal setting that can be adapted to different model types, it opens up new avenues for developing and testing explainability techniques. This theoretical contribution is essential for pushing the boundaries of what can be explained and how, ultimately leading to more sophisticated and adaptable AI systems.

3.3 Conclusion

The project addresses critical issues in both practical and theoretical contexts by aiming to enhance the trustworthiness, compliance, and effectiveness of AI systems through explainability. Its contributions to the abstract theory of explainability will serve as a cornerstone for future research and development in this rapidly evolving field.

Solutions Exploration

Contents

4.1 Introduction

In this study, we look closely at how a learning model works, employing various techniques to explain it and unravel its underlying mechanisms. After training the model and using it to make predictions on test data, it is important to evaluate how accurate these predictions are. This evaluation is not just about performance metrics; it is about assessing whether the predictions are based on real, meaningful patterns within the data, rather than being the result of random chance.

Amgoud *et al.* [ACD24] explore a range of *explainer* methods that offer feature-based explanations derived from individual samples or entire datasets. An *explainer* is a *function* that helps to understand and interpret the *decisions* made by a machine learning *classifier*. These *explainers* help to clarify which features or attributes significantly influence the model's predictions. By analyzing the significance and impact of each feature, these methods allow users to understand the underlying reasons behind the model's decisions. This understanding is crucial for validating the model's behavior, ensuring transparency, and enhancing trust in the model's outcomes. Additionally, feature-based explanations can help in identifying potential biases or errors in the model, facilitating improvements and ensuring more reliable predictions.

The authors in [ACD24] suggested various *explainers* to meet the specific properties. In this research, we aim to propose an additional *explainer* for the same properties, enhancing the understanding and clarity of how these features influence the model's predictions. This new *explainer* will complement and extend the current methods.

4.2 Understanding an Explainer

According to [ACD24], an *explainer* is a function that, given a question as input, generates a set of explanations as output. It helps to interpret and comprehend the decisions made by a machine learning classifier. When a *classifier* processes inputs to make *predictions* or *decisions*, the *explainer* takes these same inputs (a question) and provides detailed *explanations* about how and why the *classifier* reached its conclusions. This is particularly useful in complex models like deep neural networks, where the decision-making process is not immediately transparent. For instance, consider a *classifier* trained to identify whether an email is spam or not. The inputs might be features like the presence of certain keywords, the email's sender, or the structure of the email. If the *classifier* decides an email is spam, the *explainer* function can analyze these features and indicate which ones were most influential in making that decision.

The purpose of an explainer is to make machine learning models more interpretable and transparent, thereby increasing trust and accountability. This is essential in applications where understanding the rationale behind a decision is crucial, such as in medical diagnoses, financial decisions, or legal judgments. By providing insights into the *classifier*'s decision-making process, explainers help ensure that the decisions are fair, unbiased, and justifiable, ultimately enhancing the usability and reliability of machine learning systems.

To describe how an *explainer* works, we need to introduce several key concepts, such as *classification theory*, *literal*, *question*, *feature space*, and *explainer* function. The followings are the mathematical representation of these concepts.

Classification Theory

A *classification theory* is represented as a tuple $\mathbb{T} = \langle \mathtt{F}, \mathtt{dom}, \mathtt{C} \rangle$, where:

- $\mathtt{F}$ is a finite set of *features*.
- $\mathtt{dom}$ is a function that returnes the *domain* of each *feature*, with $|\mathtt{dom}(.)| > 1$ and $\mathtt{dom}(.)$ is finite.
- $\mathtt{C}$ is a finite set of *classes* with $|\mathtt{C}| \geq 2$.

Literal

A *literal* in $\mathbb{T}$ is a pair (f, v), where:

- f is a *feature* in $\mathtt{F}$
- v is a value from the domain of feature f in $\mathtt{dom}(f)$

Question

A *question* is defined as a tuple $\mathbf{Q} = \langle \mathbb{T}, \kappa, \mathcal{D}, x \rangle$, where:

- $\mathbb{T}$ is a *classification theory*.
- κ is a classification on $\mathbb{T}$.
- $\mathcal{D}$ is the dataset used for our explanation and $\mathcal{D} \subseteq \mathbb{F}(\mathbb{T})$.
- x is an instance from $\mathcal{D}$.

Feature Space

The *feature space*, denoted by $\mathbb{F}(\mathbb{T})$, is the set of all instances in theory $\mathbb{T}$.

Classifier

A *classifier* on a theory $\mathbb{T}$ is a function $\kappa : \mathbb{F}(\mathbb{T}) \to \mathbf{Q}$, which maps every instance in $\mathbb{F}(\mathbb{T})$ to a *classs* in the set C.

Explainer

An *explainer* is a *function* $\mathbf{L}$ that takes every *question* $\mathbf{Q}$ into a subset of $\mathbb{E}(\mathbb{T})$ (the set of all possible partial assignments) and provides explanation E for the *classifier*'s predictions on specific sample of instances. The explanation E is a subset of the features and their values that are most relevant to the prediction under the dataset $\mathcal{D}$.

$\mathcal{D}_1$	age	$income$	$student$	$credit\_rating$	$\kappa(x_i)$
x_1	youth	high	yes	fair	deny
x_2	middle-aged	medium	no	excellent	approve
x_3	senior	low	yes	fair	deny
x_4	youth	low	no	excellent	approve

Table 4.1: Credit application example

The dataset $\mathcal{D}_1$ shown in Table 4.1 can be utilized to train a model for predicting the outcomes of credit applications. Consider a *classifier* κ_1 that predicts the outcome C = {approve, deny} based on each *instance* x in the *dataset*, where each *instance* represents a credit application with *features* $\mathbb{F} = \{age, income, student, credit\_rating\}$. The *features* have different domains; for instance, the age feature has three categories: youth, middle-aged, and senior.

Example 1. *Suppose that* $\mathbf{Q} = \langle \mathbb{T}, \kappa_1, \mathcal{D}_1, x_1 \rangle$ *and an explainer* $\mathbf{L}$ *such that the* function $\mathbf{L}(\mathbf{Q})$ *returns the explanations* $E = \{(age = youth), (student = yes)\}$ *and* $E' = \{(credit\_rating = fair)\}$ *of* $\kappa_1(x_1)$ *for the dataset* $\mathcal{D}_1$*. In this scenario, the explainer provided two explanations. Explanation* E *indicates that the* classifier*'s decision to approve or deny the credit application of sample* x_1 *is primarily based on the applicant's* age *and* $student$ *status. Another explanation,* E'*, suggests that the* classifier*'s decision for sample* x_1 *is based on the applicant's* $credit\_rating$*.*

4.3 Challenges of existing explainers

Among the most popular explanations, abductive explanation (AXp[1]) is discussed most extensively in the literature. According to [Mil19], AXp is described as the process of reasoning to the best explanation. Specifically, in the context of AI, it involves selecting the most plausible hypothesis or explanation for a given set of observations or outputs generated by an AI model. Also, AXp refers to a type of explanation that highlights feature-values sufficient to account for a given classification decision, as described in [Amg23]. This approach focuses on identifying the most plausible causes or factors that could explain why a specific outcome was reached by the classifier, even if they are not definitively provable, thereby making the AI's decisions more understandable and interpretable to users.

The existing literature has shown that the sample-based explanation approach is highly challenging due to the various issues it encounters. According to [CA23]; [Amg23], dataset-based abductive explanation (dAXp[2]) can lead to global inconsistency in the following technical sense: two explanations for instances belonging to different classes might be compatible, implying that a third instance could satisfy both sufficient reasons for conflicting classes. As stated in [AMT23], it is also impossible to define an explanation function that produces subset-minimal sufficient reasons while simultaneously guaranteeing the existence of explanations (success) and maintaining their global consistency. The authors also introduced functions that prioritize consistency over success, while those in [CA23] ensure success but compromise consistency. However, no existing explanation function in the literature guarantees both properties for dAXp.

In the recent study, Amgoud *et al.* [ACD24] proposed some novel explanation functions that produce feature-based explanations from samples or datasets while achieving both of the above properties. These explanations are considered adequate for justifying the predicted classifications of instances within a *classification theory*.

Here we explain two explainers from [ACD24]: *Irrefutable Explainer* and *Surrogate Explainer*.

4.3.1 Irrefutable Explainer

The *irrefutable explainer* relies on an *irrefutable envelope* that encompasses all weak abductive explanations (dwAXp[3]), ensuring there are no conflicts among them. This *envelope* includes the complete dataset for which it is designed and represents the intersection of all the largest possible *coherent* sub-envelopes. In essence, it guarantees that every explanation that does not conflict with others is considered within the full context of the data.

Suppose, an *irrefutable explainer* is a *function* $\mathbf{L}_{irr}$ such that for any *question* $\mathbf{Q} = \langle \mathbb{T}, \kappa, \mathcal{D}, x \rangle$,

$$\mathbf{L}_{irr}(\mathbf{Q}) = \{E \in \texttt{Irr}(\mathcal{D}, \kappa) \mid E \subseteq x_i\}. \tag{4.1}$$

Example 2. *Consider the theory* $\mathbb{T}_1$ *consisting of two binary features and a binary classifier* κ_2 *which*

[1] ***A**bductive e**Xp**lanation*
[2] ***D**ata **A**bductive e**Xp**lanation*
[3] ***D**ata **W**eak **A**bductive e**Xp**lanation*

provides the predictions below for three instances in $\mathcal{D}_2$. Let $\mathbf{Q}_i = \langle \mathbb{T}_1, \kappa_2, \mathcal{D}_2, x_i \rangle$, with $i \in \{1, 2, 3\}$.

