Influence-Driven Explanations for Bayesian Network Classifiers

Antonio Rago[1], Emanuele Albini[1], Pietro Baroni[2] and Francesca Toni[1]
[1] Dept. of Computing, Imperial College London, UK
[2] Dip.to di Ingegneria dell'Informazione, Università degli Studi di Brescia, Italy
{a.rago15, emanuele, ft}@imperial.ac.uk, pietro.baroni@unibs.it

Abstract

One of the most pressing issues in AI in recent years has been the need to address the lack of explainability of many of its models. We focus on explanations for discrete Bayesian network classifiers (BCs), targeting greater transparency of their inner workings by including *intermediate* variables in explanations, rather than just the input and output variables as is standard practice. The proposed *influence-driven explanations* (IDXs) for BCs are systematically generated using the causal relationships between variables *within the BC*, called influences, which are then categorised by logical requirements, called *relation properties*, according to their behaviour. These relation properties both provide guarantees beyond heuristic explanation methods and allow the information underpinning an explanation to be tailored to a particular context's and user's requirements, e.g., IDXs may be *dialectical* or *counterfactual*. We demonstrate IDXs' capability to explain various forms of BCs, e.g., naive or multi-label, binary or categorical, and also integrate recent approaches to explanations for BCs from the literature. We evaluate IDXs with theoretical and empirical analyses, demonstrating their considerable advantages when compared with existing explanation methods.

1 Introduction

The need for explainability has been one of the fastest growing concerns in AI of late, driven by academia, industry and governments, as various stakeholders become more conscious of the dangers posed by the widespread implementation of systems we do not fully understand. In response to this pressure, a plethora of explanations for AI methods have been proposed, with diverse strengths and weaknesses.

Some approaches to Explainable AI are heuristic and model-agnostic, e.g., [Ribeiro *et al.*, 2016; Lundberg and Lee, 2017]. Although model-agnosticism leads to broad applicability and uniformity across contexts, the heuristic nature of these approaches leads to some weaknesses, especially concerning user trust [Ignatiev, 2020]. Another limitation, brought about in part by the model-agnosticism, is the fact that these explanations are restricted to representations of how inputs influence outputs, neglecting the influences of intermediate components of models. This can lead to the undesirable situation where explanations of the outputs from two very diverse systems, e.g., a Bayesian network classifier (BC) and a neural model, performing the same task may be identical, despite completely different underpinnings. This trend towards explanations focusing on inputs and outputs exclusively is not limited to model-agnostic explanations, however. Explanations tailored towards specific AI methods are often also restricted to inputs' influence on outputs, e.g., the methods of [Shih *et al.*, 2018] for BCs or of [Bach *et al.*, 2015] for neural networks. Various methods have been devised for interpreting the intermediate components of neural networks (e.g., see [Bau *et al.*, 2017]), and [Olah *et al.*, 2018] have shown the benefits of identifying relations between these components and inputs/outputs. Some methods exist for accommodating intermediate components in counterfactual explanations (CFXs) of BCs [Albini *et al.*, 2020].

We propose the novel formalism of *influence-driven explanations* (IDXs) for systematically providing various forms of explanations for a variety of BCs, and admitting CFXs as instances, amongst others. The influences provide insight into the causal relationships between variables *within the BC*, which are then categorised by logical requirements, called, *relation properties*, according to their behaviour. The relation properties provide formal guarantees on the inclusion of the required types of influences included in IDXs. IDXs are thus fully customisable to the explanatory requirements of a particular application. Our contribution is threefold: we give (1) a systematic approach for generating IDXs from BCs, generalising existing work and offering great flexibility with regards to the BC model being explained and the nature of the explanation; (2) various instantiations of IDXs, including two based on the cardinal principle of dialectical monotonicity and one capturing (and extending) CFXs; and (3) theoretical and empirical analyses, showing the strengths of IDXs with respect to existing methods, along with illustrations of real world cases where the exploitation of these benefits may be particularly advantageous.

2 Related Work

There are a multitude of methods in the literature for providing explanations (e.g., see the recent survey undertaken by [Guidotti *et al.*, 2019]). Many are model-agnostic, including: *attribution* methods such as *LIME* [Ribeiro *et al.*, 2016] and *SHAP* [Lundberg and Lee, 2017], which assign each feature an *attribution value* indicating its contribution towards a prediction; *CXPlain* [Schwab and Karlen, 2019], using a causal model for generating attributions; and methods giving *counterfactual explanations*, such as *CERTIFAI* [Sharma *et al.*, 2020], using a custom genetic algorithm, *CLEAR* [White and d'Avila Garcez, 2020], using a local regression model, and *FACE* [Poyiadzi *et al.*, 2020], giving actionable explanations. Some of the model-agnostic methods rely upon symbolic representations, either to define explanations (e.g., *anchors* [Ribeiro *et al.*, 2018] determine sufficient conditions (inputs) for predictions (outputs)), or for logic-based counterparts of the underlying models from which explanations are drawn (e.g., [Ignatiev *et al.*, 2019a; Ignatiev *et al.*, 2019b]). Due to their model-agnosticism, all these methods restrict explanations to "correlations" between *inputs* and *outputs*. Instead, our focus on a specific method (BCs) allows us to define explanations providing a deeper representation of how the model is functioning via relations between input, output and *intermediate* variables.

Regarding BCs, [Shih *et al.*, 2018] define *minimum cardinality* and *prime implicant* explanations (extended to any model in [Darwiche and Hirth, 2020]) to ascertain pertinent features based on a complete set of classifications, i.e., a decision function representing the BC [Shih *et al.*, 2019]. These explanations are formally defined for binary variables only and again only explain outputs in terms of inputs. CFXs [Albini *et al.*, 2020] may include also intermediate variables and their relations: we will show that they can be captured as instantiations of our method. Explanation trees for causal Bayesian networks [Nielsen *et al.*, 2008] represent causal relations between variables, and links between causality and explanation have also been studied by [Halpern and Pearl, 2001a; Halpern and Pearl, 2001b]. While our influences are causal wrt the BC (as opposed to some model of the real world), we do not restrict the extracted relations in IDXs exclusively to those of a causal nature. Finally, [Timmer *et al.*, 2015] use support graphs (argumentation frameworks with support relations) as explanations showing the interplay between variables (as we do) in Bayesian networks, but (differently from us) commit to specific types of relations.

We will explore instances of our method giving *dialectical* forms of explanations for BCs, i.e., those which represent relationships between variables as being positive or negative. Various types of argumentation-based, dialectical explanations have been defined in the literature, but in different contexts and of different forms than ours, e.g., [García *et al.*, 2013] propose explanations as dialectical trees of arguments; [Fan and Toni, 2015] explain the *admissibility* [Dung, 1995] of arguments using dispute trees; several, e.g., [Teze *et al.*, 2018; Naveed *et al.*, 2018; Rago *et al.*, 2018], draw dialectical explanations to explain the outputs of recommender systems, while others focus on argumentation-based explanations of review aggregation [Cocarascu *et al.*, 2019], decision-making [Zeng *et al.*, 2018] and scheduling [Cyras *et al.*, 2019].

3 Bayesian Network Classifiers and Influences

We first define (discrete) BCs and their *decision functions*:

Definition 1. *A* BC *is a tuple* $\langle \mathcal{O}, \mathcal{C}, \mathcal{V}, \mathcal{D}, \mathcal{A} \rangle$ *such that:*

- $\mathcal{O}$ *is a (finite) set of* observations*;*
- $\mathcal{C}$ *is a (finite) set of* classifications*; we refer to* $\mathcal{X} = \mathcal{O} \cup \mathcal{C}$ *as the set of* variables*;*
- $\mathcal{V}$ *is a set of sets such that for any* $x \in \mathcal{X}$ *there is a unique* $V \in \mathcal{V}$ *associated to* x*, called* values *of* x *(*$\mathcal{V}(x)$ *for short);*
- $\mathcal{D} \subseteq \mathcal{X} \times \mathcal{X}$ *is a set of* conditional dependencies *such that* $\langle \mathcal{X}, \mathcal{D} \rangle$ *is an acyclic directed graph (we refer to this as the underlying* Bayesian network*); for any* $x \in \mathcal{X}$*,* $\mathcal{D}(x) = \{y \in \mathcal{X} | (y, x) \in \mathcal{D}\}$ *are the* parents *of* x*;*
- *For each* $x \in \mathcal{X}$*, each* $x_i \in \mathcal{V}(x)$ *is equipped with a* prior probability $P(x_i) \in [0, 1]$ *where* $\sum_{x_i \in \mathcal{V}(x)} P(x_i) = 1$*;*
- *For each* $x \in \mathcal{X}$*, each* $x_i \in \mathcal{V}(x)$ *is equipped with a set of* conditional probabilities *where if* $\mathcal{D}(x) = \{y, \ldots, z\}$*, for every* $y_m, \ldots, z_n \in \mathcal{V}(y) \times \ldots \times \mathcal{V}(z)$*, we have* $P(x_i | y_m, \ldots, z_n)$*, again with* $\sum_{x_i \in \mathcal{V}(x)} P(x_i | y_m, \ldots, z_n) = 1$*;*
- $\mathcal{A}$ *is the set of all possible* input assignments*: any* $a \in \mathcal{A}$ *is a (possibly partial) mapping* $a : \mathcal{X} \mapsto \bigcup_{x \in \mathcal{X}} \mathcal{V}(x)$ *such that, for every* $x \in \mathcal{O}$*, a assigns a value* $a(x) \in \mathcal{V}(x)$ *to* x*, and for every* $x \in \mathcal{X}$*, for every* $x_i \in \mathcal{V}(x)$*,* $P(x_i | a)$ *is the* posterior probability *of the value of* x *being* x_i*, given* a*.*[1]

Then, the decision function *(of the BC) is* $\sigma : \mathcal{A} \times \mathcal{X} \mapsto \bigcup_{x \in \mathcal{X}} \mathcal{V}(x)$ *where, for any* $a \in \mathcal{A}$ *and any* $x \in \mathcal{X}$*,* $\sigma(a, x) = argmax_{x_i \in \mathcal{V}(x)} P(x_i | a)$*.*

Thus, a BC consists of *variables*, which may be *classifications* or *observations*, *conditional dependencies* between them, *values* that can be ascribed to variables, and associated probability distributions resulting in a *decision function*, i.e., a mapping from inputs (assignments of values to variables) to outputs (assignments of values to classifications). Note that, differently from [Shih *et al.*, 2018; Albini *et al.*, 2020], BCs are not equated to decision functions, as probabilistic information is explicit in Definition 1.

