Enhancing SHAP Explanation Interpretability Using Subgroup Discovery

Maëlle Moranges, Thomas Guyet

Lyon 1

AIstroSight

HCL
HOSPICES CIVILS
DE LYON

Inria

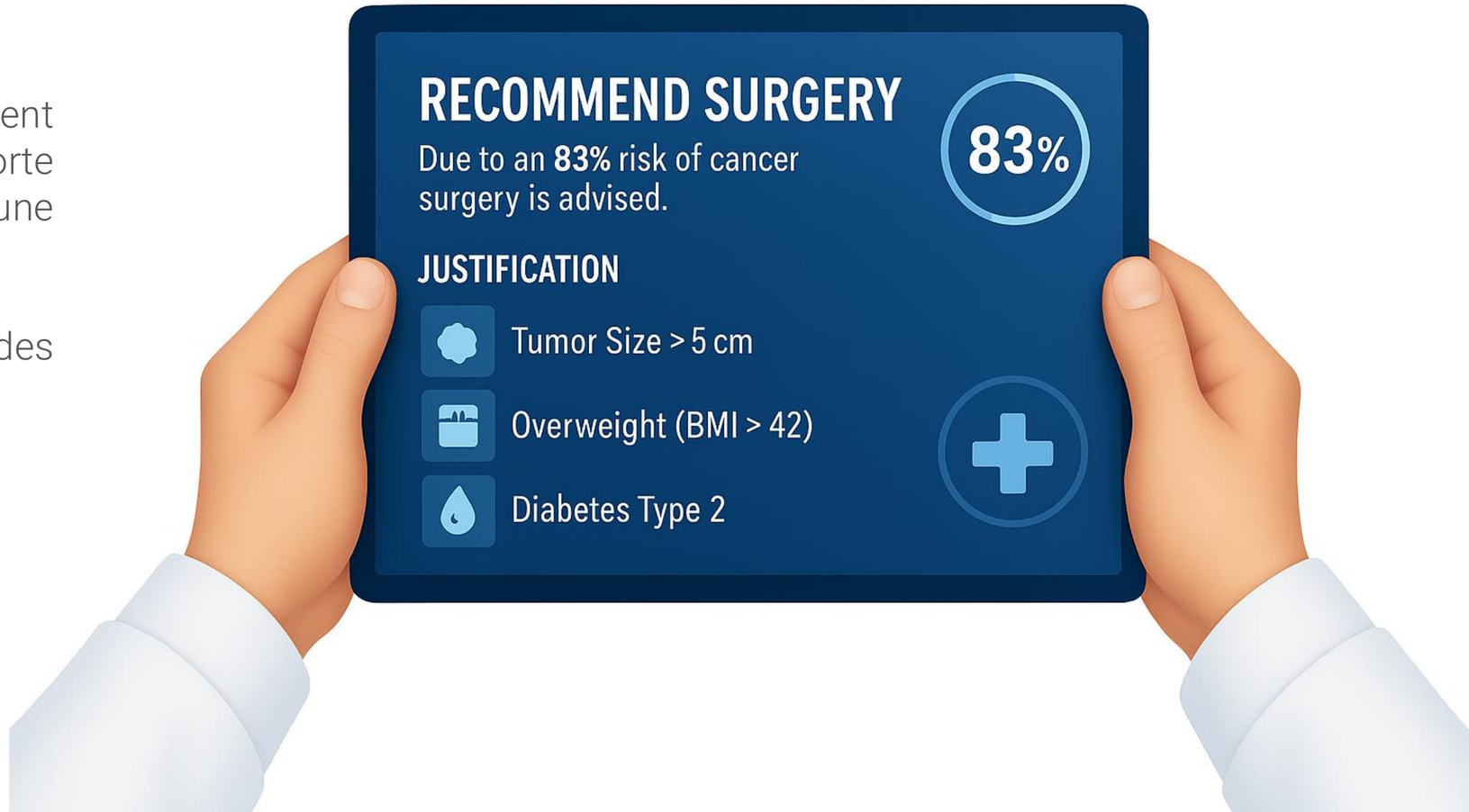
27/01/2026

Classification en médecine

Les modèles prédictifs sont largement employés en médecine, mais leur forte performance s'accompagne souvent d'une opacité qui limite leur adoption clinique.