$\mathcal{D}_2$	f_1	f_2	$\kappa_2(x_i)$
x_1	0	1	1
x_2	1	0	0
x_3	1	1	1

Table 4.2: Predictions of *classifier* κ_2 for the instances in dataset $\mathcal{D}_2$.

$$E_1 = \{(f_1, 0)\}$$
$$E_2 = \{(f_2, 1)\}$$
$$E_3 = \{(f_2, 0)\}$$
$$E_4 = x_1 \;\; E_5 = x_2 \; E_6 = x_3$$

Note that, the function $\mathbf{L}_{dw}$ generate dwAXp's for $\bigcup_{i=1}^{3} \mathbf{L}_{dw}(\mathbf{Q}_i) = \{E_1, \ldots, E_6\}$. It is also important to note that the set $\{E_1, E_3\}$ is incoherent, and thus, the irreducible explanations for $\texttt{Irr}(\mathcal{D}_2, \kappa_2)$ are $\{E_2, E_4, E_5, E_6\}$. Consequently, $\mathbf{L}_{irr}(\mathbf{Q}_1) = \{E_2, E_4\}$, $\mathbf{L}_{irr}(\mathbf{Q}_2) = \{E_5\}$, $\mathbf{L}_{irr}(\mathbf{Q}_3) = \{E_2, E_6\}$.

Output Analysis

In the above context of the example 2 provided the weak abductive explanations $E_1, E_2, E_3, E_4, E_5, E_6$ for the dataset $\mathcal{D}_2$.

The incoherent set is $\{E_1, E_3\}$ because they cannot coexist under the same *classifier* prediction $\kappa_2(x_i)$, so the irreducible explanations are:

$$\texttt{Irr}(\mathcal{D}_2, \kappa_2) = \{E_2, E_4, E_5, E_6\}$$

Query for the question $\mathbf{Q}_1 = \langle \mathbb{T}_1, \kappa_2, \mathcal{D}_2, x_1 \rangle$

- **Instance** x_1 has $\kappa_2(x_1) = 1$.
- **Relevant explanations for** x_1: $\{E_1, E_2, E_4\}$
- **Irreducible explanations for** x_1 (excluding incoherent E_1):

$$\mathbf{L}_{irr}(\mathbf{Q}_1) = \{E_2, E_4\}$$

Query for question $\mathbf{Q}_2 = \langle \mathbb{T}_1, \kappa_2, \mathcal{D}_2, x_2 \rangle$

- **Instance** x_2 has $\kappa_2(x_2) = 0$.
- **Relevant explanations for** x_2: $\{E_3, E_5\}$
- **Irreducible explanations for** x_2 (excluding incoherent E_3):

$$\mathbf{L}_{irr}(\mathbf{Q}_2) = \{E_5\}$$

Query for question $\mathbf{Q}_3 = \langle \mathbb{T}_1, \kappa_2, \mathcal{D}_2, x_3 \rangle$

- **Instance** x_3 has $\kappa_2(x_3) = 1$.
- **Relevant explanations for** x_3: $\{E_1, E_2, E_6\}$
- **Irreducible explanations** for x_3 (excluding incoherent E_1):

$$\mathbf{L}_{irr}(\mathbf{Q}_3) = \{E_2, E_6\}$$

In the analysis of *irrefutable explanations* for dataset $\mathcal{D}_2$ under *classifier* κ_2, we identified the incoherent set $\{E_1, E_3\}$ and the *irreducible explanations* $\{E_2, E_4, E_5, E_6\}$. For each query, we determined the relevant and *irreducible explanations* for the given instances. This approach ensures that only *coherent* and essential *explanations* are considered for each instance, enhancing the interpretability of the *classifier*'s predictions.

4.3.2 Surrogate Explainer

A *surrogate explainer* is a method for providing *coherent* explanations for a *classifier*'s decisions by using a *surrogate classifier* that mimics the original *classifier* on a given dataset. Specifically, the *surrogate classifier*, often a *decision tree*, is chosen because it allows for tractable and understandable dwAXp's over the entire *feature space*. The *surrogate classifier*, denoted by σ, finds a *decision tree* for any model κ with 100% accuracy on any dataset $\mathcal{D}$. The final explanations produced by the *surrogate explainer*, denoted by $\mathbf{L}_{su}$, will be derived from the dwAXp's values of $\mathbf{L}_{dw}$ obtained from the *surrogate classifier* σ. Even though the *surrogate* may not perfectly match the original *classifier* outside the dataset, it ensures *coherence* and provides explanations that are simpler and computationally efficient.

Example 3. *Let $\mathbb{T}_2$ be a theory, $\mathcal{D}_3 \subseteq \mathbb{T}_2$, and we have an original κ_3 and a* surrogate σ classifiers *on $\mathbb{T}_2$ such that $\forall y \in \mathcal{D}_3$, $\sigma(y) = \kappa(y)$. A surrogate explainer is a function $\mathbf{L}_{su}$ such that*

$$\forall x \in \mathcal{D}_3, \quad \mathbf{L}_{su}(\langle \mathbb{T}_2, \kappa_3, \mathcal{D}_3, x \rangle) = \mathbf{L}_{dw}(\langle \mathbb{T}_2, \sigma, \mathbb{F}(\mathbb{T}_2), x \rangle) \tag{4.2}$$

Suppose, the original classifier κ_3 *takes two binary input features f_1 and f_2 from the dataset $\mathcal{D}_3$ and outputs a binary class label* $\mathtt{C}$ *(0 or 1). The* classifier κ_3 *works as follows:*

- *If $f_1 = 1$ and $f_2 = 0$, then the output is 1.*
- *For all other input combinations, the output is 0.*

We can summarize this behavior in Table 4.3.

$\mathcal{D}_3$	f_1	f_2	$\kappa_3(x_i)$
x_1	0	1	0
x_2	1	0	1
x_3	1	1	0

Table 4.3: Predictions of original *classifier* κ_3 for the instances in dataset $\mathcal{D}_3$.

4.3.2.1 Surrogate Classifier (σ)

Now, we will create a *surrogate classifier* σ using a *decision tree* that mimics the behavior of the original *classifier* κ_3 on the given dataset $\mathcal{D}_3$. The *decision tree* will provide a *coherent* and understandable explanation for the *classifier's decisions*.

The *decision tree* for σ is constructed as follows:

Root Node: Checks if $f_1 = 1$.

- If **No** (i.e., $f_1 = 0$), the output is 0.
- If **Yes** (i.e., $f_1 = 1$), it checks if $f_2 = 0$.

Second Level: Checks if $f_2 = 0$ (only if $f_1 = 1$).

- If **Yes** (i.e., $f_2 = 0$), the output is 1.
- If **No** (i.e., $f_2 = 1$), the output is 0.

This *decision tree* accurately represents the *classifier* κ_3's logic for the given *dataset* $\mathcal{D}_3$ and it can be represented visually in Figure 4.1.

The corresponding *decision tree* in textual form:

```
if f1 == 1:
    if f2 == 0:
        return 1
    else:
        return 0
else:
    return 0
```

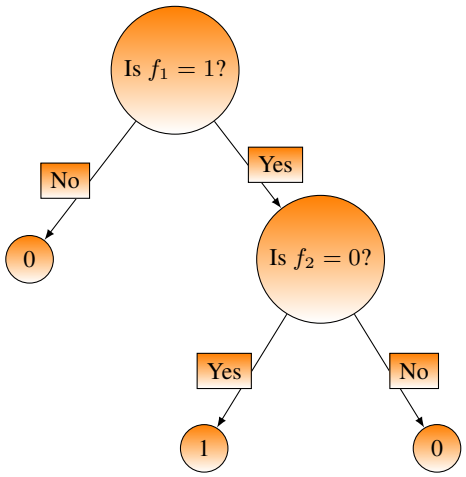

Figure 4.1: *Decision tree* generated by the *Surrogate Classifier* σ

From the above descriptions, we know that *surrogate explainer function* $\mathbf{L}_{su}$ generate *explanation's* from the *surrogate classifier's decision tree*. So, our $\mathbf{L}_{su}$ *function* gives us this *explanation*

$$\mathbf{L}_{su}(\langle \mathbb{T}_2, \kappa_3, \mathcal{D}_3, x_i \rangle) = \{E_1, E_2, E_3\}$$

Explanations* for Each *instance

- For $x_1 = (0, 1)$:
$$\mathbf{L}_{su}(\langle \mathbb{T}_2, \kappa_3, \mathcal{D}_3, (0, 1) \rangle) = E_1 = \{(f_1, 0)\}$$
- For $x_2 = (1, 0)$:
$$\mathbf{L}_{su}(\langle \mathbb{T}_2, \kappa_3, \mathcal{D}_3, (1, 0) \rangle) = E_2 = \{x_2\}$$
- For $x_3 = (1, 1)$:
$$\mathbf{L}_{su}(\langle \mathbb{T}_2, \kappa_3, \mathcal{D}_3, (1, 1) \rangle) = E_3 = \{(f_2, 1)\}$$

Minimal Explanation Set for $\mathbf{L}_{su}$

The minimal explanation set that covers all the decision rules used by the *surrogate classifier* σ is:

- If $f_1 = 1$ and $f_2 = 0$, then predict 1.
  - This rule covers the instance $x_2 = (1, 0)$
- Otherwise, predict 0.
  - This rule covers the instances $x_1 = (0, 1)$ and $x_3 = (1, 1)$

In literal form, the minimal explanation set $\mathbb{E}$ for the whole dataset is:

- $(f_1, 1)$ and $(f_2, 0)$: This ensures that when both conditions are met, the prediction is 1.
- $(f_2, 1)$: This ensures that when $f_2 = 1$, the prediction is 0 (since it contradicts the condition for predicting 1).