We will consider various concrete BCs throughout the paper, all special cases of Definition 1 satisfying, in addition, an *independence property* among the parents of each variable. For all these BCs, and in the remainder of the paper, the *conditional probabilities* can be defined, for each $x \in \mathcal{X}, x_i \in \mathcal{V}(x)$, $y \in \mathcal{D}(x), y_m \in \mathcal{V}(y)$, as $P(x_i | y_m)$ with $\sum_{x_i \in \mathcal{V}(x)} P(x_i | y_m) = 1$. Specifically, for single-label classification we use Naive Bayes Classifiers (NBCs), with $\mathcal{C} = \{c\}$ and $\mathcal{D} = \{(c, x) | x \in \mathcal{O}\}$. For multi-label classification we use a variant of Bayesian network-based Chain Classifiers (BCCs) [Enrique Sucar *et al.*, 2014] in which leaves of the network are considered observations, the remaining variables classifications, and every

[1] Posterior probabilities may be estimated from the prior and conditional probabilities. Note that, if a is defined for x and $a(x) = x_i$, then $P(x_i | a) = 1$ and, for all $x_j \in \mathcal{V}(x) \setminus \{x_i\}$, $P(x_j | a) = 0$.

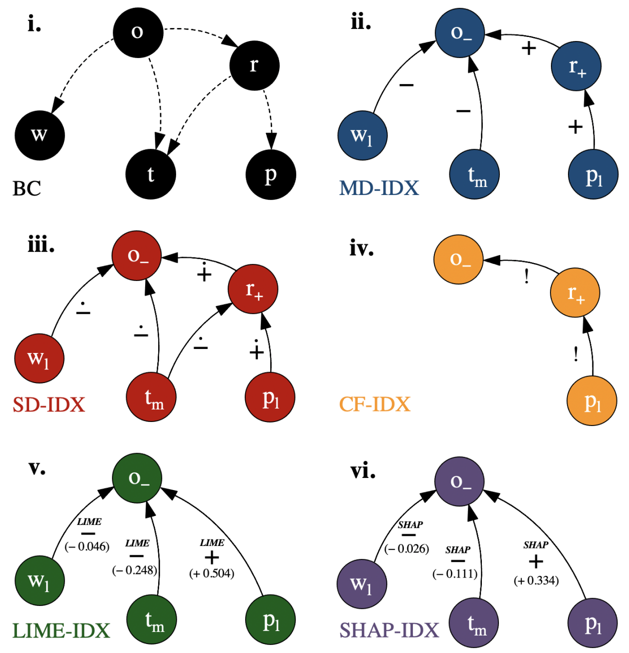

Figure 1: (i) Bayesian network for the Play-outside BC, with conditional dependencies as dashed arrows (see Table 1 for the BC's probabilities and Table 2 for its decision function σ), and (ii-vi) corresponding explanations (shown as graphs, with relations given by edges labelled with their type) for input *low wind* (w_l), *medium temperature* (t_m), and *low pressure* (p_l), with relevant variables, extracted relations and LIME/SHAP attribution values indicated. All explanations are instances of IDXs, illustrating our method's flexibility and potential to accommodate different explanatory requirements.

classification $c \in \mathcal{C}$ is estimated with a NBC in which the children of c are the inputs.

In the remainder of the paper, unless specified otherwise, we assume as given a generic BC $\langle \mathcal{O}, \mathcal{C}, \mathcal{V}, \mathcal{D}, \mathcal{A} \rangle$ satisfying the aforementioned independence property.

For illustration, consider the *play-outside* BCC in Figure 1i (for now ignoring the other subfigures, which will be introduced later), in which classifications *play outside* and *raining* are determined from observations *wind*, *temperature* and *pressure*. Here, $\mathcal{C} = \{o, r\}$, $\mathcal{O} = \{w, t, p\}$ and $\mathcal{D}$ is as in the figure. Then, let $\mathcal{V}$ be such that $\mathcal{V}(w) = \mathcal{V}(t) = \{low, medium, high\}$, $\mathcal{V}(p) = \{low, high\}$ and $\mathcal{V}(r) = \mathcal{V}(o) = \{-, +\}$, i.e., w and t are categorical while p, r and o are binary. Table 1 shows the prior and conditional probabilities for this BCC leading, in turn, to posterior probabilities in Table 2. For example, for input *low wind*, *medium temperature* and *low pressure*, the BCC's posterior probabilites for *raining* and *play outside* may be calculated as (with values of variables as subscripts)[2] $P(r_+|t_m, p_l) = 0.94 \propto P(r_+) \cdot P(t_m|r_+) \cdot P(p_l|r_+)$ and $P(o_-|w_l, t_m, p_l) = 0.99 \propto P(o_-) \cdot P(w_l|o_-) \cdot P(t_m|o_-) \cdot P(r_+|o_-)$.

	w_l	w_m	w_h	t_l	t_m	t_h	p_l	p_h	r_+	r_-	o_+	o_-
$P(\cdot)$	.33	.33	.33	.33	.33	.33	.50	.50	.67	.33	.22	.78
$P(\cdot\|r_+)$	×	×	×	.25	.25	.50	.75	.25	×	×	×	×
$P(\cdot\|r_-)$	×	×	×	.49	.49	.02	.02	.98	×	×	×	×
$P(\cdot\|o_+)$	.48	.26	.26	.26	.72	.02	×	×	.02	.98	×	×
$P(\cdot\|o_-)$	.28	.36	.36	.36	.22	.42	×	×	.85	.15	×	×

Table 1: Prior and conditional probabilities for the play-outside BCC estimated with Laplace smoothing ($\alpha = 0.1$) from the dataset/decision function σ in Table 2.

σ					SD-/**MD**-IDX		CF-IDX	
w	t	p	r	o	$-(r)/+(r)$	$-(o)/+(o)$	$!(r)/*(r)$	$!(o)/*(o)$
l	l	l	+.94	−.99	$\{t\}/\{\boldsymbol{p}\}$	$\{\boldsymbol{w}\}/\{t, \boldsymbol{r}\}$	$\{p\}/\{\}$	$\{r\}/\{t\}$
m	l	l	+.94	−.99	$\{t\}/\{\boldsymbol{p}\}$	$\{\}/\{w, t, \boldsymbol{r}\}$	$\{p\}/\{\}$	$\{\}/\{w, t, r\}$
l	m	l	+.94	−.99	$\{t\}/\{\boldsymbol{p}\}$	$\{\boldsymbol{w}, \boldsymbol{t}\}/\{\boldsymbol{r}\}$	$\{p\}/\{\}$	$\{r\}/\{\}$
m	m	l	+.79	−.51	$\{t\}/\{\boldsymbol{p}\}$	$\{\boldsymbol{t}\}/\{w, \boldsymbol{r}\}$	$\{p\}/\{\}$	$\{r\}/\{w\}$
h	l	l	+.79	−.69	$\{t\}/\{\boldsymbol{p}\}$	$\{\}/\{w, t, \boldsymbol{r}\}$	$\{p\}/\{\}$	$\{\}/\{w, t, r\}$
l	h	l	+.79	−.51	$\{\}/\{\boldsymbol{t}, \boldsymbol{p}\}$	$\{\boldsymbol{w}\}/\{\boldsymbol{t}, \boldsymbol{r}\}$	$\{\}/\{t, p\}$	$\{\}/\{r\}$
h	m	l	+.79	−.81	$\{t\}/\{\boldsymbol{p}\}$	$\{\boldsymbol{t}\}/\{w, \boldsymbol{r}\}$	$\{p\}/\{\}$	$\{r\}/\{w\}$
m	h	l	+.79	−.91	$\{\}/\{\boldsymbol{t}, \boldsymbol{p}\}$	$\{\}/\{w, \boldsymbol{t}, \boldsymbol{r}\}$	$\{\}/\{t, p\}$	$\{\}/\{w, r\}$
h	h	l	+.79	−.81	$\{\}/\{\boldsymbol{t}, \boldsymbol{p}\}$	$\{\}/\{w, \boldsymbol{t}, \boldsymbol{r}\}$	$\{\}/\{t, p\}$	$\{\}/\{w, r\}$
l	l	h	−.99	+.99	$\{\}/\{t, \boldsymbol{p}\}$	$\{\}/\{\boldsymbol{w}, t, \boldsymbol{r}\}$	$\{p\}/\{t\}$	$\{w, r\}/\{\}$
m	l	h	−.99	−.99	$\{\}/\{t, \boldsymbol{p}\}$	$\{t, \boldsymbol{r}\}/\{w\}$	$\{p\}/\{t\}$	$\{t\}/\{w\}$
l	m	h	−.99	+.99	$\{\}/\{t, \boldsymbol{p}\}$	$\{\}/\{\boldsymbol{w}, \boldsymbol{t}, \boldsymbol{r}\}$	$\{p\}/\{t\}$	$\{r\}/\{w, t\}$
m	m	h	−.97	+.99	$\{\}/\{t, \boldsymbol{p}\}$	$\{w\}/\{\boldsymbol{t}, \boldsymbol{r}\}$	$\{p\}/\{t\}$	$\{t, r\}/\{\}$
h	l	h	−.97	−.99	$\{\}/\{t, \boldsymbol{p}\}$	$\{t, \boldsymbol{r}\}/\{w\}$	$\{p\}/\{t\}$	$\{t\}/\{w\}$
l	h	h	+.97	−.99	$\{\boldsymbol{p}\}/\{\boldsymbol{t}\}$	$\{\boldsymbol{w}\}/\{\boldsymbol{t}, \boldsymbol{r}\}$	$\{t\}/\{\}$	$\{\}/\{r\}$
h	m	h	−.97	+.98	$\{\}/\{t, \boldsymbol{p}\}$	$\{w\}/\{\boldsymbol{t}, \boldsymbol{r}\}$	$\{p\}/\{t\}$	$\{t, r\}/\{\}$
m	h	h	+.97	−.95	$\{\boldsymbol{p}\}/\{\boldsymbol{t}\}$	$\{\}/\{w, \boldsymbol{t}, \boldsymbol{r}\}$	$\{t\}/\{\}$	$\{\}/\{w, r\}$
h	h	h	+.97	−.98	$\{\boldsymbol{p}\}/\{\boldsymbol{t}\}$	$\{\}/\{w, \boldsymbol{t}, \boldsymbol{r}\}$	$\{t\}/\{\}$	$\{\}/\{w, r\}$

Table 2: Decision function σ (with the probability of the most probable class) and explanations (SD-IDX and CF-IDX, that will be introduced in Section 4) for the play-outside BC in Figure 1i, where, for any variable $x \in \mathcal{X}$ and relation type t, $t(x) = \{y \in \mathcal{X} | (y, x) \in \mathcal{R}_t\}$. We indicate in bold variables in the corresponding monotonic attackers and supporters sets in the MD-IDX for this example (also introduced in Section 4).