Leur utilisation en pratique nécessite des explications :

- compréhensibles,
- fiable
- cliniquement pertinentes.



XAI en médecine

Dominant Method in Healthcare: *SHAP (83% of studies with XAI)*

✅ Model-agnostic, post-hoc

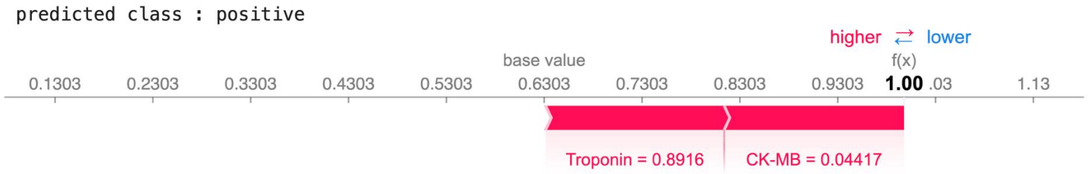
✅ Local (patient) & global (model) explanations

✅ Attractive visualizations

❌ Limitations:

- Oversimplifies
- Lack of variable interactions
- SHAP values remain abstract for clinicians
- Global averages mask patient diversity

Example:



Rule-based method : *Anchors, LORE and SuRE*

✅ Model-agnostic, post-hoc

✅ Close to clinical reasoning

✅ Capturing variable interactions

❌ Limitations:

- Only local or global explanations
- Rules are often long and therefore complex

Example: CK-MB $\geq -0.21 \ \wedge$ Troponin $\geq -0.30 \rightarrow \ positive$

Références : Lundberg et Lee. 2017. ; Ribeiro et al. 2018. ; Guidotti et al. 2019. ; Yuan et al. 2022.

Purpose

Objective: producing explanations that are:

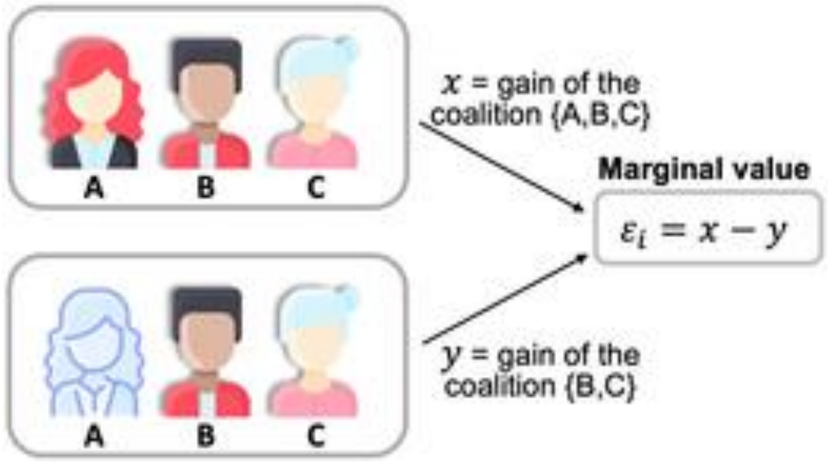
- Model-agnostic, post-hoc
- Clinically interpretable
- Integrating interactions between variables
- Accounting for individual variability
- Both global and local

Proposed approach: Generation of explicit IF-THEN rules by combining 2 complementary approaches in AI:

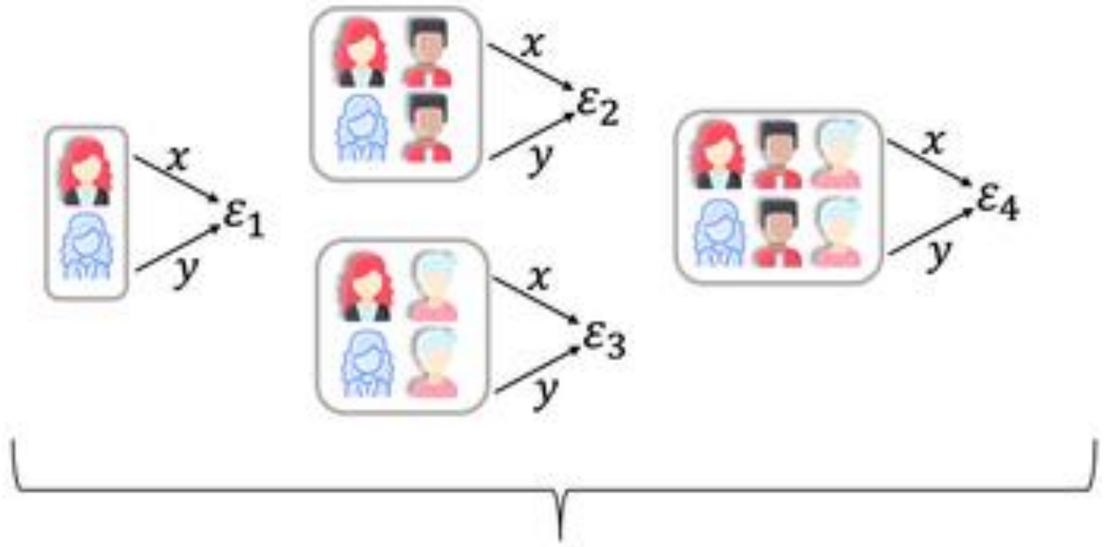
- **Explainability (XAI)** : explanation of internal reasoning (local & global) using *SHAP*
- **Interpretability (Data Mining)** : extraction of descriptive rules through *Subgroup Discovery*

Shapley value (in game theory)

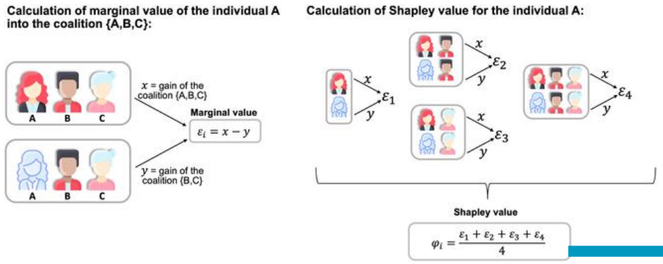
Calculation of marginal value of the individual A into the coalition {A,B,C}:



Calculation of Shapley value for the individual A:



Shapley value (in XAI)



=feature i

=other features

shapley value for a feature i for an instance

For the label c :

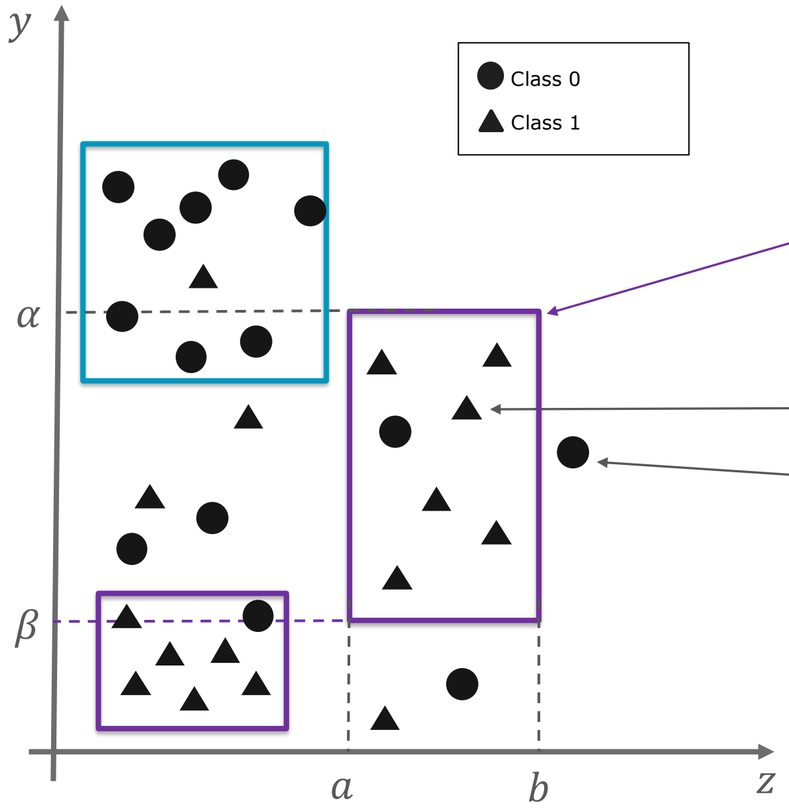
	Features				
Instance		f_1	f_2	...	f_k
	x_1	φ_{x_1,f_1}	φ_{x_1,f_2}	...	φ_{x_1,f_k}
	x_2	φ_{x_2,f_1}	φ_{x_2,f_2}	...	φ_{x_2,f_k}
	...	...	...	...	...
	x_n	φ_{x_n,f_1}	φ_{x_n,f_2}	...	φ_{x_n,f_k}