This set of rules is derived from the decision tree and represents the complete logic of the *surrogate classifier* σ for the dataset $\mathcal{D}_3$.

In this example, we demonstrated how a *surrogate explainer* can be used to provide *coherent* explanations for a *classifier*'s decisions using a *decision tree*. The *surrogate classifier* σ accurately mimics the original *classifier* κ_3 on the dataset $\mathcal{D}_3$, ensuring understandable and tractable explanations over the feature space.

4.4 Conclusion

In this chapter, we discussed the mechanisms of *explainers* used to interpret machine learning models, focusing on their ability to provide transparent and meaningful explanations for model predictions. We highlighted the importance of evaluating not just the performance metrics of a model but also whether its predictions are grounded in substantial patterns rather than random chance. We explored two *explainers* named *irrefutable explainer* and *surrogate explainer*, discussing their effectiveness in addressing issues such as global consistency and coherence of explanations. The *irrefutable explainer* was shown to produce consistent explanations by filtering out incoherent ones, while the *surrogate explainer* provided understandable and coherent explanations by mimicking the original *classifier's* behavior with simpler models like *decision trees*. This comparative analysis underscores the critical need for effective and reliable explanation methods to enhance the interpretability of machine learning models, ensuring that their predictions are not only accurate but also justifiable and comprehensible.

Proposed Solutions

Contents

5.1 Proposed solutions

In this internship work, I contributed to study, analyze, and evaluate the method proposed by my supervisor. My work involved adapting the methodology to the context, analyzing the results, and identifying potential improvements. So, we studied a novel explainer that generate a new explanation method called the *Cross Validated Data Abductive Explanation* (Cv-dAXp). This method provides explanations for a classifier's output by ensuring consistency through cross validation against features from other instances within the dataset.

We assume that the cross validated data abductive explainer is denoted by L_{cv}, which produces a collection of Cv-dAXp's. Each explanation in this set is composed of a specific subset of *literals*. The cross-validated data abductive explainer L_{cv} is defined as a function that, given any *question* $\mathbf{Q} = \langle \mathbb{T}, \kappa, \mathcal{D}, v \rangle$, generates a set of *explanations* for the classifier $\kappa(v)$.

Let $X \subseteq v$ and X is a Cv-dAXp of v if the following condition holds:

$$\forall u \in \mathcal{D}, \quad \kappa(v[\tilde{X}] \cup u[F \setminus \tilde{X}]) = \kappa(v). \tag{5.1}$$

Here, we denote the set of features as $\tilde{X}$, which are present in the set of literals X. To verify whether X qualifies as a Cv-dAXp for v, it must satisfy a following condition. For every instance u in the dataset $\mathcal{D}$, the result of applying the *classifier* κ to the combined set of $v[\tilde{X}]$ and $u[F \setminus \tilde{X}]$ should be the same as the result of applying κ to v alone. The condition ensures that the classifier κ returns the same value as v on any feature-vector which is identical to v on the features of X and the same as any other instance u in the dataset $\mathcal{D}$ on the other features.

Fundamentally, X is considered cross validated because it consistently captures the core charac-

teristics of v's performance across various combinations of data instances and subsets of features. This process highlights the stability and resilience of X in terms of feature selection. By reliably reflecting v's behavior under different data scenarios, X is validated as a significant and representative subset of features for the specific instance v within the dataset $\mathcal{D}$. The explanations concerning Cv-dAXp are further illustrated in Example 4.

Example 4. *Consider the theory $\mathbb{T}_3$ consisting of two binary features and a binary classifier $\kappa_4(f_1, f_2) = f_1 \vee (\neg f_2)$ which provides the predictions in Table 5.1 for three instances in $\mathcal{D}_4$. Let we focus on the first* instance x_1 *and want to find out the Cv-dAXp's of* question $\mathbf{Q}_1 = \langle \mathbb{T}_3, \kappa_4, \mathcal{D}_4, x_1 \rangle$.

$\mathcal{D}_4$	f_1	f_2	$\kappa_4(x_i)$
x_1	0	1	0
x_2	1	0	1
x_3	1	1	1

Table 5.1: Predictions of *classifier* κ_4 for the instances in dataset $\mathcal{D}_4$.

Explanations of instances: $x_1..x_3$,

$x_1 : E_1 = \{(f_1, 0), (f_2, 1)\}$ *or* x_1

$x_2 : E_2 = \{(f_1, 1)\}$

$x_2 : E_3 = \{(f_2, 0)\}$

$x_2 : E_4 = x_2$

$x_3 : E_5 = \{(f_1, 1)\}$ *or* E_2

$x_3 : E_6 = x_3$

Here, $L_{cv}(\mathbf{Q}_1) = \{E_1\} = \{x_1\}$, explains that x_1 is only an Cv-dAXp itself i.e.$\{(f_1, 0), (f_2, 1)\}$ because individually $(f_1, 0)$ and $(f_2, 1)$ does not satisfy the condition of Cv-dAXp. So, the Cv-dAXp for the entire dataset $\mathcal{D}_4$ comprises the explanation set $\mathbb{E} = \{E_1, \ldots, E_6\}$. Specifically, $L_{cv}(\mathbf{Q}_1) = \{x_1\}$, $L_{cv}(\mathbf{Q}_2) = \{E_2, E_3, x_2\}$, and $L_{cv}(\mathbf{Q}_3) = \{E_5, x_3\}$. For more explanations of these results, please see Section 5.1.1.

5.1.1 Result Analysis of Cv-dAXp

We begin by considering X as the subset of *literals* associated with each instance in the dataset $\mathcal{D}_4$ presented in Table 5.2. Our goal is to analyze which X satisfies as a Cv-dAXp for a given instance. The *classifier* is formally defined as:

$$\kappa_4(f_1, f_2) = f_1 \vee (\neg f_2).$$

Next, we explore the Cv-dAXp's corresponding to the instance x_1, where:

$$x_1 = \{(f_1, 0), (f_2, 1)\} = (0, 1).$$

We aim to investigate how this particular subset of *literals*, x_1, behaves under the defined *classifier* κ_4 and identify the conditions under which it qualifies as a Cv-dAXp.

Assuming $v = x_1$ and considering other instances as u_i where $i = 1, 2, ...(n-1)$ and n represents total number of instances. We examine various subsets X of *literals* and evaluate their behaviors under the classifier κ_4.

1. If subset of *literal* $X = \{(f_1, 0)\}$ and $\tilde{X} = f_1$ that indicated by blue colored box in Table 5.2 then

- $v[\tilde{X}] = x_1[\tilde{X}] = x_1[f_1] = (0)$
- Possible combinations of v with all u:
  - if $u_1 = x_2(1, 0)$ then $\kappa_4(v[\tilde{X}] \cup u_1[F \setminus \tilde{X}]) \rightarrow \kappa_4(0, x_2[f_2]) \rightarrow \kappa_4(0, 0) = 1$ (Mismatch) $\neq \kappa_4(v)$
  - if $u_2 = x_3(1, 1)$ then $\kappa_4(v[\tilde{X}] \cup u_2[F \setminus \tilde{X}]) \rightarrow \kappa_4(0, x_3[f_2]) \rightarrow \kappa_4(0, 1) = 0$ (Matches) $= \kappa_4(v)$
- Result: Mismatch found, so this subset does not satisfy the condition of Cv-dAXp shown in Table 5.2

	$\mathcal{D}_4$	f_1	f_2	$\kappa_4(x_i)$	$\kappa_4(v[\tilde{X}] \cup u_i[F \setminus \tilde{X}])$
v	x_1	**0**	1	**0**	
u_1	x_2	1	**0**	1	1 ✗
u_2	x_3	1	**1**	1	0

Table 5.2: Presents the comparison of classifier κ_4 outcomes for instance v and its interactions with instances u_1 and u_2, based on subset literal $X = \{(f_1, 0)\}$. The very last column highlights inconsistency, indicated by a cross symbol (✗) for u_1, between the predicted and actual classifier results when $v[\tilde{X}]$ is combined with the remaining features from other instances.

2. If subset of *literal* $X = \{(f_2, 1)\}$ and $\tilde{X} = f_2$ that indicated by blue colored box in Table 5.3 then

- $v[\tilde{X}] = x_1[\tilde{X}] = x_1[f_2] = (1)$
- Possible combinations of v with all u:
  - if $u_1 = x_2(1, 0)$ then $\kappa_4(v[\tilde{X}] \cup u_1[F \setminus \tilde{X}]) \rightarrow \kappa_4(x_2[f_1], 1) \rightarrow \kappa_4(1, 1) = 1$ (Mismatch) $\neq \kappa_4(v)$
  - if $u_2 = x_3(1, 1)$ then $\kappa_4(v[\tilde{X}] \cup u_2[F \setminus \tilde{X}]) \rightarrow \kappa_4(x_3[f_1], 1) \rightarrow \kappa_4(1, 1) = 1$ (Mismatch) $\neq \kappa_4(v)$
- Result: Mismatch found, so this subset does not satisfy the condition of Cv-dAXp shown in Table 5.3

	$\mathcal{D}_4$	f_1	f_2	$\kappa_4(x_i)$	$\kappa_4(v[\tilde{X}] \cup u_i[F \setminus \tilde{X}])$
v	x_1	0	**1**	**0**	
u_1	x_2	**1**	0	1	1 ✗
u_2	x_3	**1**	1	1	1 ✗

Table 5.3: Presents the comparison of classifier κ_4 outcomes for instance v and its interactions with instances u_1 and u_2, based on subset literal $X = \{(f_2, 1)\}$. The very last column highlights inconsistency, indicated by a cross symbol (✗) for u_1 and u_2, between the predicted and actual classifier results when $v[\tilde{X}]$ is combined with the remaining features from other instances.