This gives the decision function σ in Table 2, where the earlier example input results in r_+ and o_-.

Our method for generating explanations relies on modelling how the variables within a BC *influence* one another. For this we adapt the following from [Albini *et al.*, 2020].

Definition 2. *The set of* influences *is the (acyclic) relation* $\mathcal{I} = \{(x, c) \in \mathcal{X} \times \mathcal{C} | (c, x) \in \mathcal{D}\}$.

Influences thus indicate the direction of the inferences in determining the values of classifications, neglecting, for example, dependencies between observations as considered in *tree-augmented naive BCs* [Friedman *et al.*, 1997]. Note that other forms of influence, which we leave to future work, may be required for more complex BCs where inferences are made in other ways, e.g., for *Markov Blanket-based BCs* [Koller and Sahami, 1996] where inferences may follow dependencies or be drawn between variables not linked by them. Notice also that the influences correspond to the structure of the model as it has been built or learned from data, any imprecision in the construction of the model or bias in the data with respect to the real world will be reflected in the influences too.

[2] We indicate with $\propto$ the normalized posterior probability (i.e., such that $\sum_{c_i \in \mathcal{V}(c)} P(c_i|\cdot) = 1$). $P(r_+|t_m, p_l)$ is the posterior probability of an NBC predicting r from observations t, p as input whilst $P(o_-|w_l, t_m, p_l)$ is that of another NBC predicting o from, as inputs, observations w, t and classification r (as predicted by the first NBC).

While the explanations we will define may amount to non-shallow graphs[3] of relations between variables, in order to integrate other, shallow explanation methods, i.e., connecting only inputs and outputs, we define a restricted form of influences.

Definition 3. *The set of* input-output influences*, where $\mathcal{C}_o \subseteq \mathcal{C}$ are* outputs*, is the (acyclic) relation $\mathcal{I}_{io} = \mathcal{O} \times \mathcal{C}_o$.*

For illustration, in the running example in Figure 1i, $\mathcal{I} = \{(w,o),(t,o),(r,o),(t,r),(p,r)\}$ and $\mathcal{I}_{io} = \{(w,o),(t,o),(p,o)\}$ for $\mathcal{C}_o = \{o\}$, while $\mathcal{I}_{io} = \{(w,r),(t,r),(p,r),(w,o),(t,o),(p,o)\}$ for $\mathcal{C}_o = \{o,r\}$. Note that in the former $\mathcal{I}_{io}$ case, r is neglected, while in the latter, the influence (w,r) is extracted despite the fact that *wind* cannot influence *raining* in this BC, highlighting that using $\mathcal{I}_{io}$, instead of the full $\mathcal{I}$, may have drawbacks for non-naive BCs, except when the notions coincide, as characterised next:[4]

Proposition 1. *Given outputs $\mathcal{C}_o \subseteq \mathcal{C}$, $\mathcal{I} = \mathcal{I}_{io}$ iff $\mathcal{D} = \mathcal{C}_o \times \mathcal{O}$.*

4 Influence-Driven Explanations

We introduce a general method for generating explanations of BCs, and then several instantiations thereof. Our explanations are constructed by categorising influences as specific types of relations depending on the satisfaction of relation properties characterising those relations, defined as follows.

Definition 4. *Given influences $\mathcal{I}$, an* explanation kit *is a finite set of pairs $\{\langle t_1, \pi_1\rangle, \ldots \langle t_n, \pi_n\rangle\}$ where $\{t_1, \ldots, t_n\}$ is a set of* relation types *and $\{\pi_1, \ldots, \pi_n\}$ is a set of* relation properties *with $\pi_i : \mathcal{I} \times \mathcal{A} \rightarrow \{true, false\}$, for $i = 1, \ldots, n$.*

Intuitively, relation property π_i is satisfied for $(x,y) \in \mathcal{I}$ and $a \in \mathcal{A}$ iff $\pi((x,y),a) = true$.

Definition 5. *Given influences $\mathcal{I}$ and an explanation kit $\{\langle t_1, \pi_1\rangle, \ldots \langle t_n, \pi_n\rangle\}$, an* influence-driven explanation *(IDX) for* explanandum *$e \in \mathcal{C}$ with input assignment $a \in \mathcal{A}$ is a tuple $\langle \mathcal{X}_r, \mathcal{R}_{t_1}, \ldots, \mathcal{R}_{t_n}\rangle$ with:*
- *$\mathcal{X}_r \subseteq \mathcal{X}$ such that $e \in \mathcal{X}_r$;*
- *$\mathcal{R}_{t_1}, \ldots \mathcal{R}_{t_n} \subseteq \mathcal{I} \cap (\mathcal{X}_r \times \mathcal{X}_r)$ such that for any $i = 1 \ldots n$, for every $(x,y) \in \mathcal{R}_{t_i}$, $\pi_i((x,y),a) = true$;*
- *$\forall x \in \mathcal{X}_r$ there is a sequence $x_1, \ldots, x_k$, $k \geq 1$, such that $x_1 = x$, $x_k = e$, and $\forall 1 \leq i < k$ $(x_i, x_{i+1}) \in \mathcal{R}_{t_1} \cup \ldots \cup \mathcal{R}_{t_n}$.*

An IDX is thus guaranteed to be a set of *relevant* variables ($\mathcal{X}_r$), including the explanandum, connected to one another in a graph by the relations from the explanation kit. We follow the findings of [Nielsen *et al.*, 2008] that only some of the variables in $\mathcal{X}$ (potentially very few [Miller, 2019]) may suffice to explain the explanandum's value.

The computational cost of producing IDXs essentially depends on the computational cost of computing the relation properties (for each influence), namely, once an input assignment a is given, $\pi_1((x,y),a), \ldots, \pi_n((x,y))$ for every influence $(x,y) \in \mathcal{I}$. In fact, we can see from Algorithm 1 that the rest of the explanation process mainly amounts to a deep-first search in the graph of the influences. We will discuss further the computational costs of our explanations in Section 5.2.

Algorithm 1 Influence-Driven Explanations

```
function PARENTS($e, \mathcal{I}$)
    $parents \leftarrow \varnothing$
    for $(x,c) \in \mathcal{I}$ do
        if $c == e$ then
            $parents \leftarrow parents \cup \{x\}$
        end if
    end for
    return $parents$
end function

function XP($\mathcal{O}, \mathcal{C}, \mathcal{I}, e, a, \{\langle t_1, \pi_1\rangle, \ldots, \langle t_n, \pi_n\rangle\}$)    ▷ Generate the explanation for the explanandum $e \in \mathcal{C}$ given an input assignment $a \in \mathcal{A}$ using the explanation kit $\{\langle t_1, \pi_1\rangle, \ldots, \langle t_n, \pi_n\rangle\}$
    $\mathcal{X}_r \leftarrow \varnothing$
    $\mathcal{R} \leftarrow \{\}$
    for $t_i \in \{t_1, \ldots, t_n\}$ do
        $\mathcal{R}[t_i] = \varnothing$
    end for
    for $x \in$ PARENTS($e, \mathcal{I}$) do
        for $\pi_i \in \{\pi_1, \ldots, \pi_n\}$ do
            if $\pi_i((x,e),a)$ then
                $\mathcal{X}_r \leftarrow \mathcal{X}_r \cup \{x\}$
                $\mathcal{R}[t_i] \leftarrow \mathcal{R}[t_i] \cup \{(x,e)\}$
            end if
        end for
        $\widehat{\mathcal{X}_r}, \widehat{\mathcal{R}} \leftarrow$ XP($\mathcal{O}, \mathcal{C}, \mathcal{I}, x, a, \{\langle t_1, \pi_1\rangle, \ldots, \langle t_n, \pi_n\rangle\}$)
        $\mathcal{X}_r \leftarrow \mathcal{X}_r \cup \widehat{\mathcal{X}_r}$
        for $t_i \in \{t_1, \ldots, t_n\}$ do
            $\mathcal{R}[t_i] \leftarrow \mathcal{R}[t_i] \cup \widehat{\mathcal{R}[t_i]}$
        end for
    end for
    return $\mathcal{X}_r, \mathcal{R}$
end function
```

We will now demonstrate the flexibility of our approach by instantiating various IDXs, which are illustrated in Table 2 and in Figures 1ii-vi for the running example. In doing so, we will make use of the following notion.

Definition 6. *Given a BC $\langle \mathcal{O}, \mathcal{C}, \mathcal{V}, \mathcal{D}, \mathcal{A}\rangle$ with influences $\mathcal{I}$, a variable $x \in \mathcal{X}$ and an input $a \in \mathcal{A}$, the* modified input *$a'_{x_k} \in \mathcal{A}$ by $x_k \in \mathcal{V}(x)$ is such that, for any $z \in \mathcal{X}$: $a'_{x_k}(z) = x_k$ if $z = x$, and $a'_{x_k}(z) = a(z)$ otherwise.*

A modified input thus assigns a desired value (x_k) to a specified variable (x), keeping the preexisting input assignments unchanged. For example, if $a \in \mathcal{A}$ amounts to *low wind*, *medium temperature* and *low pressure* in the running example (i.e., $a(w) = l$, $a(t) = m$, $a(p) = l$), then $a'_{w_h} \in \mathcal{A}$ refers to *high wind*, *medium temperature* and *low pressure*.