$\varphi_{x,f} > 0$ if f contributes to predicting c for x

$\varphi_{x,f} = 0$ if f does not contribute to predicting c for x

$\varphi_{x,f} < 0$ if f contributes to predicting $\neg c$ for x

Subgroup Discovery



subgroup :
$z \in [a, b]\ and\ y \in [\alpha, \beta] \rightarrow 1$

Target class

Negative (non-target) classes

$$WRAcc\ (R) = \frac{\#}{\#} \left(\frac{\#}{\#} - \frac{\#}{\#} \right)$$

Wrobel , 1997

Lavrač, 2004

Aim

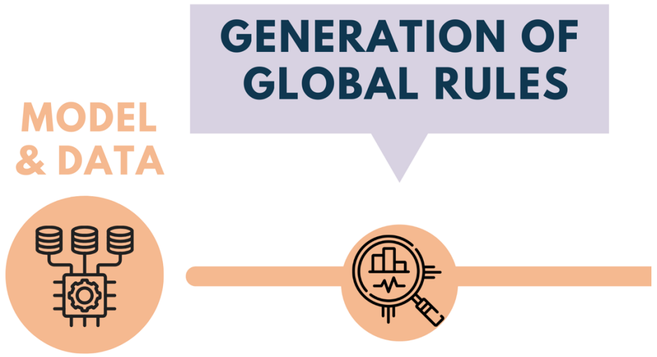
Dataset

	f_1	...	f_k	Predicted class
x_1	2.1		M	1
x_2	3		F	0
...				
x_n	2.6		M	1

subgroup : $f_1 \in [2,3]$ ^ $f_k = M \rightarrow 1$

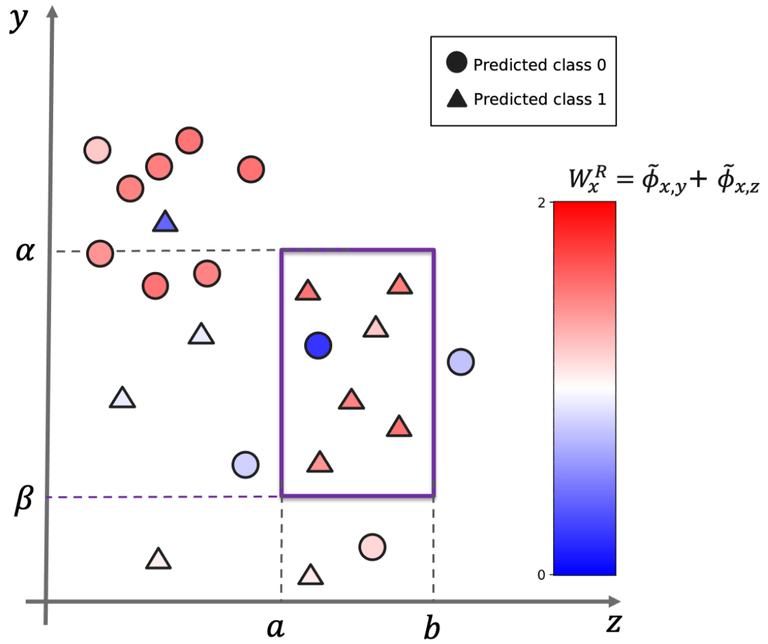
Find the rules present in the dataset that rely on attributes that have actually contributed to the model's prediction.