3. If subset of *literal* $X = \{(f_1, 0), (f_2, 1)\}$ and $\tilde{X} = f_1, f_2$ that indicated by blue colored box in Table 5.4 then

- $v[\tilde{X}] = x_1[\tilde{X}] = x_1[f_1, f_2] = (0, 1)$
- Possible combinations of v with all u:
  - if $u_1 = x_2(1, 0)$ then $\kappa_4(v[\tilde{X}] \cup u_1[F \setminus \tilde{X}]) \rightarrow \kappa_4(0, 1, x_2[\emptyset]) \rightarrow \kappa_4(0, 1) = 0$ (Matches) $= \kappa_4(v)$
  - if $u_2 = x_3(1, 1)$ then $\kappa_4(v[\tilde{X}] \cup u_2[F \setminus \tilde{X}]) \rightarrow \kappa_4(0, 1, x_3[\emptyset]) \rightarrow \kappa_4(0, 1) = 0$ (Matches) $= \kappa_4(v)$
- Result: Matches for all combinations. So this subset is an Cv-dAXp for instance x_1 shown in Table 5.4.

	$\mathcal{D}_4$	f_1	f_2	$\kappa_4(x_i)$	$\kappa_4(v[\tilde{X}] \cup u_i[F \setminus \tilde{X}])$
v	x_1	**0**	**1**	**0**	
u_1	x_2	1	0	1	0
u_2	x_3	1	1	1	0

Table 5.4: Presents the comparison of classifier κ_4 outcomes for instance v and its interactions with instances u_1 and u_2, based on subset literal $X = \{(f_1, 0), (f_2, 1)\}$. The very last column highlights consistency, between the predicted and actual classifier results when $v[\tilde{X}]$ is combined with the remaining features from other instances.

So, for instance x_1, the function $L_{cv}(\mathbf{Q}_1) = \{E_1\} = \{(f_1, 0), (f_2, 1)\} = \{x_1\}$ indicates that instance x_1 serves as a self-explanations of Cv-dAXp. For further explanations of others instances result shown in Table 5.5 and final result of Cv-dAXp shows in Table 5.6

Instance x_i	**Subset of literal** X	**Check** $\kappa(v[\tilde{X}] \cup u[F \setminus \tilde{X}]) = \kappa(v)$ **for all** u**, where** $\kappa_4(f_1, f_2) = f_1 \vee (\neg f_2)$	**Cv-dAXp**
if, $v = x_1 = (0, 1)$	$\{(f_1, 0)\}$	if, $u_1 = x_2(1, 0) \rightarrow \kappa_4(x_1[f_1] \cup x_2[f_2])$ $\rightarrow \kappa_4(0, 0) = 1$ **(Mismatch)** $\neq \kappa_4(x_1)$ if, $u_2 = x_3(1, 1) \rightarrow \kappa_4(x_1[f_1] \cup x_3[f_2])$ $\rightarrow \kappa_4(0, 1) = 0$ **(Matches)** $= \kappa_4(x_1)$	$\times$
	$\{(f_2, 1)\}$	if, $u_1 = x_2(1, 0) \rightarrow \kappa_4(x_2[f_1] \cup x_1[f_2])$ $\rightarrow \kappa_4(1, 1) = 1$ **(Mismatch)** $\neq \kappa_4(x_1)$ if, $u_2 = x_3(1, 1) \rightarrow \kappa_4(x_3[f_1] \cup x_1[f_2])$ $\rightarrow \kappa_4(1, 1) = 1$ **(Mismatch)** $\neq \kappa_4(x_1)$	$\times$
	$\{(f_1, 0), (f_2, 1)\}$	if, $u_1 = x_2(1, 0) \rightarrow \kappa_4(x_1[f_1, f_2] \cup x_2[\emptyset]) \rightarrow \kappa_4(0, 1) = 0$ **(Matches)** $= \kappa_4(x_1)$ if, $u_2 = x_3(1, 1) \rightarrow \kappa_4(x_1[f_1, f_2] \cup x_3[\emptyset]) \rightarrow \kappa_4(0, 1) = 0$ **(Matches)** $= \kappa_4(x_1)$	✓
if, $v = x_2 = (1, 0)$	$\{(f_1, 1)\}$	if, $u_1 = x_1(0, 1) \rightarrow \kappa_4(x_2[f_1] \cup x_1[f_2])$ $\rightarrow \kappa_4(1, 1) = 1$ **(Matches)** $= \kappa_4(x_2)$ if, $u_2 = x_3(1, 1) \rightarrow \kappa_4(x_2[f_1] \cup x_3[f_2])$ $\rightarrow \kappa_4(1, 1) = 1$ **(Matches)** $= \kappa_4(x_2)$	✓
	$\{(f_2, 0)\}$	if, $u_1 = x_1(0, 1) \rightarrow \kappa_4(x_1[f_1] \cup x_2[f_2])$ $\rightarrow \kappa_4(0, 0) = 1$ **(Matches)** $= \kappa_4(x_2)$ if, $u_2 = x_3(1, 1) \rightarrow \kappa_4(x_3[f_1] \cup x_2[f_2])$ $\rightarrow \kappa_4(1, 0) = 1$ **(Matches)** $= \kappa_4(x_2)$	✓
	$\{(f_1, 1), (f_2, 0)\}$	if, $u_1 = x_1(0, 1) \rightarrow \kappa_4(x_2[f_1, f_2] \cup x_1[\emptyset]) \rightarrow \kappa_4(1, 0) = 1$ **(Matches)** $= \kappa_4(x_2)$ if, $u_2 = x_3(1, 1) \rightarrow \kappa_4(x_2[f_1, f_2] \cup x_3[\emptyset]) \rightarrow \kappa_4(1, 0) = 1$ **(Matches)** $= \kappa_4(x_2)$	✓

Instance x_i	**Subset of literal** X	**Check** $\kappa(v[\tilde{X}] \cup u[F \setminus \tilde{X}]) = \kappa(v)$ **for all** u**, where** $\kappa_4(f_1, f_2) = f_1 \vee (\neg f_2)$	**Cv-dAXp**
if, $v = x_3 = (1, 1)$	$\{(f_1, 1)\}$	if, $u_1 = x_1(0, 1) \rightarrow \kappa_4(x_3[f_1] \cup x_1[f_2])$ $\rightarrow \kappa_4(1, 1) = 1$ **(Matches)** $= \kappa_4(x_3)$ if, $u_2 = x_2(1, 0) \rightarrow \kappa_4(x_3[f_1] \cup x_2[f_2])$ $\rightarrow \kappa_4(1, 0) = 1$ **(Matches)** $= \kappa_4(x_3)$	✓
	$\{(f_2, 1)\}$	if, $u_1 = x_1(0, 1) \rightarrow \kappa_4(x_1[f_1] \cup x_3[f_2])$ $\rightarrow \kappa_4(0, 1) = 0$ **(Mismatch)** $\neq \kappa_4(x_3)$ if, $u_2 = x_2(1, 0) \rightarrow \kappa_4(x_2[f_1] \cup x_3[f_2])$ $\rightarrow \kappa_4(1, 1) = 1$ **(Matches)** $= \kappa_4(x_3)$	×
	$\{(f_1, 1), (f_2, 1)\}$	if, $u_1 = x_1(0, 1) \rightarrow \kappa_4(x_3[f_1, f_2] \cup x_1[\emptyset]) \rightarrow \kappa_4(1, 1) = 1$ **(Matches)** $= \kappa_4(x_3)$ if, $u_2 = x_2(1, 0) \rightarrow \kappa_4(x_3[f_1, f_2] \cup x_2[\emptyset]) \rightarrow \kappa_4(1, 1) = 1$ **(Matches)** $= \kappa_4(x_3)$	✓

Table 5.5: Result Analysis of Cv-dAXp for the instances of dataset $\mathcal{D}_4$

Instance x_i	$L_{cv}(\mathbf{Q})$	**Explanations**
$x_1 = (0, 1)$	$\{E_1\}$	$\{(f_1, 0), (f_2, 1)\} = x_1$
$x_2 = (1, 0)$	$\{E_2, E_3, E_4\}$	$\{\{(f_1, 1)\}, \{(f_2, 0)\}, \{(f_1, 1), (f_2, 0)\}\}$
$x_3 = (1, 1)$	$\{E_5, E_6\}$	$\{\{(f_1, 1)\}, \{(f_1, 1), (f_2, 1)\}\}$

Table 5.6: Result of Cv-dAXp's for dataset $\mathcal{D}_4$

5.1.2 Time Complexity Analysis of Cv-dAXp

Analyzing the time complexity of an explainer in XAI is crucial for ensuring that explanations are generated efficiently, are practical for real-world use, and fit within the constraints of the system and user needs. It helps in achieving a balance between the quality of explanations and computational resources, thereby enhancing the overall effectiveness of the AI system.

The Cv-dAXp result analysis of dataset $\mathcal{D}_4$ in the Table 5.5 involves evaluating the satisfaction of a certain condition over various instances and subsets of literals. To determine the time complexity of this evaluation process, we need to analyze the operations involved and their impact on computational efficiency.