4.1 Monotonically Dialectical IDXs

Motivated by recent uses of argumentation for explanation (see Section 2), we draw inspiration from *bipolar argumentation frameworks* [Cayrol and Lagasquie-Schiex, 2005] to define a *dialectical* instance of the notion of explanation kit, with attack and support relations defined by imposing properties

[3] A non-shallow graph is a graph with one or more paths (between any two vertices) of length > 1.

[4] The proofs of all results are in Appendix A.

of *dialectical monotonicity*, requiring that attackers (supporters) have a negative (positive, resp.) effect on the variables they influence. Concretely, we require that an influencer is an attacker (a supporter) if its assigned value minimises (maximises, resp.) the posterior probability of the influencee's current value (with all other influencers' values unchanged):

Definition 7. *A* monotonically dialectical explanation kit *is a pair* $\{\langle -, \pi_- \rangle, \langle +, \pi_+ \rangle\}$ *with relation types of* monotonic attack $-$ *and* monotonic support $+$ *characterised by the following relation properties* π_-, π_+. *For any* $(x, y) \in \mathcal{I}$, $a \in \mathcal{A}$:

- $\pi_-((x, y), a) = true$ *iff* $\forall x_k \in \mathcal{V}(x) \setminus \{\sigma(a, x)\}$:
  $P(\sigma(a, y)|a) < P(\sigma(a, y)|a'_{x_k})$;
- $\pi_+((x, y), a) = true$ *iff* $\forall x_k \in \mathcal{V}(x) \setminus \{\sigma(a, x)\}$:
  $P(\sigma(a, y)|a) > P(\sigma(a, y)|a'_{x_k})$.

Thus, a *monotonically dialectical IDX (MD-IDX)* is a IDX drawn from a monotonically dialectical kit.

For illustration, consider the MD-IDX in Figure 1ii: here p_l monotonically supports (i.e., increases the probability of) r_+, which in turn monotonically supports o_-; while, w_l and t_m reduce the chance of o_-, leading to a monotonic attack.

It should be noted that since the dialectical monotonicity requirement here is rather strong, for some BCs (in particular those with variables with large domains) the MD-IDX could be empty, i.e., comprises only the explanandum. We examine the prevalence of these relations for a range of datasets in Section 5. This form of explanation is appropriate in contexts where the users prefer sharp explanations with monotonic properties, when available, and may accept the absence of explanations, when these properties are not satisfied.

4.2 Stochastically Dialectical IDXs

We believe that monotonicity is a property humans naturally expect from explanations . Nonetheless, monotonicity is a strong requirement that may lead, for some BCs and contexts, to very few influences playing a role in MD-IDXs. We now introduce a weaker form of dialectical explanation, where an influencer is an attacker (supporter) if the posterior probability of the influencee's current value is lower (higher, resp.) than the average of those resulting from the influencer's other values, weighted by their prior probabilities (while all other influencers' values remain unchanged):

Definition 8. *A* stochastically dialectical explanation kit *is a pair* $\{\langle \dot{-}, \pi_{\dot{-}} \rangle, \langle \dot{+}, \pi_{\dot{+}} \rangle\}$ *with relation types of* stochastic attack $\dot{-}$ *and* stochastic support $\dot{+}$ *characterised by the following relation properties* $\pi_{\dot{-}}$, $\pi_{\dot{+}}$. *For any* $(x, y) \in \mathcal{I}$, $a \in \mathcal{A}$:

- $\pi_{\dot{-}}((x, y), a) = true$ *iff*

$$P(\sigma(a, y)|a) < \frac{\sum_{x_k \in \mathcal{V}(x) \setminus \{\sigma(a,x)\}} \left[P(x_k) \cdot P(\sigma(a,y)|a'_{x_k})\right]}{\sum_{x_k \in \mathcal{V}(x) \setminus \{\sigma(a,x)\}} P(x_k)};$$

- $\pi_{\dot{+}}((x, y), a) = true$ *iff*

$$P(\sigma(a, y)|a) > \frac{\sum_{x_k \in \mathcal{V}(x) \setminus \{\sigma(a,x)\}} \left[P(x_k) \cdot P(\sigma(a,y)|a'_{x_k})\right]}{\sum_{x_k \in \mathcal{V}(x) \setminus \{\sigma(a,x)\}} P(x_k)}.$$

Then, a *stochastically dialectical IDX (SD-IDX)* is a IDX drawn from a stochastically dialectical kit. SD-IDXs are *stochastic* in that they weaken the monotonicity constraint (compared to MD-IDXs) by taking into account the prior probabilities of the possible changes of the influencers.

For illustration, the SD-IDX in Figure 1iii extends the MD-IDX in Figure 1ii by including the negative (stochastic) effect which t_m has on r_+.

The weakening of the monotonicity requirement for SD-IDXs means that SD-IDXs will not be empty except in some special, improbable cases, e.g., when every influencing variable of the explanandum has all values with equal posterior probability. We therefore expect this form of explanation to be appropriate in contexts where users are looking for explanations featuring some degree of dialectically monotonic behaviour in the influences, but do not require strong properties and then would prefer to receive a more populated IDX than MD-IDXs.

4.3 Counterfactual IDXs

We can also naturally instantiate explanation kits to define counterfactual explanations capturing the CFXs of [Albini *et al.*, 2020]. For this we use two relations: one indicating influencers whose assigned value is critical to the influencee's current value, and another indicating influencers whose assigned value can potentially contribute to the change (together with the changes of other influencers).

Definition 9. *A* counterfactual explanation kit *is a pair* $\{\langle !, \pi_! \rangle, \langle *, \pi_* \rangle\}$ *with relation types of* critical influence ! *and* potential influence $*$ *characterised by the following relation properties* $\pi_!$, π_*. *For any* $(x, y) \in \mathcal{I}$, $a \in \mathcal{A}$:

- $\pi_!((x,y), a) = true$ *iff* $\exists a' \in \mathcal{A}$ *such that* $\sigma(a',x) \neq \sigma(a, x)$ *and* $\forall z \in \mathcal{I}(y) \backslash \{x\}$ $\sigma(a',z) = \sigma(a,z)$, *and* $\forall a' \in \mathcal{A}$, $\sigma(a,y) \neq \sigma(a',y)$.
- $\pi_*((x,y), a) = true$ *iff* $\pi_!((x,y), a) = false$ *and* $\exists a', a'' \in \mathcal{A}$ *such that* $\sigma(a, x) = \sigma(a', x) \neq \sigma(a'', x)$, $\forall z \in \mathcal{I}(y) \backslash \{x\}$ $\sigma(a', z) = \sigma(a'', z)$, *and* $\sigma(a, y) = \sigma(a', y) \neq \sigma(a'', y)$.

Thus a *counterfactual IDX (CF-IDX)* is an IDX drawn from a counterfactual kit. For illustration, Figure 1iv shows the CF-IDX for the running example, with no potential influences and with critical influences indicating that if p_l were to change (i.e., to p_h), this would force r_- and, in turn, o_+.

CF-IDXs are not built around the principle of dialectical monotonicity, and instead reflect the effects that changes to the values of variables would have on the variables that they influence. We suggest that this form of explanation is appropriate when the users wish to highlight the factors which led to a prediction and which can be changed in order to reverse it.

4.4 Attribution Method Based Dialectical IDXs

We now further show the versatility of the notion of explanation kit by instantiating it to integrate attribution methods, e.g., LIME (see Section 2), which are widely used in academia and industry. To reflect attribution methods' focus on input-output variables, these instances are defined in terms of input-output influences $\mathcal{I}_{io}$ and outputs $\mathcal{C}_o$ as follows:

Definition 10. *A* LIME explanation kit *is a pair* $\{\langle \overset{\scriptscriptstyle LIME}{-}, \pi_{\overset{\scriptscriptstyle LIME}{-}} \rangle, \langle \overset{\scriptscriptstyle LIME}{+}, \pi_{\overset{\scriptscriptstyle LIME}{+}} \rangle\}$ *with relation types* LIME-attack $\overset{\scriptscriptstyle LIME}{-}$ *and* LIME-support $\overset{\scriptscriptstyle LIME}{+}$ *characterised by the following relation properties* $\pi_{\overset{\scriptscriptstyle LIME}{-}}$, $\pi_{\overset{\scriptscriptstyle LIME}{+}}$. *For any* $(x, y) \in \mathcal{I}_{io}$, $a \in \mathcal{A}$:

- $\pi_{\overset{\scriptscriptstyle LIME}{+}}((x, y), a) = true$ *iff* $v_{LIME}(a, x, y) > 0$, *and*
- $\pi_{\overset{\scriptscriptstyle LIME}{-}}((x, y), a) = true$ *iff* $v_{LIME}(a, x, y) < 0$,

where $v_{LIME} : \mathcal{O} \times \mathcal{A} \times \mathcal{C}_o \mapsto \mathbb{R}$ *is such that* $v_{LIME}(a, x, y)$ *is the value that LIME assigns to the observation* x *with an input assignment* a *with respect to the output value* $\sigma(a, y)$*.*

Then, a *LIME-IDX* is an IDX drawn from a LIME explanation kit. Note that LIME-IDXs are still *dialectical.* We can similarly define (dialectical) explanation kits and IDXs for other attribution methods (see Section 2), e.g., a *SHAP explanation kit* and resulting *SHAP-IDX.* For illustration, Figures 1v and 1vi show the LIME-IDX and SHAP-IDX, resp., for the play-outside BCC, with input-output influences annotated with the respective attribution values. A notable difference is that LIME and SHAP also provide weightings on the relations (LIME contributions and Shapley values, resp.), while IDXs do not include any such scoring. Enhancing IDXs with quantitative information is an important direction of future work. As an example, in SD-IDXs importance scores could be defined based on the difference between $P(\sigma(a, y)|a)$ and its (weighted) expected probability with other inputs (the right-hand side of the inequalities in Definition 8).