METHOD



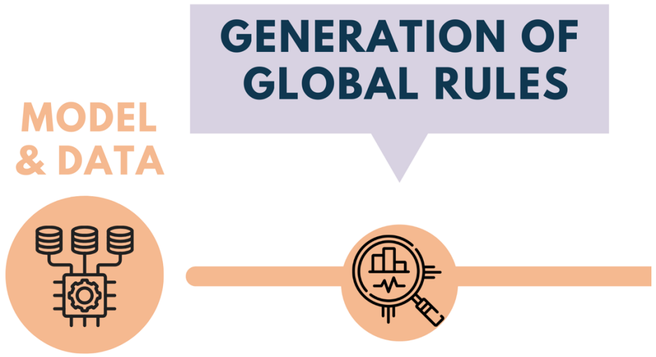
Extraction of class-wise descriptive rules via Subgroup Discovery, with optimization of a SHAP-weighted WRAcc

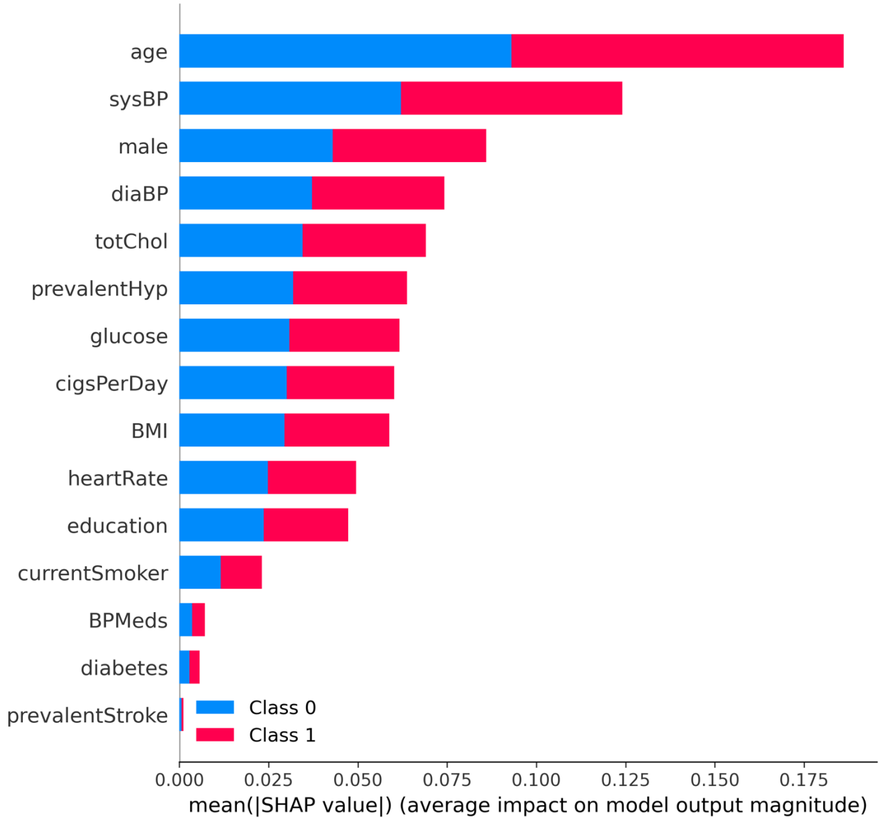
$$R = z \in [a, b]\ AND\ y \in [\alpha, \beta] \rightarrow 1$$



$$Wracc_{\phi}(R) = \frac{\sum}{\sum}\left(\frac{\sum}{\sum} - \frac{\sum}{\sum}\right)$$