Problem Description

Given a set of instances $\{x_1, x_2, ..., x_n\}$, and a classifier κ_4 defined as $\kappa_4(f_1, f_2) = f_1 \vee (\neg f_2)$, our task was to check whether for a given subset of literals X, the condition $\kappa(v[\tilde{X}] \cup u[F \setminus \tilde{X}]) = \kappa(v)$ is true for all possible instances u. Here, v represents a specific instance, X is a subset of literals (features with values), $\tilde{X}$ subset of feature and F denotes the set of all features.

Components of the Analysis

1. **Number of Instances**: Assume there are n distinct instances to consider.
2. **Number of Literal Subsets**: For a set of m literals, there are 2^m possible subsets. Each subset represents a different configuration of literals, and thus, a different scenario to evaluate.
3. **Function Evaluation**: Evaluating the function of classifier κ involves logical operations. Specifically, for each subset X and instance u, we need to compute $\kappa(v[\tilde{X}] \cup u[F \setminus \tilde{X}])$. This computation is a constant-time operation, involving Boolean algebra (logical OR and NOT).
4. **Comparisons**: Each computed result from the classifier κ is compared to the value of $\kappa(v)$ for the instance v. Each comparison is also a constant-time operation.

Total Computational Effort

To compute whether the condition $\kappa(v[\tilde{X}] \cup u[F \setminus \tilde{X}])$ is satisfied for all instances u, the following steps are performed:

- For each of the n instances, we evaluate κ for all 2^m subsets of literals X.
- For each subset, we check the condition across all instances u, which involves $O(1)$ operations per check.

Thus, the overall computational effort is dominated by the number of subset evaluations, which is exponential in the number of literals due to 2^m possible subsets. The time complexity for processing all instances and subsets is $O(n \cdot 2^m)$. This reflects the challenge of dealing with exponential growth in the number of subsets as the number of literals increases, combined with the linear factor related to the number of instances.

5.1.3 Axioms / Formal Properties

In the literature [ACD24], we identified a set of axioms intended to theoretically evaluate feature-based *explainers*. These formal properties, which *explainers* can satisfy, are crucial for deepening our understanding of the explanation process overall. They clarify the fundamental assumptions of *explainers*, highlight their strengths and weaknesses, facilitate the comparison of different *explainers*, and help identify unexplored types of *explainers*.

Name of properties	Mathematical definitions
Success	$\forall \mathbf{Q} = \langle \mathbb{T}, \kappa, \mathcal{D}, x \rangle, \mathbf{L}(\mathbf{Q}) \neq \emptyset$.
Coherence	$\forall \mathbf{Q} = \langle \mathbb{T}, \kappa, \mathcal{D}, x \rangle$ and $\mathbf{Q}' = \langle \mathbb{T}, \kappa, \mathcal{D}, x' \rangle$ s.t. $\kappa(x) \neq \kappa(x')$, then $\forall E \in \mathbf{L}(\mathbf{Q}), \forall E' \in \mathbf{L}(\mathbf{Q}')$, $E \cup E'$ is inconsistent.
Irreducibility	$\forall \mathbf{Q} = \langle \mathbb{T}, \kappa, \mathcal{D}, x \rangle, \forall E \in \mathbf{L}(\mathbf{Q}), \forall l \in E, \exists x' \in \mathcal{D}$ s.t. $\kappa(x') \neq \kappa(x)$ and $E \setminus \{l\} \subseteq x'$.
Strong Irreducibility	$\forall \mathbf{Q} = \langle \mathbb{T}, \kappa, \mathcal{D}, x \rangle, \forall E \in \mathbf{L}(\mathbf{Q}), \forall l \in E, \exists x' \in \mathbb{F}(\mathbb{T})$ s.t. $\kappa(x') \neq \kappa(x)$ and $E \setminus \{l\} \subseteq x'$.
Completeness	$\forall \mathbf{Q} = \langle \mathbb{T}, \kappa, \mathcal{D}, x \rangle, \forall E \subseteq x$, if $E \notin \mathbf{L}(\mathbf{Q})$, then $\exists y \in \mathcal{D}$ s.t. $E \subseteq y$ and $\kappa(y) \neq \kappa(x)$.
Strong Completeness	$\forall \mathbf{Q} = \langle \mathbb{T}, \kappa, \mathcal{D}, x \rangle, \forall E \subseteq x$, if $E \notin \mathbf{L}(\mathbf{Q})$, then $\exists y \in \mathbb{F}(\mathbb{T})$ s.t. $E \subseteq y$ and $\kappa(y) \neq \kappa(x)$.
Feasibility	$\forall \mathbf{Q} = \langle \mathbb{T}, \kappa, \mathcal{D}, x \rangle, \forall E \in \mathbf{L}(\mathbf{Q}), E \subseteq x$.
Validity	$\forall \mathbf{Q} = \langle \mathbb{T}, \kappa, \mathcal{D}, x \rangle, \forall E \in \mathbf{L}(\mathbf{Q}), \nexists y \in \mathcal{D}$ s.t. $E \subseteq y$ and $\kappa(y) \neq \kappa(x)$.
Monotonicity	$\forall \mathbf{Q} = \langle \mathbb{T}, \kappa, \mathcal{D}, x \rangle, \forall \mathbf{Q}' = \langle \mathbb{T}, \kappa, \mathcal{D}', x \rangle$, if $\mathcal{D} \subseteq \mathcal{D}'$, then $\mathbf{L}(\mathbf{Q}) \subseteq \mathbf{L}(\mathbf{Q}')$.
Counter-Monotonicity (CM)	$\forall \mathbf{Q} = \langle \mathbb{T}, \kappa, \mathcal{D}, x \rangle, \forall \mathbf{Q}' = \langle \mathbb{T}, \kappa, \mathcal{D}', x \rangle$, if $\mathcal{D} \subseteq \mathcal{D}'$, then $\mathbf{L}(\mathbf{Q}') \subseteq \mathbf{L}(\mathbf{Q})$.

Table 5.7: Formal Properties of Explainer $\mathbf{L}$.

Table 5.7 represents the formal properties of an explainer. It contains two columns: the left-side column lists the property names (e.g., "Success", "Coherence"), while the rightcolumn provides their formal mathematical definitions or descriptions. For an explainer $\mathbf{L}$, it must satisfy to the mathematical notations and conditions associated with each property listed in Table 5.7. A brief explanationof the properties is given below.

- **Success**: For every question $\mathbf{Q}$, the explainer $\mathbf{L}(\mathbf{Q})$ produces a non-empty result.
- **Coherence**: For any two questions $\mathbf{Q}$ and $\mathbf{Q}'$ where $\kappa(x) \neq \kappa(x')$, all explanations $E \in \mathbf{L}(\mathbf{Q})$ and $E' \in \mathbf{L}(\mathbf{Q}')$ are mutually inconsistent.
- **Irreducibility**: For every explanation E in $\mathbf{L}(\mathbf{Q})$ and every element l in E, there exists an instance x' in the dataset $\mathcal{D}$ that removing l from E results in a subset of x' where $\kappa(x') \neq \kappa(x)$.
- **Strong Irreducibility**: Similar to Irreducibility, but the instance x' belongs to the set $\mathbb{F}(\mathbb{T})$.
- **Completeness**: If a subset E of x is not an explanation in $\mathbf{L}(\mathbf{Q})$, there exists an instance y in the dataset $\mathcal{D}$ that E is a subset of y and $\kappa(y) \neq \kappa(x)$.
- **Strong Completeness**: Similar to Completeness, but the instance y belongs to the set $\mathbb{F}(\mathbb{T})$.
- **Feasibility**: Every explanation E in $\mathbf{L}(\mathbf{Q})$ is a subset of x.
- **Validity**: There does not exist any instance y in the dataset $\mathcal{D}$ that E is a subset of y and $\kappa(y) \neq \kappa(x)$.
- **Monotonicity**: If the dataset $\mathcal{D}$ is a subset of another dataset $\mathcal{D}'$, then the explanations in $\mathbf{L}(\mathbf{Q})$ are also a subset of those in $\mathbf{L}(\mathbf{Q}')$.
- **Counter-Monotonicity (CM)**: If the dataset $\mathcal{D}$ is a subset of another dataset $\mathcal{D}'$, then the explanations in $\mathbf{L}(\mathbf{Q}')$ are a subset of those in $\mathbf{L}(\mathbf{Q})$.

These properties provide a rigorous framework for evaluating and comparing different explainers.

5.1.4 Proof of Cv-dAXp's properties

The newly proposed explainer, Cv-dAXp, meets several important properties, as demonstrated in Table 5.8. This table presented a comprehensive comparison of different explainers and their properties, specifying whether each explainer fulfills particular criteria. The explainers included in this comparison are $\mathbf{L}_{dA}$, $\mathbf{L}_{irr}$, $\mathbf{L}_{su}$, $\mathbf{L}_{cv}$.

Explainers / Axioms	$\mathbf{L}_{dA}$	$\mathbf{L}_{irr}$	$\mathbf{L}_{su}$	$\mathbf{L}_{cv}$
Success	✓	✓	✓	✓
Irreducibility	✓	×	×	×
Strong Irreducibility	✓	×	×	×
Coherence	×	✓	✓	×
Feasibility	✓	✓	✓	✓
Validity	✓	✓	✓	✓
Completeness	✓	×	×	×
Strong Completeness	✓	×	×	×
Monotonicity	−	×	×	×
Counter-Monotonicity	−	×	×	×

Table 5.8: This table lists the various properties of different explainer in the very-left column, while the remaining columns show different explainers and indicate whether each property is satisfied. Here, a checkmark (✓) indicates that the corresponding property is satisfied while a crossmark (×) signifies that it is not. For unknown cases, a hyphen (−) is used.