The restriction to input-output influences implies that the intermediate variable *raining* is not considered by these explanations. This may not seem essential in a simple example such as this, but in real world applications such as medical diagnosis, where BCs are particularly prevalent, the inclusion of intermediate information could be beneficial (e.g., see Figure 2). We suggest that these forms of IDX are suitable when the users prefer explanations with a simpler structure and, in particular, are not concerned about the intermediate variables of the BC nor require the dialectical relations to satisfy dialectical monotonicity.

5 Evaluation

In this section we perform a thorough evaluation of IDXs via theoretical and empirical analyses.

5.1 Theoretical Analysis

Our first two propositions show the relationship and, in special cases, equivalence between MD-IDXs and SD-IDXs.

Proposition 2. *Given an MD-IDX* $\langle \mathcal{X}_r, \mathcal{R}_-, \mathcal{R}_+ \rangle$ *and an SD-IDX* $\langle \mathcal{X}'_r, \mathcal{R}_{\dot{-}}, \mathcal{R}_{\dot{+}} \rangle$ *for an explanandum* $e \in \mathcal{X}_r \cap \mathcal{X}'_r$ *and input assignment* $a \in \mathcal{A}$*, we have* $\mathcal{X}_r \subseteq \mathcal{X}'_r$, $\mathcal{R}_- \subseteq \mathcal{R}_{\dot{-}}$ *and* $\mathcal{R}_+ \subseteq \mathcal{R}_{\dot{+}}$*.*

Proposition 2 shows that an MD-IDX (for a certain explanandum and input assignment) is always (element-wise) a subset of the corresponding SD-IDX.

As described in Section 4.1, due to the stronger assumption required by *dialectical monotonicity* (when compared to its stochastic counterpart), there are two predominant factors that contribute to MD-IDX becoming an (increasingly) smaller, and potentially empty, subset of SD-IDX: (1) the cardinalities of variables' domains, i.e., the larger the domain the less likely *dialectical monotonicity* is to hold; (2) the depth of the explanation, i.e., the deeper an explanation (i.e, the longer the path from the inputs to the explanandum is), the less likely it is for a variable to have a path to the explanandum. When all variables are binary, MD-IDXs and SD-IDXs are equivalent as shown in the following.

Proposition 3. *Given an MD-IDX* $\langle \mathcal{X}_r, \mathcal{R}_-, \mathcal{R}_+ \rangle$ *and an SD-IDX* $\langle \mathcal{X}'_r, \mathcal{R}_{\dot{-}}, \mathcal{R}_{\dot{+}} \rangle$ *for explanandum* $e \in \mathcal{X}_r \cap \mathcal{X}'_r$ *and input assignment* $a \in \mathcal{A}$*, if, for any* $x \in \mathcal{X}'_r \setminus \{e\}$, $|\mathcal{V}(x)| = 2$*, then* $\mathcal{X}_r = \mathcal{X}'_r$, $\mathcal{R}_- = \mathcal{R}_{\dot{-}}$ *and* $\mathcal{R}_+ = \mathcal{R}_{\dot{+}}$*.*

We now characterise desirable behaviour of dialectical explanation kits, requiring that attackers and supporters have a monotonic effect on the posterior probability of the assigned values to variables that they influence. Concretely, we characterise a *dialectical* version of *monotonicity* which requires that if the influence from a variable is classified as being a support (an attack) then its assigned value maximises (minimises, resp.) the posterior probability of the influenced variable's current value (with all other influencing variables' values remaining unchanged). In other words, if we were to change the value of a variable supporting (attacking) another variable, then the posterior probability of the latter's current value would decrease (increase, resp.).

Property 1. *A dialectical explanation kit* $\{\langle a, \pi_a \rangle, \langle s, \pi_s \rangle\}$[5] *satisfies* dialectical monotonicity *iff for any dialectical IDX* $\langle \mathcal{X}_r, \mathcal{R}_a, \mathcal{R}_s \rangle$ *drawn from the kit (for any explanandum* $e \in \mathcal{X}_r$*, input assignment* $a \in \mathcal{A}$*), it holds that for any* $(x,y) \in \mathcal{R}_a \cup \mathcal{R}_s$*, if* $a' \in \mathcal{A}$ *is such that* $\sigma(a',x) \neq \sigma(a,x)$ *and* $\sigma(a',z) = \sigma(a,z)$ $\forall z \in \mathcal{I}(y) \setminus \{x\}$*, then:*

- *if* $(x, y) \in \mathcal{R}_a$ *then* $P(\sigma(a, y)|a') > P(\sigma(a, y)|a)$*;*
- *if* $(x, y) \in \mathcal{R}_s$ *then* $P(\sigma(a, y)|a') < P(\sigma(a, y)|a)$*.*

Proposition 4. *Monotonically dialectical explanation kits satisfy dialectical monotonicity, while stochastically dialectical, LIME and SHAP explanation kits do not.*

The final proposition gives the relationship between CF-IDXs and MD-/SD-IDXs.

Proposition 5. *Given an MD-IDX* $\langle \mathcal{X}_r, \mathcal{R}_-, \mathcal{R}_+ \rangle$*, an SD-IDX* $\langle \mathcal{X}_r, \mathcal{R}_{\dot{-}}, \mathcal{R}_{\dot{+}} \rangle$ *and a CF-IDX* $\langle \mathcal{X}'_r, \mathcal{R}_!, \mathcal{R}_* \rangle$ *for explanandum* $e \in \mathcal{X}_r$ *and input assignment* $a \in \mathcal{A}$*, it holds that* $\mathcal{R}_! \subseteq \mathcal{R}_+$ *and* $\mathcal{R}_! \subseteq \mathcal{R}_{\dot{+}}$*.*

Proposition 5 shows that the *critical influence* relation of CF-IDXs is a stronger variant of the relation of monotonic support. In fact it requires the current value of an influencee to change (rather then only the reduction of its probability), given a change in the value of the influencer.

5.2 Empirical Analysis

We used several datasets/Bayesian networks (see Table 3 for details), for each of which we deployed an NBC (for single-label classification dataset) or a BCC (for multi-label classification datasets and non-shallow Bayesian networks). When training BCs from datasets, we split them into train and test sets (with 75/25% ratio) and optimised the hyper-parameters using 5-fold cross-validation (see Appendix B for details).

[5] Here *a* and *s* are some form of attack and support, resp., depending on the specific explanation kit; e.g., for *stochastically* dialectical explanation kits $a = \dot{-}$ and $s = \dot{+}$.

Dataset	BC[†]	Size	Variables $\lvert\mathcal{O}\rvert$	$\lvert\mathcal{C}\rvert$	Types[‡] $\mathcal{O}$	$\mathcal{C}$	Performance[§] Accuracy	F1
Shuttle[6]	NBC	278	6	1	C	B	95.7%	0.96
Votes[6]	NBC	435	16	1	B	B	90.8%	0.90
Parole[7]	NBC	675	8	1	C	B	88.8%	0.69
German[6]	NBC	750	20	1	C	B	76.4%	0.72
COMPAS[8]	NBC	6951	12	1	C	B	70.5%	0.71
Car[6]	NBC	1728	6	1	C	C	86.6%	0.76
Emotions[9]	BCC	593	72	6	C	B	80.2%	0.70
Asia[10]	BCC	4	2	6	B	B	100%	1.00
Child[10]	BCC	1080	7	13	C	C	80.6%	0.66

Table 3: Characteristics of the datasets used in the evaluation. (†) NBC (Naive BC) or BCC (Bayesian Chain Classifier); (‡) **B**inary or **C**ategorical; (§) accuracy and macro F1 score on the test set, averaged for multi-label settings.

Computational cost. MD-IDXs and SD-IDXs are model-specific explanations and can thus access the internal state of the BC. This allows them to be more efficient in extracting information than model-agnostic methods, e.g., methods based on the sampling of the input space. Formally, let t_p be the time to compute a prediction and its associated posterior probabilities (in our experiments[11], t_p ranged from $3\mu s$ for the simplest NBC to $40ms$ for the more complex BCC). The time complexity to compute whether an influence (x, y) belongs to a relation in MD-IDXs or in SD-IDXs The time complexity to compute whether an influence (x, y) belongs to a relation in MD-IDXs or an SD-IDXs, denoted as T_{1-IDX}, is a function of the number of values that x can be assigned to (i.e, $|\mathcal{V}(x)|$) because in order to assess that, the algorithm has to check how the posterior probability of y changes when changing x (as per Defs. 7 and 8).

$$T_{1-IDX}(\mathcal{V}(x))=\Theta\left(t_p \cdot [1+|\mathcal{V}(x)| - 1]\right)=\Theta\left(t_p \cdot |\mathcal{V}(x)|\right).$$

The time complexity to compute a full MD/SD-IDX, denoted as T_{IDX}, is the sum of the above for all the influences, with some potential savings that can be achieved depending on the specific type of BC that is being used (e.g., in BCC we can decompose the prediction into multiple NBCs, therefore leading to a lower cost). Thus, formally, the time complexity to compute an MD-IDX or an SD-IDX is:

$$T_{IDX}(\mathcal{I}, \mathcal{V}) = \Theta\left(t_p \cdot \left[1 + \textstyle\sum_{x \in \{x|(x,y) \in \mathcal{I}\}} (|\mathcal{V}(x)| - 1)\right]\right).$$

The asymptotic complexity can be further simplified to:

$$T_{IDX}(\mathcal{V}) = \Theta\left(t_p \cdot \textstyle\sum_{x \in \mathcal{X}} |\mathcal{V}(x)|\right)$$

which is *linear* with respect to the sum of the number of variables' values. This makes our dialectical explanations *efficient*. As a comparison, the time taken to generate an MD-IDX for the *Child BC* is less than $60 \cdot t_p$ while the time taken to generate a LIME explanation with default parameters is $5000 \cdot t_p$ because LIME asks the BC for the prediction of a sample of 5000 inputs before generating an explanation.