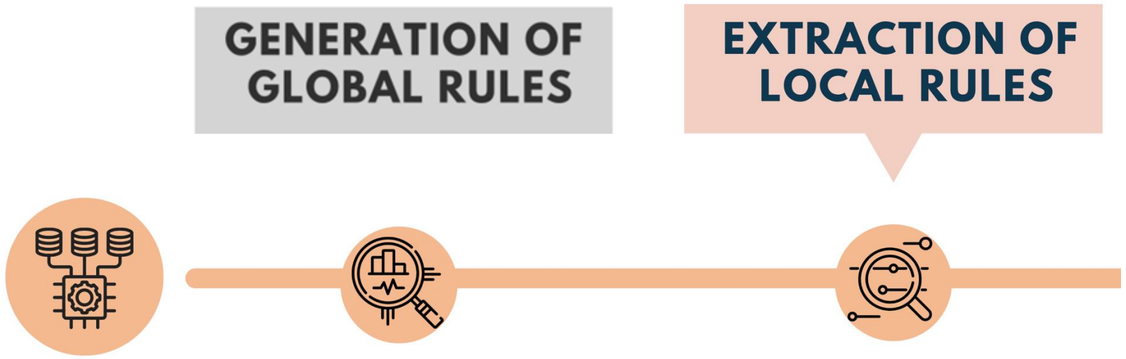
METHOD





$\mathrm{WRAcc}_\phi(R)$	R	c
0.084	prevalentHyp=0	$\rightarrow 0$
0.076	male=0 AND prevalentHyp=0	$\rightarrow 0$
0.076	male=0	$\rightarrow 0$
0.073	age < 0.26	$\rightarrow 0$
0.071	diabetes=0 AND prevalentHyp=0	$\rightarrow 0$
0.071	BPMeds=0 AND prevalentHyp=0	$\rightarrow 0$
0.069	prevalentHyp=0 AND prevalentStroke=0	$\rightarrow 0$
0.063	BPMeds=0 AND male=0	$\rightarrow 0$
0.062	diabetes=0 AND male=0	$\rightarrow 0$
0.061	age $\in [0.26 : 0.42[$	$\rightarrow 0$
0.084	prevalentHyp=1	$\rightarrow 1$
0.082	age ≥ 0.74	$\rightarrow 1$
0.077	sysBP ≥ 0.32	$\rightarrow 1$
0.076	male=1	$\rightarrow 1$
0.069	prevalentHyp=1 AND prevalentStroke=0	$\rightarrow 1$
0.068	age $\geq 0.74 AND prevalentStroke = 0$	$\rightarrow 1$
0.066	age $\geq 0.74 AND diabetes = 0$	$\rightarrow 1$
0.065	prevalentHyp=1 AND sysBP ≥ 0.32	$\rightarrow 1$
0.064	prevalentStroke=0 AND sysBP ≥ 0.32	$\rightarrow 1$
0.062	diabetes=0 AND prevalentHyp=1	$\rightarrow 1$

METHOD



For each instance, local explanations are obtained by selecting and ranking the activated global rules:

1. Selection of rules covering the instance
2. Retention of rules with positive SHAP contributions for all variables
3. Ranking of rules by mean SHAP contribution

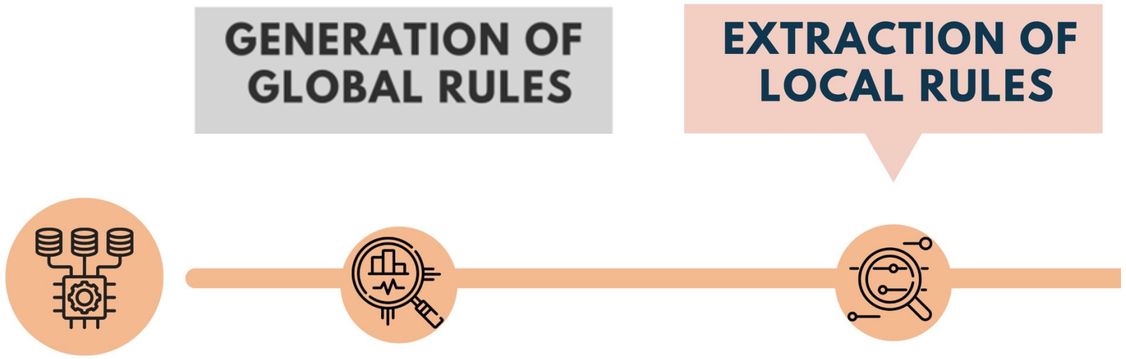
Instance

	sysBP	Prevalent Hyp	Prevalent Stroke	BPMeds	diabetes	...	age	male
x_1	0.33	1	0	0	0	...	0.42	0