The Cv-dAXp *explainer*, denoted as $\mathbf{L}_{cv}$, satisfies the principles of properties *Success*, *Validity*, and *Feasibility* but does not fulfil the conditions of the remaining axioms outlined in Table 5.8.

- ***Success***: According to the principle of *Success*, $\mathbf{L}_{cv}$ must provide an *explanation* for any *question* $\mathbf{Q} = \langle \mathbb{T}, \kappa, \mathcal{D}, v \rangle$. For example, the instance v qualifies as a Cv-dAXp by itself because feature set F always constitutes a Cv-dAXp. This is evident since $v[F] \cup u[F \setminus F] = v$ (e.g. Table 5.4). Therefore, $\mathbf{L}_{cv}$ satisfies the Success property by always providing a non-empty explanation set.
- ***Feasibility***: The *Feasibility* axiom requires that an explanation must be contained within the instance it explains. By definition of Cv-dAXp, if X is a Cv-dAXp of instance v, it must satisfy $X \subseteq v$. This ensures that the explanation is a subset of the features of v, meaning it is derived entirely from the instance itself (e.g. Table 5.6). Hence, the Cv-dAXp satisfies the *Feasibility* condition, as it inherently restricts the explanation to be a subset of the instance's features.
- ***Validity***: This property ensures that the explanations maintain local consistency across the dataset. If instance u and instance v agrees on the feature subset $\tilde{X}$, then the output of $v[\tilde{X}] \cup$

$u[F \setminus \tilde{X}]$ will be u (e.g. Table 5.3). This implies that the feature subset $\tilde{X}$, which explains v, also maintains consistency when applied to u. Thus, the Cv-dAXp satisfies the *Validity* property by ensuring that explanations remain consistent and can be generalized to other instances in the dataset that sharing the same feature subset $\tilde{X}$.

- ***Coherence***: *Coherence* requires that explanations for two different outcomes cannot simultaneously hold true. For example, consider two *questions* $\mathbf{Q}_1 = \langle \mathbb{T}, \kappa_5, \mathcal{D}_5, v \rangle$ and $\mathbf{Q}_2 = \langle \mathbb{T}, \kappa_5, \mathcal{D}_5, u \rangle$, where $\kappa_5(v) \neq \kappa_5(u)$. If the Cv-dAXp *function* $\mathbf{L}_{cv}$ yields two explanations E_1 and E_2, for these questions, the explanations E_1 and E_2 are consistent. As a result, Cv-dAXp fails to fulfill the requirement of *coherence* property.

  – Proof: Consider the *classifier* $\kappa_5(f_1, f_2, f_3) = (f_1 \wedge f_2 \wedge f_3) \vee (\neg f_1 \wedge \neg f_2 \wedge f_3)$ and the dataset $\mathcal{D}_5$ in Table 5.4.

  - $\{(f_1, 0)\}$ is a Cv-dAXp of u since $\kappa_5(0, 1, 1) = \kappa_5(0, 0, 0) = 0$ (shown in Table 5.10)
  - $\{(f_3, 1)\}$ is a Cv-dAXp of v since $\kappa_5(0, 0, 1) = \kappa_5(1, 1, 1) = 1$ (shown in Table 5.11)

$\mathcal{D}_5$	f_1	f_2	f_3	κ_5
u	0	0	0	0
v	1	1	1	1

Table 5.9: Predictions of *classifier* κ_5 for the instances in dataset $\mathcal{D}_5$.

$\mathcal{D}_5$	f_1	f_2	f_3	κ_5	$\kappa_5(u[f_1, f_2] \cup v[f_3])$
u	**0**	**0**	0	0	1
v	1	1	**1**	**1**	

Table 5.11: Prediction of *classifier* κ_5 of u when $(f_3, 1)$.

$\mathcal{D}_5$	f_1	f_2	f_3	κ_5	$\kappa_5(u[f_1] \cup v[f_2, f_3])$
u	**0**	0	0	**0**	
v	1	**1**	**1**	1	0

Table 5.10: Prediction of *classifier* κ_5 of v when $(f_1, 0)$.

For instance, $u : E_1 = \{(f_1, 0)\}$

$v : E_2 = \{(f_3, 1)\}$

Here, $\kappa_5(u) \neq \kappa_5(v)$ and the set $\{E_1, E_2\}$ is incoherent since $E_1 \cup E_2$ is consistent, thus coherence is not satisfied.

- ***(Strong) Irreducibility***: *Irreducibility* refers to the property of an explanation set E where removing any single element l from E leads to a subset that corresponds to another instance x' in the dataset $\mathcal{D}$ (resp. $x' \in \mathbb{F}(\mathbb{T})$), produces a different classification outcome $\kappa(x') \neq \kappa(x_1)$. So, *Irreducibility* requires that such a different behavior must exist for some $x' \in \mathcal{D}$ that contains a subset of E. However, Cv-dAXp does not meet the conditions of *irreducibility*, thus it satisfy the condition of *Validity* which disallows the existence of any $x' \in \mathcal{D}$ with a diiferent behavior (i.e., $\kappa(x') \neq \kappa(x)$) that contains E or any subset of it.

- ***(Strong) Completeness***: This property ensures that the *explainer* is capable of generating explanations for a wide range of instances, even when some subsets of these instances are considered invalid or do not align with the explanations for a specific *question*. In particular, it asserts that for any subset E of an instance x that is not identified as an explanation in $\mathbf{L}(\mathbf{Q})$ (i.e., E is not valid or logical according to question $\mathbf{Q}$), there must exist another instance y in the dataset $\mathcal{D}$ (or $y' \in \mathbb{F}(\mathbb{T})$) that contains the subset E and is assigned a different classification from x (i.e., $\kappa(y) \neq \kappa(x)$). On the other hand, Cv-dAXp fulfills the *Feasibility* property, which mandates that explanations be strictly confined to the features relevant to the instance they are explaining. Thus, satisfying both the *(Strong) Completeness* and *Feasibility* properties simultaneously presents a logical contradiction for an *explainer*.

- ***Monotonicity***: This property guarantees that if an explanation is valid for the smaller dataset $\mathcal{D}$, it remains valid when the dataset expands to $\mathcal{D}'$. The validity of a subset X as a Cv-dAXp explanation in $\mathbf{L}_{cv}$ depends on the dataset $\mathcal{D}$ and the condition that the classifier's output remains unchanged when $\tilde{X}$ is combined with other subsets u from $\mathcal{D}$. If the dataset expands to $\mathcal{D}'$ (where $\mathcal{D} \subseteq \mathcal{D}'$), new elements may alter whether the condition $\kappa(v[\tilde{X}] \cup u[F \setminus \tilde{X}]) = \kappa(v)$ holds. Since explanations are tied to the specific dataset $\mathcal{D}$, expanding to $\mathcal{D}'$ can invalidate explanations that were previously valid. Therefore, $\mathbf{L}_{cv}$ can violate *Monotonicity* because explanations valid for $\mathcal{D}$ may no longer meet the required conditions when the dataset changes, disrupting consistency.

- ***Counter-Monotonicity***: The *Counter-Monotonicity* property asserts that if $\mathcal{D} \subseteq \mathcal{D}'$, then the set of explanations $\mathbf{L}(\mathbf{Q}')$ is a subset of $\mathbf{L}(\mathbf{Q})$. The explainer $\mathbf{L}_{cv}$ potentially violates the *Counter-Monotonicity* property because, as $\mathcal{D}$ increases to $\mathcal{D}'$, explanations may be removed rather than refined, that contradicting the expectation that $\mathbf{L}(\mathbf{Q}')$ should always be a subset of $\mathbf{L}(\mathbf{Q})$. The condition $\forall u \in \mathcal{D}$ forces explanations to adapt or be excluded as the dataset changes, making $\mathbf{L}(\mathbf{Q}')$ possibly larger or fundamentally different from $\mathbf{L}(\mathbf{Q})$, rather than contained within it.

In summary, while the Cv-dAXp explainer, $\mathbf{L}_{cv}$, demonstrates adherence to the principles of *Success*, *Feasibility*, and *Validity* by ensuring consistent, relevant, and localized explanations, it fails to meet the criteria for *Coherence*, *(Strong) Irreducibility*, *(Strong) Completeness*, *Monotonicity*, and *Counter-Monotonicity*.

5.2 Evaluation and discussion

In this study, we evaluated the performance of Cv-dAXp, using the well-established zoo dataset [Zoo], which contains 101 instances of zoo animals categorized into seven distinct classes: Mammal, Bird, Reptile, Fish, Amphibian, Bug, and Invertebrate. Each animal is represented by 16 features, and the classifier κ is assumed to classify the animals accurately across all 101 instances.