Stability. All IDXs are *unique* (for each input assignment-explanandum pair), but, unlike LIME and SHAP, MD-IDXs, SD-IDXs and CF-IDXs use deterministic methods (i.e., not based on sampling) and thus their explanations are also *stable* [Sokol and Flach, 2020].

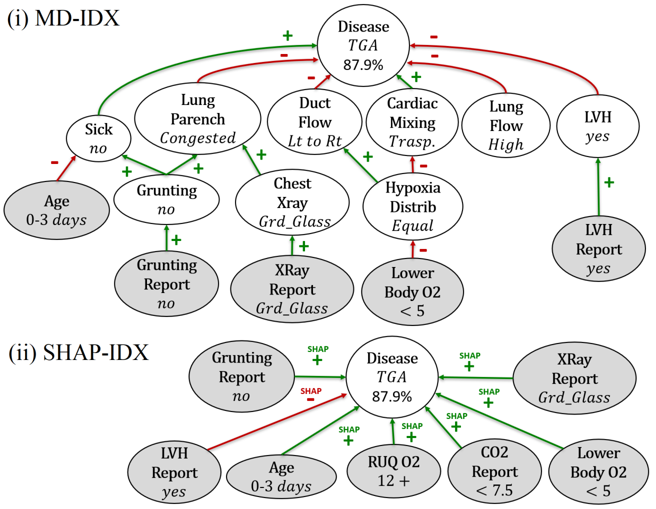

Figure 2: Example MD-IDX (i) and SHAP-IDX (ii), in graphical form, for explanandum *Disease* in the *Child* BCC. Each node represents a variable with the assigned/estimated value in italics. Grey and white nodes indicate observations and classifications, resp. $+/\overset{\text{SHAP}}{+}$ and $-/\overset{\text{SHAP}}{-}$ indicate supports and attacks, resp. Note that other ways to present IDXs to users may be suitable in some settings: this is left for future work.

Size of the explanations. In order to *understand how many influences translated into relations and their prevalence* we calculated the number of relations of each of the instantiated IDXs from Section 4; the results are reported in Table 4, showing the percentage of influences that are categorised into each relation. We note that: **(1)** MD-IDXs, SD-IDXs and CF-IDX are non-shallow, thus, when non-naive BCs are used, they also find relationships between pairs of classifications (see $\mathcal{R}^{\mathcal{C}}_{-+}$, $\mathcal{R}^{\mathcal{C}}_{\underset{-+}{\cdot\,\cdot}}$ and $\mathcal{R}^{\mathcal{C}}_{*!}$ in Table 4) that, depending on the structure of the Bayesian network underlying the BC, can be also numerous, e.g., for *Child* and *Asia*; this shows that our notion of influences provide a deeper insight into the intermediate variables of the model, that are otherwise neglected by *input-output influences*; **(2)** SD-IDXs and LIME-IDXs tend to behave similarly here, while MD-IDXs tend to include fewer influences than SD-IDXs (in line with Proposition 2), which in turn include fewer influences than critical influences in CF-IDXs (in line with Proposition 5); note that, CF-IDXs could not be computed in all settings because their computational complexity is *exponential* wrt the cardinality of the variables' sets of values, while the complexity of SM-IDXs and MD-IDXs is *linear* (see above); **(3)** in some settings, SHAP-IDXs fail to capture the majority of attacks captured by the other dialectical explanations (e.g., for *Votes* and *Emotions*).

Agreement of Dialectical Explanations. To show *how*

[6] Machine Learning Repository [UCI, 2020]

[7] National Archive of Criminal Justice Data [NACJD, 2004]

[8] ProRepublica Data Store [ProPublica, 2016]

[9] Multi-Label Classification Dataset Repository [Moyano, 2020]

[10] Bayesian Network Repository [BNlearn, 2020]

[11] We used a machine with *Intel i9-9900X* at 3.5*Ghz* and 32*GB* of RAM with no GPU acceleration. For BCCs we did not use production-ready code optimized for performances.

Dataset	SD-IDX			MD-IDX			CF-IDX			LIME-IDX		SHAP-IDX	
	$\mathcal{R}_+$	$\mathcal{R}_-$	$\mathcal{R}^{\mathcal{C}}_{-+}$	$\mathcal{R}_+$	$\mathcal{R}_-$	$\mathcal{R}^{\mathcal{C}}_{-+}$	$\mathcal{R}_!$	$\mathcal{R}_*$	$\mathcal{R}^{\mathcal{C}}_{*!}$	$\mathcal{R}_{\underset{+}{\text{LIME}}}$	$\mathcal{R}_{\underset{-}{\text{LIME}}}$	$\mathcal{R}_{\underset{+}{\text{SHAP}}}$	$\mathcal{R}_{\underset{-}{\text{SHAP}}}$
Shuttle	59.0%	41.0%	×	51.2%	32.4%	×	17.9%	46.0%	×	59.8%	40.2%	61.9%	38.1%
Votes	77.1%	22.9%	×	77.1%	22.9%	×	3.0%	74.1%	×	77.1%	22.9%	73.2%	7.3%
Parole	61.5%	38.5%	×	38.5%	27.4%	×	2.1%	70.5%	×	55.6%	44.4%	56.4%	43.6%
German	59.3%	40.7%	×	29.6%	22.0%	×	×	×	×	55.9%	44.1%	46.9%	36.4%
COMPAS	67.0%	33.0%	×	45.4%	20.3%	×	×	×	×	65.7%	34.3%	35.6%	19.1%
Car	57.4%	42.6%	×	39.3%	21.1%	×	14.2%	66.4%	×	55.3%	44.7%	55.9%	44.1%
Emotions	56.9%	24.0%	1.1%	10.3%	5.4%	1.1%	1.9%	1.5%	×	60.6%	39.4%	56.8%	10.3%
Child	77.5%	22.5%	64.0%	65.4%	15.1%	64.0%	34.6%	27.6%	64.0%	54.0%	41.3%	24.4%	9.7%
Asia	87.5%	12.5%	62.5%	87.5%	12.5%	62.5%	46.9%	9.4%	62.5%	70.8%	29.2%	54.2%	20.8%

Table 4: Average percentage of the influences that constitute relations for several instantiated IDXs. For any relation type t, $\mathcal{R}^{\mathcal{C}}_t = \{(x, y) \in \mathcal{R}_t | x, y \in \mathcal{C}\}$. × indicates results that cannot be computed in these settings or that must be 0 due to the BC type. Note that percentages may not sum to 100% because not all the influences are categorised into relations.

Dataset	MD-/LIME-IDX		MD-/SHAP-IDX		SD-/LIME-IDX		MD-/SHAP-IDX	
	Agree	Disagree	Agree	Disagree	Agree	Disagree	Agree	Disagree
Shuttle	82.4%	17.6%	83.6%	16.4%	98.8%	1.2%	97.1%	2.9%
Votes	99.1%	0.9%	78.8%	21.2%	99.1%	0.9%	78.8%	21.2%
Parole	65.8%	34.2%	65.8%	34.2%	91.1%	8.9%	94.9%	5.1%
German	50.1%	49.9%	38.2%	61.8%	83.6%	16.4%	73.5%	26.5%
COMPAS	65.7%	34.3%	20.4%	79.6%	98.3%	1.7%	54.1%	45.9%
Car	59.1%	40.9%	60.3%	39.7%	94.7%	5.3%	97.1%	2.9%
Emotions	15.0%	85.0%	12.3%	87.7%	65.0%	35.0%	51.9%	48.1%
Child	8.3%	91.7%	8.3%	91.7%	9.2%	90.8%	9.3%	90.7%
Asia	25.0%	75.0%	25.0%	75.0%	25.0%	75.0%	25.0%	75.0%

Table 5: (Dis)Agreement between MD-IDXs/SD-IDXs and LIME-IDXs/SHAP-IDXs. Agreement means identifying the same (possibly empty) relations for the same influences, in practice $|\mathcal{R}_\cdot|/|I|$, where $\cdot$ is the relation type and, accordingly with Defs. 2 and 3, $I = \mathcal{I}$ for SD-/MD-/CF-IDX and $I = \mathcal{I}_{io}$ for LIME-/SHAP-IDX.

our explanations differ from existing dialectical explanations we compared the relations in MD-IDXs/SD-IDXs with those in LIME-IDXs/SHAP-IDXs, analysing how often they agree in identifying attacks or supports between observations and classifications. Table 5 shows the results in percentage for each pair, e.g, between MD-IDX and LIME-IDX $|(\mathcal{R}_- \cap \mathcal{R}_{\underset{-}{\text{LIME}}}) \cup (\mathcal{R}_+ \cap \mathcal{R}_{\underset{+}{\text{LIME}}})|/|\mathcal{I}_{io}|$. We note that: **(1)** MD-IDXs agree on average 52.30% and 43.6% of the time while SD-IDXs agree on average 73.9% and 64.63% of the time with LIME-IDXs and SHAP-IDXs, respectively, due to MD-IDXs' stronger constraints on the selection of attacks and supports; **(2)** when a BCC with many classifications is used (as in *Child*, *Asia*, *Emotions*), the agreement decreases considerably, due to LIME-IDXs and SHAP-IDXs being shallow, and thus selecting different influences from non-shallow MD-IDXs and SD-IDXs, as described in Section 4.4 and exemplified by Figure 2.

Satisfaction of Dialectical Monotonicity. We checked the number of influences in relations in CF-IDXs, LIME-IDXs and SHAP-IDXs which do not satisfy *dialectical monotonicity* (Property 1), which holds for MD-IDXs. The results are in Table 6. We note that: **(1)** SM-IDXs violate the dialectical monotonicity constraint significantly ($p < 0.05$) less than other methods for all the NBCs, while their increase in the number of violations of dialectical monotonicity for BCCs is due to SM-IDXs being non-shallow, unlike LIME-IDXs and SHAP-IDXs; **(2)** all three methods violate the dialectical monotonicity constraint. The violation of dialectical monotonicity gives rise to counter-intuitive results from a dialectical perspective. Consider the example in Figure 1: if *raining* is considered to be an output, LIME draws a negative contribution from *windy* to *raining*, which makes little sense given the structure of the BC (this also indicates the benefits of incorporating a model's structure into IDXs via influences). Further, for the SHAP-IDX in Figure 2ii, changing the value of *Disease*'s supporter, *Age*, so that it is no longer a supporter, results in the probability of *Disease* being *TGA* increasing.