Globales explanations

$\text{WRAcc}_\phi(R)$	R	c
0.084	prevalentHyp=0	$\rightarrow 0$
0.076	male=0 AND prevalentHyp=0	$\rightarrow 0$
0.076	male=0	$\rightarrow 0$
0.073	age < 0.26	$\rightarrow 0$
0.071	diabetes=0 AND prevalentHyp=0	$\rightarrow 0$
0.071	BPMeds=0 AND prevalentHyp=0	$\rightarrow 0$
0.069	prevalentHyp=0 AND prevalentStroke=0	$\rightarrow 0$
0.063	BPMeds=0 AND male=0	$\rightarrow 0$
0.062	diabetes=0 AND male=0	$\rightarrow 0$
0.061	age $\in [0.26 : 0.42[$	$\rightarrow 0$
0.084	prevalentHyp=1	$\rightarrow 1$
0.082	age ≥ 0.74	$\rightarrow 1$
0.077	sysBP ≥ 0.32	$\rightarrow 1$
0.076	male=1	$\rightarrow 1$
0.069	prevalentHyp=1 AND prevalentStroke=0	$\rightarrow 1$
0.068	age $\geq 0.74 AND prevalentStroke = 0$	$\rightarrow 1$
0.066	age $\geq 0.74 AND diabetes = 0$	$\rightarrow 1$
0.065	prevalentHyp=1 AND sysBP ≥ 0.32	$\rightarrow 1$
0.064	prevalentStroke=0 AND sysBP ≥ 0.32	$\rightarrow 1$
0.062	diabetes=0 AND prevalentHyp=1	$\rightarrow 1$

METHOD



For each instance, local explanations are obtained by selecting and ranking the activated global rules:

1. Selection of rules covering the instance
2. Retention of rules with positive SHAP contributions for all variables
3. Ranking of rules by mean SHAP contribution