We assume that the classifier κ correctly classifies all 101 instances in the dataset. The complexity of exhaustively evaluating Cv-dAXp across all feature combinations for each instance highlights the challenges posed by the large search space. It is impractical to show an exhaustive search over all 101 instances in the Table 5.12. Consequently, instead of presenting a full exhaustive analysis, we selected

		animal_name	hair	feathers	eggs	milk	airborne	aquatic	predator	toothed	backbone	breathes	venomous	fins	legs	tail	domestic	catsize	class_type
v	x_1	aardvark	1	0	0	1	0	0	1	1	1	1	0	0	4	0	0	1	1
u_1	x_2	antelope	1	0	0	1	0	0	0	1	1	1	0	0	4	1	0	1	1
u_2	x_3	bass	0	0	1	0	0	1	1	1	1	0	0	1	0	1	0	0	4
u_3	x_4	bear	1	0	0	1	0	0	1	1	1	1	0	0	4	0	0	1	1
u_4	x_5	boar	1	0	0	1	0	0	1	1	1	1	0	0	4	1	0	1	1
u_5	x_6	buffalo	1	0	0	1	0	0	0	1	1	1	0	0	4	1	0	1	1
u_6	x_7	calf	1	0	0	1	0	0	0	1	1	1	0	0	4	1	1	1	1
u_7	x_8	carp	0	0	1	0	0	1	0	1	1	0	0	1	0	1	1	0	4
u_8	x_9	catfish	0	0	1	0	0	1	1	1	1	0	0	1	0	1	0	0	4
u_9	x_{10}	cavy	1	0	0	1	0	0	0	1	1	1	0	0	4	0	1	0	1
u_{10}	x_{11}	cheetah	1	0	0	1	0	0	1	1	1	1	0	0	4	1	0	1	1
u_{11}	x_{12}	chicken	0	1	1	0	1	0	0	0	1	1	0	0	2	1	1	0	2
u_{12}	x_{13}	chub	0	0	1	0	0	1	1	1	1	0	0	1	0	1	0	0	4
u_{13}	x_{14}	clam	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	7
u_{14}	x_{15}	crab	0	0	1	0	0	1	1	0	0	0	0	0	4	0	0	0	7

Table 5.12: This table comprises 15 of 101 instances of zoo animals, each characterized by 16 features and classified into one of seven categories (class_type): (1) Mammal, (2) Bird, (3) Reptile, (4) Fish, (5) Amphibian, (6) Bug, and (7) Invertebrate.

a small subset of the dataset to illustrate the explainer's functionality. So, we begin by considering some examples help clarify how Cv-dAXp operates in practical terms. By doing so, we proposed insight into how Cv-dAXp generates explanations and evaluates which features satisfy the condition.

To perform this evaluation, we establish the following classifier based on the pattern analysis of the zoo dataset. Assume the classifier $\kappa(x_i)$ is defined as follows:

$$\kappa(x_i) = \begin{cases} \textbf{\textit{Mammal}} : \{(milk, 1)\} \\ \textbf{\textit{Bird}} : \{(feathers, 1)\} \\ \textbf{\textit{Reptile}} : \{(milk, 0), (airborne, 0), (fins, 0), ((legs, 0), (legs, 4)), (tails, 1)\} \\ \textbf{\textit{Fish}} : \{(breathes, 0), (fins, 1)\} \\ \textbf{\textit{Amphibian}} : \{(eggs, 1), (aquatic, 1), (toothed, 1), (backbone, 1), (breathes, 1)\} \\ \textbf{\textit{Bug}} : \{(breathes, 1), (legs, 6)\} \\ \textbf{\textit{Invertebrate}} : \{(airborne, 0), (backbone, 0), (not((breathes, 1), (legs, 6)))\} \end{cases}$$

Where κ denotes a classifier and x_i represents each individual instance, with i ranging from 1 to n.

For instance, the explanation generated by Cv-dAXp for the animal *aardvark* focuses on the feature {(milk, 1)}. This means that, for $v = aardvark$ and $X = \{(milk, 1)\}$, after evaluating all other

u_i (where i ranges from 1 to n), assuming $\{(milk, 1)\}$ for all u_i, produces instance that would all be classified as a mammal by the classifier $\kappa(x_i)$. So, all other animals in the dataset that produce milk are also mammals and *literal* $X = \{(milk, 1)\}$ is a Cv-dAXp for $aardvark$.

Similarly, the Cv-dAXp for the animal *bass* is {(breathes, 0), (fins, 1)} as all animals in the dataset that do not breathe and have fins are classified as fish. So, the model successfully identifies mammal and fish class. Likewise, we evaluate the model's success across all other classes, thereby reinforcing the validity of the explanation under the $\mathbf{L}_{cv}$ function. This demonstrates the explainer's ability to identify key features crucial for classification.

The examples discussed reveal both the strengths and limitations of Cv-dAXp. On one hand, the explainer effectively identifies crucial features that differentiate classes, providing clear and concise justifications for the classifications made by the black-box classifier. On the other hand, the inherent complexity of evaluating large datasets with many features underscores the difficulty of achieving global completeness in every case. While the examples provided demonstrate that Cv-dAXp satisfies the formal properties for individual instances, the challenge of extending this to a broader dataset remains significant.

This analysis highlights the importance of selecting representative subsets of data to evaluate and explain complex classifiers. By narrowing the focus, we were able to gain a better understanding of Cv-dAXp's capabilities without being overwhelmed by the vast combinatorial possibilities. Future work could explore strategies to optimize the evaluation process for larger datasets while ensuring that the explainer remains robust across a wide range of instances.

Innovation Aspects

Contents

6.1 Innovative idea

In this internship project, we study on a novel approach to explainability in machine learning by exploring an explainer, Cv-dAXp, designed to generate concise, abductive explanations for complex classification tasks. The main characteristics of Cv-dAXp lies in its ability to identify and present sufficient reasons for the classification decisions made by black-box classifiers. Traditional machine learning models often operate as black boxes, providing little insight into how they reach their decisions. This lack of transparency can be problematic, particularly in sensitive domains such as healthcare or finance. Cv-dAXp addresses this issue by systematically identifying key features that contribute to a given classification, offering users a clearer understanding of the underlying logic behind the model's decisions.

The innovative aspect of Cv-dAXp is its use of abductive reasoning, which differs from traditional explainability methods. Instead of simply correlating features with outcomes, Cv-dAXp attempts to infer the most plausible explanations for the classification of a given instance, grounded in a formal set of properties. This approach enables the explainer to focus on the sufficiency of features in generating explanations, ensuring that the explanation not only justifies the model's decision but does so in a way that is concise and informative. This method allows users to gain insights into the model's reasoning process without overwhelming them with extraneous details, making Cv-dAXp particularly useful for interpreting complex models in data-rich environments.

Importantly, while evaluating the cross-validation of a feature in Section 5.2 (e.g., $v =$ aardvark and $X = \{(milk, 1)\}$), Cv-dAXp produces a new instance for all other u_i by changing the value of milk into 1. So, this approach represents an innovative aspect of the method. This means that Cv-dAXp is also capable of generating and classifying new instances.

6.2 Implementation

The implementation of the Cv-dAXp method was carried out in several steps, beginning with the formalization of the explainer and then building a framework to test its efficacy. The explanation method, represented by L_{cv}, generates a set of Cv-dAXp explanations by applying a logical and mathematical formulation to the classifier's output.

We can define the cross-validated explainer L_{cv} as a function that, given any "question" or instance $\mathbf{Q} = \langle \mathbb{T}, \kappa, \mathcal{D}, v \rangle$, will generate a set of explanations $\mathbb{E}$. These explanations consist of literals—specific feature-value pairs—that the classifier κ uses to make a decision about the instance v. The key challenge here is ensuring that the explanation holds true across different combinations of data instances and features within the dataset $\mathcal{D}$.

Mathematically, this is expressed through the condition:

$$\forall u \in \mathcal{D}, \quad \kappa(v[\tilde{X}] \cup u[F \setminus \tilde{X}]) = \kappa(v)$$

This condition ensures that for every instance u in the dataset $\mathcal{D}$, the classifier's decision on the combined feature set $v[\tilde{X}]$ and $u[F \setminus \tilde{X}]$ remains consistent with the classifier's decision on the original instance v. If this condition holds, X, the selected set of literals, qualifies as a cross-validated abductive explanation.

6.2.1 Dataset and Model Selection

The zoo [Zoo] dataset, a well-known dataset in machine learning containing 101 instances of animals with 16 features each, was selected for evaluating the performance of Cv-dAXp. The animals in the dataset are classified into one of seven classes: Mammal, Bird, Reptile, Fish, Amphibian, Bug, and Invertebrate.

For this implementation, we assume a pre-trained classifier κ, which had been optimized for this dataset and was assumed to classify the animals correctly across all instances. The classifier's output served as the basis for generating explanations using the Cv-dAXp method.

6.2.2 Algorithm Design and Computation

The algorithm begins by selecting a specific instance v from the dataset. It then identifies a subset of features that are likely to explain the classifier's decision for this instance. The next step involves cross-validating these features by combining them with features from other instances u within the dataset. If the classifier's decision remains unchanged across these combinations, the feature subset is validated as an explanation.

The cross-validation step is computationally intensive, as it requires evaluating all possible combinations of features for each instance in the dataset. Therefore, we do not present all possible examples. To make the discussion more concrete, we provided a few selected examples. For example, in the case of explaining why an animal is classified as a "mammal," the explainer might identify the feature $(milk = 1)$ as critical. This feature is then cross-validated by testing whether it consistently explains

the classification of all mammals in the dataset, while not misclassifying any other class of animals. If the cross-validation holds, the feature is confirmed as a valid explanation. Similarly, for the animal bass, the Cv-dAXp explanation is {(breathes, 0), (fins, 1)}, highlighting that all animals in the dataset that do not breathe and possess fins are classified as fish. Likewise, we assess the model's performance across all other classes, further supporting the explanation's validity under the $\mathbf{L}_{cv}$ function. This highlights the explainer's capability to identify key features essential for accurate classification.