Dataset	SM-IDX	LIME-IDX	SHAP-IDX
Shuttle	5.8%	6.8%	6.5%
Votes	0.0%	0.2%	0.1%
Parole	13.3%	13.2%	13.8%
German	18.5%	20.8%	19.8%
COMPAS	12.3%	12.5%	22.7%
Car	16.8%	18.5%	17.4%
Emotions	12.0%	11.9%	8.9%
Child	7.1%	2.5%	5.6%
Asia	0.0%	0.0%	0.0%

Table 6: Average percentage of influences in relations *not* satisfying the dialectical monotonicity property, obtained with a sample of 25,000 influences in relations for 250 data-points.

6 Conclusions

We have introduced IDXs, a general formalism for producing various forms of explanations for BCs. We have demonstrated IDXs' versatility by giving novel forms of dialectical explanations as IDX instances, based on the principle of monotonicity, and integrating existing notions of explanation. We have performed a wide-ranging evaluation with theoretical and empirical analyses of IDXs with respect to existing methods, identifying some advantages of IDXs with non-naive BCs.

Acknowledgements

Francesca Toni was partially supported by the Royal Academy of Engineering and by JP Morgan under the Research Chairs and Senior Research Fellowships scheme.

References

[Albini *et al.*, 2020] Emanuele Albini, Antonio Rago, Pietro Baroni, and Francesca Toni. Relation-based counterfactual explanations for bayesian network classifiers. In *Proceedings of the Twenty-Ninth Int. Joint Conf. on Artificial Intelligence, IJCAI 2020*, pages 451–457, 2020.

[Bach *et al.*, 2015] Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLOS ONE*, 10(7):1–46, 07 2015.

[Bau *et al.*, 2017] David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. Network dissection: Quantifying interpretability of deep visual representations. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, pages 3319–3327, 2017.

[BNlearn, 2020] BNlearn. Bayesian network repository - an r package for bayesian network learning and inference, 2020.

[Cayrol and Lagasquie-Schiex, 2005] Claudette Cayrol and Marie-Christine Lagasquie-Schiex. On the acceptability of arguments in bipolar argumentation frameworks. In *Proceedings of the 8th European Conf. on Symbolic and Quantitative Approaches to Reasoning with Uncertainty*, pages 378–389. Springer Berlin Heidelberg, 2005.

[Cocarascu *et al.*, 2019] Oana Cocarascu, Antonio Rago, and Francesca Toni. Extracting dialogical explanations for review aggregations with argumentative dialogical agents. In *Proceedings of the 18th Int. Conf. on Autonomous Agents and MultiAgent Systems, AAMAS '19*, pages 1261–1269, 2019.

[Cyras *et al.*, 2019] Kristijonas Cyras, Dimitrios Letsios, Ruth Misener, and Francesca Toni. Argumentation for explainable scheduling. In *The Thirty-Third AAAI Conf. on Artificial Intelligence, AAAI 2019*, pages 2752–2759, 2019.

[Darwiche and Hirth, 2020] Adnan Darwiche and Auguste Hirth. On the reasons behind decisions. In *24th European Conference on Artificial Intelligence (ECAI 2020)*, 2020.

[Dung, 1995] Phan Minh Dung. On the Acceptability of Arguments and its Fundamental Role in Nonmonotonic Reasoning, Logic Programming and n-Person Games. *Artificial Intelligence*, 77(2):321–358, 1995.

[Enrique Sucar *et al.*, 2014] L. Enrique Sucar, Concha Bielza, Eduardo F. Morales, Pablo Hernandez-Leal, Julio H. Zaragoza, and Pedro Larrañaga. Multi-label classification with bayesian network-based chain classifiers. *Pattern Recognition Letters*, 41:14 – 22, 2014.

[Fan and Toni, 2015] Xiuyi Fan and Francesca Toni. On computing explanations in argumentation. In *Proceedings of the Twenty-Ninth AAAI Conf. on Artificial Intelligence*, pages 1496–1502, 2015.

[Friedman *et al.*, 1997] Nir Friedman, Dan Geiger, and Moisés Goldszmidt. Bayesian network classifiers. *Mach. Learn.*, 29(2-3):131–163, 1997.

[García *et al.*, 2013] Alejandro Javier García, Carlos Iván Chesñevar, Nicolás D. Rotstein, and Guillermo Ricardo Simari. Formalizing dialectical explanation support for argument-based reasoning in knowledge-based systems. *Expert Syst. Appl.*, 40(8):3233–3247, 2013.

[Guidotti *et al.*, 2019] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. A survey of methods for explaining black box models. *ACM Comput. Surv.*, 51(5):93:1–93:42, 2019.

[Halpern and Pearl, 2001a] Joseph Y. Halpern and Judea Pearl. Causes and explanations: A structural-model approach - part II: explanations. In *Proceedings of the Seventeenth Int. Joint Conf. on Artificial Intelligence, IJCAI 2001*, pages 27–34, 2001.

[Halpern and Pearl, 2001b] Joseph Y. Halpern and Judea Pearl. Causes and explanations: A structural-model approach: Part 1: Causes. In *UAI '01: Proceedings of the 17th Conf. in Uncertainty in Artificial Intelligence*, pages 194–202, 2001.

[Ignatiev *et al.*, 2019a] Alexey Ignatiev, Nina Narodytska, and João Marques-Silva. Abduction-based explanations for machine learning models. In *The Thirty-Third AAAI Conf. on Artificial Intelligence, AAAI 2019*, pages 1511–1519, 2019.

[Ignatiev *et al.*, 2019b] Alexey Ignatiev, Nina Narodytska, and João Marques-Silva. On relating explanations and adversarial examples. In *Advances in Neural Information Processing Systems 32: Annual Conf. on Neural Information Processing Systems 2019, NeurIPS 2019*, pages 15857–15867, 2019.

[Ignatiev, 2020] Alexey Ignatiev. Towards trustable explainable AI. In *Proceedings of the Twenty-Ninth Int. Joint Conf. on Artificial Intelligence, IJCAI 2020*, pages 5154–5158, 2020.

[Koller and Sahami, 1996] Daphne Koller and Mehran Sahami. Toward optimal feature selection. In *Machine Learning, Proceedings of the Thirteenth Int. Conf. (ICML '96)*, pages 284–292, 1996.

[Lundberg and Lee, 2017] Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In

Advances in Neural Information Processing Systems 30: Annual Conf. on Neural Information Processing Systems 2017, pages 4765–4774, 2017.

[Miller, 2019] Tim Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267:1–38, 2019.

[Moyano, 2020] Jose M. Moyano. Multi-label classification dataset repository, 2020.

[NACJD, 2004] National Archive of Criminal Justice Data NACJD. National corrections reporting program, 2004.

[Naveed *et al.*, 2018] Sidra Naveed, Tim Donkers, and Jürgen Ziegler. Argumentation-based explanations in recommender systems: Conceptual framework and empirical results. In *Adjunct Publication of the 26th Conf. on User Modeling, Adaptation and Personalization, UMAP 2018*, pages 293–298, 2018.

[Nielsen *et al.*, 2008] Ulf H. Nielsen, Jean-Philippe Pellet, and André Elisseeff. Explanation trees for causal bayesian networks. In *UAI 2008, Proceedings of the 24th Conf. in Uncertainty in Artificial Intelligence*, pages 427–434, 2008.

[Olah *et al.*, 2018] Chris Olah, Arvind Satyanarayan, Ian Johnson, Shan Carter, Ludwig Schubert, Katherine Ye, and Alexander Mordvintsev. The building blocks of interpretability. *Distill*, 3(3):e10, 2018.

[Poyiadzi *et al.*, 2020] Rafael Poyiadzi, Kacper Sokol, Raúl Santos-Rodríguez, Tijl De Bie, and Peter A. Flach. FACE: feasible and actionable counterfactual explanations. In *AIES '20: AAAI/ACM Conf. on AI, Ethics, and Society*, pages 344–350, 2020.

[ProPublica, 2016] Data Store ProPublica. Compas recidivism risk score data and analysis, 2016.

[Rago *et al.*, 2018] Antonio Rago, Oana Cocarascu, and Francesca Toni. Argumentation-based recommendations: Fantastic explanations and how to find them. In *Proceedings of the Twenty-Seventh Int. Joint Conf. on Artificial Intelligence, IJCAI 2018*, pages 1949–1955, 2018.

[Ribeiro *et al.*, 2016] Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin. "Why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, pages 1135–1144, 2016.

[Ribeiro *et al.*, 2018] Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin. Anchors: High-precision model-agnostic explanations. In *Proceedings of the Thirty-Second AAAI Conf. on Artificial Intelligence (AAAI-18)*, pages 1527–1535, 2018.

[Schwab and Karlen, 2019] Patrick Schwab and Walter Karlen. Cxplain: Causal explanations for model interpretation under uncertainty. In *Advances in Neural Information Processing Systems 32: Annual Conf. on Neural Information Processing Systems 2019, NeurIPS 2019*, pages 10220–10230, 2019.

[Sharma *et al.*, 2020] Shubham Sharma, Jette Henderson, and Joydeep Ghosh. CERTIFAI: A common framework to provide explanations and analyse the fairness and robustness of black-box models. In *AIES '20: AAAI/ACM Conf. on AI, Ethics, and Society*, pages 166–172, 2020.

[Shih *et al.*, 2018] Andy Shih, Arthur Choi, and Adnan Darwiche. A symbolic approach to explaining bayesian network classifiers. In *Proceedings of the Twenty-Seventh Int. Joint Conf. on Artificial Intelligence, IJCAI 2018*, pages 5103–5111, 2018.

[Shih *et al.*, 2019] Andy Shih, Arthur Choi, and Adnan Darwiche. Compiling bayesian network classifiers into decision graphs. In *The Thirty-Third AAAI Conf. on Artificial Intelligence, AAAI 2019*, pages 7966–7974, 2019.

[Sokol and Flach, 2020] Kacper Sokol and Peter A. Flach. Explainability fact sheets: a framework for systematic assessment of explainable approaches. In *FAT* '20: Conference on Fairness, Accountability, and Transparency, Barcelona, Spain, January 27-30, 2020*, pages 56–67, 2020.

[Teze *et al.*, 2018] Juan Carlos Teze, Lluis Godo, and Guillermo Ricardo Simari. An argumentative recommendation approach based on contextual aspects. In *Scalable Uncertainty Management - 12th Int. Conf., SUM 2018*, pages 405–412, 2018.

[Timmer *et al.*, 2015] Sjoerd T. Timmer, John-Jules Ch. Meyer, Henry Prakken, Silja Renooij, and Bart Verheij. Explaining bayesian networks using argumentation. In *Symbolic and Quantitative Approaches to Reasoning with Uncertainty - 13th European Conf., ECSQARU 2015*, pages 83–92, 2015.

[UCI, 2020] Center for Machine Learning and Intelligent Systems UCI. Machine Learning Repository, 2020.

[White and d'Avila Garcez, 2020] Adam White and Artur S. d'Avila Garcez. Measurable counterfactual local explanations for any classifier. In *24th European Conference on Artificial Intelligence (ECAI 2020)*, 2020.

[Zeng *et al.*, 2018] Zhiwei Zeng, Xiuyi Fan, Chunyan Miao, Cyril Leung, Jing Jih Chin, and Yew-Soon Ong. Context-based and explainable decision making with argumentation. In *Proceedings of the 17th Int. Conf. on Autonomous Agents and MultiAgent Systems, AAMAS 2018*, pages 1114–1122, 2018.

Appendix A: Theoretical analysis proofs

Proposition 1.

Proof. If $\mathcal{I} = \mathcal{I}_{io}$ then from Defs. 2 and 3, $\{(x,c) \in \mathcal{X} \times \mathcal{C} | (c,x) \in \mathcal{D}\} = \mathcal{O} \times \mathcal{C}_o$ and it follows that $\mathcal{D} = \mathcal{C}_o \times \mathcal{O}$. If $\mathcal{D} = \mathcal{C}_o \times \mathcal{O}$ then from Defs. 2 and 3, $\mathcal{I} = \mathcal{O} \times \mathcal{C}_o = \mathcal{I}_{io}$. □

Proposition 2.

Proof. For any $(x,y) \in \mathcal{I}$ and $a \in \mathcal{A}$, by Def. 7, if $\pi_-((x,y),a) = true$ then $\forall x_k \in \mathcal{V}(x) \setminus \{\sigma(a,x)\}, P(\sigma(a,y)|a) < P(\sigma(a,y)|a'_{x_k})$.

Setting	Data Source*	Classifier Implementation†
Shuttle	DT	`skl`
Parole	DT	`skl`
German	DT	`skl`
COMPAS	DT	`skl`
Car	DT	`skl`
Emotions	DT	Chain of `skl`
Asia	BN	Chain of `pgm`
Child	BN	Chain of `pgm`

Table 7: Experimental settings details (∗) **D**a**T**aset (DT) or **B**ayesian **N**etwork (BN); (†) `sklearn.CategoricalNB` (`skl`) or `pgmpy.BayesianModel` (`pgp`);

Then we get: $\frac{\sum_{x_k \in \mathcal{V}(x) \setminus \{\sigma(a,x)\}} \left[P(x_k) \cdot P(\sigma(a,y)|a'_{x_k})\right]}{\sum_{x_k \in \mathcal{V}(x) \setminus \{\sigma(a,x)\}} P(x_k)} \geq \frac{\sum_{x_k \in \mathcal{V}(x) \setminus \{\sigma(a,x)\}} \left[P(x_k) \cdot \min_{x_k \in \mathcal{V}(x) \setminus \{\sigma(a,x)\}} P(\sigma(a,y)|a'_{x_k})\right]}{\sum_{x_k \in \mathcal{V}(x) \setminus \{\sigma(a,x)\}} P(x_k)} = \min_{x_k \in \mathcal{V}(x) \setminus \{\sigma(a,x)\}} P(\sigma(a,y)|a'_{x_k}) > P(\sigma(a,y)|a)$.

Then by Def. 8, $\pi_{\dot{-}}((x,y),a) = \mathit{true}$. The proof for $\mathcal{R}_{+} \subseteq \mathcal{R}_{\dot{+}}$ is analogous. It follows from Def. 7 that $\mathcal{X}_r \subseteq \mathcal{X}'_r$. □

Proposition 3.

Proof. If for any $x \in \mathcal{X}'_r \setminus \{e\}$, $|\mathcal{V}(x)| = 2$ then from Defs. 7 and 8 $\frac{\sum_{x_k \in \mathcal{V}(x) \setminus \{\sigma(a,x)\}} \left[P(x_k) \cdot P(\sigma(a,y)|a'_{x_k})\right]}{\sum_{x_k \in \mathcal{V}(x) \setminus \{\sigma(a,x)\}} P(x_k)} = \frac{P(x_k) \cdot P(\sigma(a,y)|a'_{x_k})}{P(x_k)} = P(\sigma(a,y)|a'_{x_k})$, where x_k is the only element of $\mathcal{V}(x) \setminus \{\sigma(a,x)\}$. Thus, $\pi_{-} = \pi_{\dot{-}}$ and $\pi_{+} = \pi_{\dot{+}}$, giving $\mathcal{R}_{-} = \mathcal{R}_{\dot{-}}$ and $\mathcal{R}_{+} = \mathcal{R}_{\dot{+}}$, resp. It follows from Def. 7 that $\mathcal{X}_r = \mathcal{X}'_r$. □

Proposition 4.

Proof. Monotonically dialectical explanation kits are monotonic by inspection of Definition 7. Stochastically dialectical, LIME and SHAP explanation kits are not monotonic as proved by the results in Table 6. □

Proposition 5.

Proof. For any $(x,y) \in \mathcal{I}$ and $a \in \mathcal{A}$, by Def. 9, if $\pi_{!}((x,y),a) = \mathit{true}$ then $\forall a' \in \mathcal{A}$ such that $\sigma(a',x) \neq \sigma(a,x)$ and $\forall z \in \mathcal{I}(y) \backslash \{x\}$ $\sigma(a',z) = \sigma(a,z)$, it holds that $\sigma(a,y) \neq \sigma(a',y)$. Thus, it must be the case that $\forall x_k \in \mathcal{V}(x) \setminus \{\sigma(a,x)\}, P(\sigma(a,y)|a) > P(\sigma(a,y)|a'_{x_k})$ and so, by Def. 7, $\pi_{+}((x,y),a) = \mathit{true}$. □

Appendix B: Empirical Experiments Setup

Datasets and dataset splits

As reported is Table 7 we ran experiments in several settings. We divided the datasets in a randomly stratified train and test sets with a split 75/25% split. When the data source was a Bayesian network we artificially generated the corresponding combinatorial dataset using variable elimination (an exact inference algorithm for Bayesian network).

Note that the experiments on the explanations were run on the test set, and if the number of samples in the test set was greater than 250 samples, we ran them on a random sample of 250 samples.

Execution details

LIME explanations were generated using default parameters. SHAP explanations were generated using default parameters. For all the random computations (`sklearn`, `pandas`, `numpy`, `random`) we used a random seed of 0.

Software Platform and Computing Infrastructure

To run our experiments, we used a machine with Ubuntu 18.04, a single Intel i9-9900X processor, 32GB of RAM and no GPU acceleration. We used a Python 3.6 environment with networkx 2.2, pgmpy 0.1.11, scikit-learn 0.22.1, shap 0.35.0 and lime 0.1.1.37 as software platform to run our experiments.

Hyper-parameter training

The classifier `sklearn CategoricalNB` has 2 hyper-parameters: $\alpha \in \mathbb{R}$ (Laplace smoothing coefficient) and a tuple of probabilities $\beta = \langle \beta_1, \ldots \beta_m \rangle$ (classes prior probabilities) such that $\beta_i \in [0,1]$ and m is the number of classes. `pgmpy BayesianModel` has no learning parameters (all probabilities are given in the Bayesian network, they do not need to be learnt). Datasets with numeric features have more hyper-parameters: the type of bins to make numeric variables categorical, that we denote as $\gamma \in \Gamma = \{SS, SL, custom\}$ where SS denotes Same Size (i.e. each bucket has the same number of elements from the dataset), SL denotes Same Length (i.e. each bucket covers an interval of the same length) and *custom* denotes a custom discretization, the number of buckets used in the discretization, denoted with δ_i for each numeric feature i.

Shuttle. $\alpha = 5$, $\beta = \langle 0.8, 0.2 \rangle$.

Votes. $\alpha = 1$, $\beta = \langle 0.4, 0.6 \rangle$.

Parole.$\alpha = 0.1$, $\beta = \langle 0.9, 0.1 \rangle$, $\gamma = SL$, $\delta_{timeserved} = 2$, $\delta_{maxsentence} = 11$, $\delta_{age} = 9$.

German. $\alpha = 0.00001$, $\beta = \langle 0.35, 0.65 \rangle$, $\gamma = SL$, $\delta_{age} = 10$, $\delta_{amount} = 10$, $\delta_{duration} = 9$.

COMPAS. $\alpha = 0.1$, $\beta = \mathit{auto}$, $\gamma = \mathit{custom}$. We used a custom discretization for the dataset obtained with manual trials because it was performing better than automatic ones (same size or same length bins).

Car. $\alpha = 1$, $\beta = \langle 0.2, 0.1, 0.6, 0.1 \rangle$.

Emotions. $\alpha = 0.00001$, $\beta = \langle 0.6, 0.4 \rangle$, $\gamma = SS$, $\delta_i = 9$ for all the features. We built the classes tree using the mutual information coefficient as described in [Enrique Sucar *et al.*, 2014], but instead of creating an ensemble of chain classifiers we considered only a single one and we considered the root of the tree as an hyper-parameter, (`amazed-lonely` was deemed as root in our case).