6.3 Working method

During the internship, I was responsible for adapting the original method proposed by my supervisors to better suit the specific challenges presented by the zoo [Zoo] dataset and the classifier κ. This involved analysis of the mathematical formulation of the explainer and implementing cross-validation procedures.

The working method was iterative, following an agile development approach. I began by testing the original explainer on small subsets of the dataset to understand its behavior and limitations. From these initial trials, I was able to identify several key areas for improvement, including the need for better feature selection techniques and more efficient cross-validation procedures.

One of the challenges encountered during this process was ensuring that the explanations generated were both consistent and meaningful. In some cases, the cross-validation procedure would return explanations that were overly complex, involving large subsets of features. To address this, I analyze explanation generation process to make it smaller, more interpretable feature subsets.

Evaluation and Testing: The method was thoroughly evaluated using the full zoo dataset. Each instance of the dataset was used as a test case, with the classifier κ providing the baseline decision. For each instance, the Cv-dAXp method generated an explanation, which was then cross-validated against the rest of the dataset.

The results were promising, showing that the Cv-dAXp method was able to generate consistent and reliable explanations for the vast majority of instances in the dataset.

Assessment

Contents

7.1 Achievements

Through this internship and research project, I studied a framework for Explainable Artificial Intelligence (XAI), specifically addressing the explainability of black-box AI models. This work involved studying an explainer called Cross Validated Data Abductive Explanation (Cv-dAXp) and enhancing the transparency and interpretability of machine learning models. My efforts contributed to bridging the gap between accurate performance and contextual understanding in AI systems, fostering trust and ethical usage in various real world applications.

7.2 Learning Outcomes

- **Understanding Black-Box AI Models:** Developed a deep understanding of the complexities and limitations of black-box AI models, particularly in terms of their decision-making processes.
- **Explainability Function:** Gained the ability to create and define a formal function for explaining the behavior of AI models, ensuring transparency and comprehensibility of automated systems.
- **Explainer Design:** Learned to design and analyze explainers, specifically Cv-dAXp, which enhance the interpretability of machine learning predictions by offering coherent and contextually relevant insights.
- **Feature-Based Explanation Techniques:** Acquired knowledge in feature-based explanation methods, understanding how individual model features influence predictions and how to identify biases and errors in AI systems.
- **Interdisciplinary Approach:** Mastered the ability to apply theoretical AI research in practical, making AI systems more accessible to both experts and non-experts.

7.3 Skill Improvements

- **AI Model Analysis and Interpretation:** Enhanced skills in analyzing complex AI models and deciphering their decision-making processes.
- **Explainability Algorithm Development:** Improved proficiency in developing and evaluating algorithms for explainability, specifically focusing on creating novel explainers that provide meaningful interpretations of AI models.
- **Cross Validated Data Abductive Explanation (Cv-dAXp) Techniques:** Advanced the ability to use cross validated data abductive explanation (Cv-dAXp) techniques in the context of explanation generation, ensuring that AI model predictions are consistent and reliable.
- **Research and Modeling:** Strengthened research skills, especially in formalizing abstract theories for AI explainability and developing universal principles applicable across different AI models.
- **Communication of Complex Concepts:** Refined the ability to explain complex technical concepts in AI and machine learning to a diverse audience, from experts to non-experts, enhancing clarity and accessibility.

7.4 Relevant Courses

- **Advanced Topics in Artificial Intelligence:** This course provides an in-depth understanding of cutting-edge AI models, especially black-box models like deep neural networks. The knowledge gained here helps in developing the theoretical framework for understanding and explaining AI models' decision-making processes. Topics such as machine learning algorithms, neural networks, and optimization directly inform your work on enhancing model interpretability.
- **Innovative Software Methods:** Developing a novel explainer for black-box models requires proficiency in innovative software engineering practices. This course equips with techniques for building scalable, flexible, and sustainable software architectures that can accommodate complex AI systems and their explainers. It also supports the integration of explainability methods within real-world AI applications, ensuring their practical viability.
- **Innovative Data Management:** Explainability often involves managing and interpreting large datasets used to train AI models. This course helps to understand how to efficiently handle and process data to ensure that our explainability models are built on solid data management principles. It also ensures that the explanations provided by AI systems are based on meaningful patterns within the data rather than noise or irrelevant factors.
- **Internet of Things (IoT):** AI is increasingly integrated into IoT systems, where trust and explainability are essential for decision-making in areas like smart homes, autonomous vehicles, and healthcare devices. This internship benefited from this course by focusing on how explainable AI can be applied to IoT systems, ensuring that the decisions made by AI in these environments are both accurate and understandable to users.

- **Research Workshop:** This internship is research-focused, and this course provides with the methodological skills necessary to conduct rigorous research. It helps to learned how to structure the research, design experiments, and evaluate the outcomes, which is essential for advancing the theoretical and practical aspects of explainability in AI systems.
- **Strategy:** The strategy course is crucial for positioning any work within the broader AI landscape. It helps this internship understand how to align this internship research with the strategic goals of organizations that implement AI systems, ensuring that explainability becomes a core component of AI strategies. This perspective is vital for driving the adoption of XAI in various industries, enhancing trust and compliance with regulatory frameworks.
- **Sustainable Information Systems:** One of your internship's goals is to ensure that AI models are sustainable for future use. This course equips with the knowledge to design information systems, including AI models, that are environmentally, socially, and economically sustainable. Ensuring explainability contributes to the long-term sustainability of AI systems by improving transparency, trust, and accountability.
- **User Interface and User Experience (UI/UX):** Explainable AI is about making AI decisions understandable to users. These UI/UX skills are essential for designing interfaces that effectively communicate the inner workings of AI models to non-experts. This is especially important in developing user-friendly explainability tools that can bridge the gap between complex AI models and the users who interact with them, ensuring that explanations are accessible, intuitive, and actionable.

Each course I completed plays a key role in enhancing my ability to tackle the challenges of XAI in my internship, contributing to the development of a robust, transparent, and sustainable AI framework.

Conclusion

8.1 Conclusion

In conclusion, this internship has been an invaluable exploration into the critical and rapidly growing field of Explainable Artificial Intelligence (XAI). While AI models are increasingly being integrated into decision-making processes across industries, the challenge remains in ensuring that these models are not just accurate, but also transparent and interpretable. Throughout this internship, significant progress was made in developing a framework that treats AI models as black-box entities while focusing on explaining their decision-making processes. This framework, along with the novel *Cross Validated Data Abductive Explanation (Cv-dAXp)* method proposed by my supervisors, contributes to demystifying complex models, providing consistent and transparent explanations that enhance trust and accountability.

This work not only bridges the gap between performance accuracy and interpretability but also lays the foundation for more robust and reliable AI systems. By establishing a universal framework for explainability, this internship enables AI models to become more accessible and understandable to both technical experts and non-experts, fostering greater trust in automated systems. The contributions made here are vital for ensuring compliance with legal and ethical standards and for promoting the responsible use of AI in critical domains such as healthcare, finance, and autonomous systems.

8.2 Future work

Looking ahead, several key areas for future work have emerged from this internship. First, there is potential to further refine and extend the Cv-dAXp explainer to improve its efficiency and applicability across a wider range of AI models. Future research could focus on optimizing the explainer's performance in real-world scenarios, particularly in large-scale and high-stakes applications.

Additionally, the formal framework developed in this internship can be expanded to encompass new types of AI models and explainers, exploring how it can be adapted to emerging AI technologies such as reinforcement learning and generative models.

Finally, a deeper exploration of the ethical implications of AI explainability is warranted, particularly in sensitive fields where decisions impact human lives. Ensuring that AI models not only perform accurately but also adhere to principles of fairness, transparency, and accountability will be essential for the continued advancement and responsible use of AI technologies.

This internship has laid a foundation for future advancements in XAI, offering valuable contri-

butions to the theoretical and practical understanding of AI explainability and setting the stage for continued innovation in this critical field.

Bibliography

[ACD24] Leila Amgouda, Martin C Cooper, and Salim Debbaoui. "Axiomatic Characterisations of Sample-based Explainers". In: *arXiv preprint arXiv:2408.04903* (2024) (cit. on pp. 11, 12, 14, 27).

[Amg23] Leila Amgoud. "Explaining black-box classifiers: Properties and functions". In: *International Journal of Approximate Reasoning* 155 (2023), pp. 40–65 (cit. on p. 14).

[AMT23] Leila Amgoud, Philippe Muller, and Henri Trenquier. "Leveraging Argumentation for Generating Robust Sample-based Explanations." In: *IJCAI*. 2023, pp. 3104–3111 (cit. on p. 14).

[CA23] Martin Cooper and Leila Amgoud. "Abductive Explanations of Classifiers under Constraints: Complexity and Properties". In: *26th European Conference on Artificial Intelligence (ECAI 2023)*. IOS Press. 2023, à–paraître (cit. on p. 14).

[Ins24] Institut de Recherche en Informatique de Toulouse. *IRIT - Overview*. https://www.irit.fr. Accessed: 2024-08-10. 2024 (cit. on p. 3).

[Kam21] Margot E Kaminski. "The right to explanation, explained". In: *Research Handbook on Information Law and Governance*. Edward Elgar Publishing, 2021, pp. 278–299 (cit. on p. 8).

[Mil19] Tim Miller. "Explanation in artificial intelligence: Insights from the social sciences". In: *Artificial intelligence* 267 (2019), pp. 1–38 (cit. on p. 14).

[Zoo] *UCI Machine Learning Repository*. Accessed: 2024-08-25. 2024 (cit. on pp. 31, 36, 37).