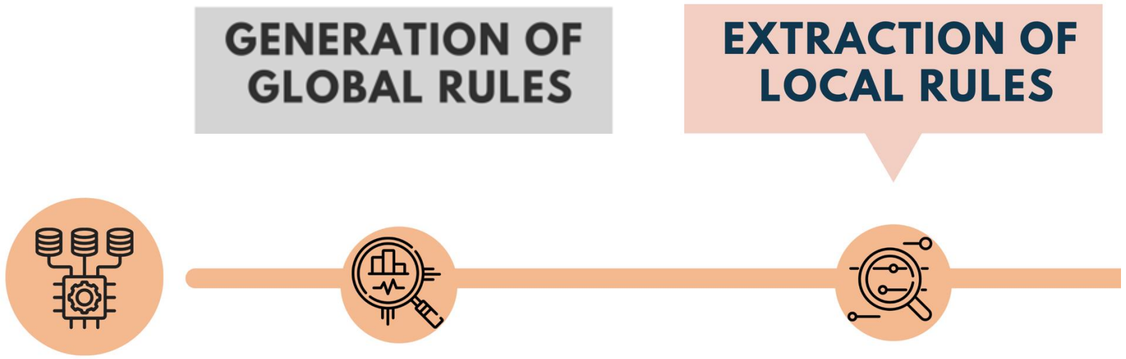
Instance

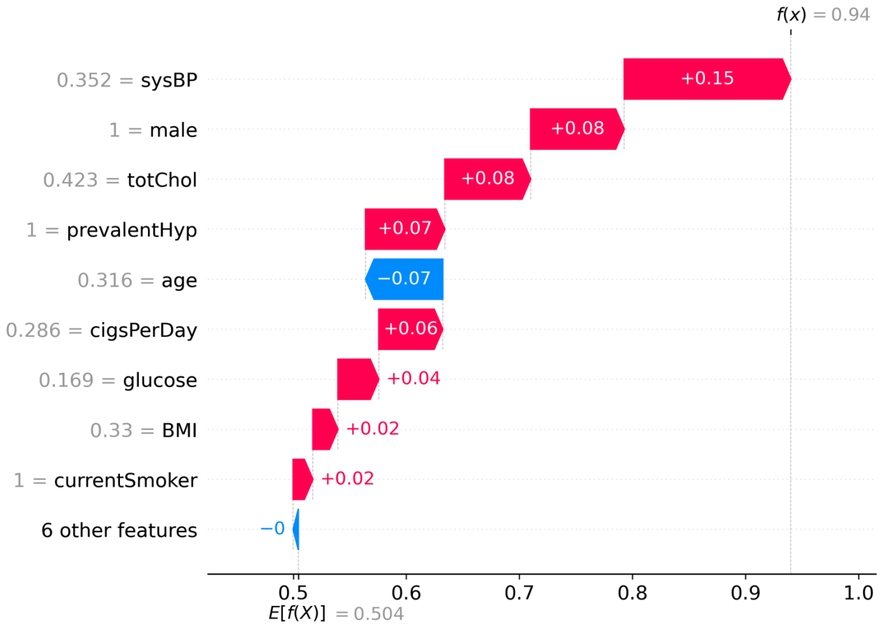
	sysBP	Prevalent Hyp	Prevalent Stroke	BPMeds	diabetes	...	age	male
x_1	0.33	1	0	0	0	...	0.42	0
φ_{x_1}	-0.024	-0.021	0.001	0.002	0.003		0.109	0.058

Globales explanations

$\mathrm{WRAcc}_\phi(R)$	R	c
0.084	prevalentHyp=0	$\rightarrow 0$
0.076	male=0 AND prevalentHyp=0	$\rightarrow 0$
0.076	male=0	$\rightarrow 0$
0.073	age < 0.26	$\rightarrow 0$
0.071	diabetes=0 AND prevalentHyp=0	$\rightarrow 0$
0.071	BPMeds=0 AND prevalentHyp=0	$\rightarrow 0$
0.069	prevalentHyp=0 AND prevalentStroke=0	$\rightarrow 0$
0.063	BPMeds=0 AND male=0	$\rightarrow 0$
0.062	diabetes=0 AND male=0	$\rightarrow 0$
0.061	age $\in [0.26 : 0.42[$	$\rightarrow 0$
0.084	prevalentHyp=1	$\rightarrow 1$
0.082	age ≥ 0.74	$\rightarrow 1$
0.077	sysBP ≥ 0.32	$\rightarrow 1$
0.076	male=1	$\rightarrow 1$
0.069	prevalentHyp=1 AND prevalentStroke=0	$\rightarrow 1$
0.068	age $\geq 0.74 AND prevalentStroke = 0$	$\rightarrow 1$
0.066	age $\geq 0.74 AND diabetes = 0$	$\rightarrow 1$
0.065	prevalentHyp=1 AND sysBP ≥ 0.32	$\rightarrow 1$
0.064	prevalentStroke=0 AND sysBP ≥ 0.32	$\rightarrow 1$
0.062	diabetes=0 AND prevalentHyp=1	$\rightarrow 1$

METHOD

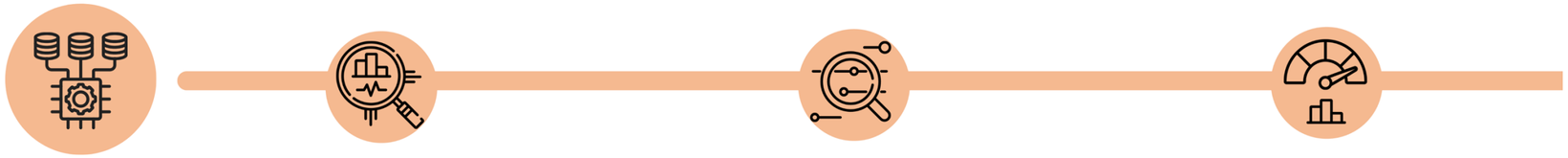


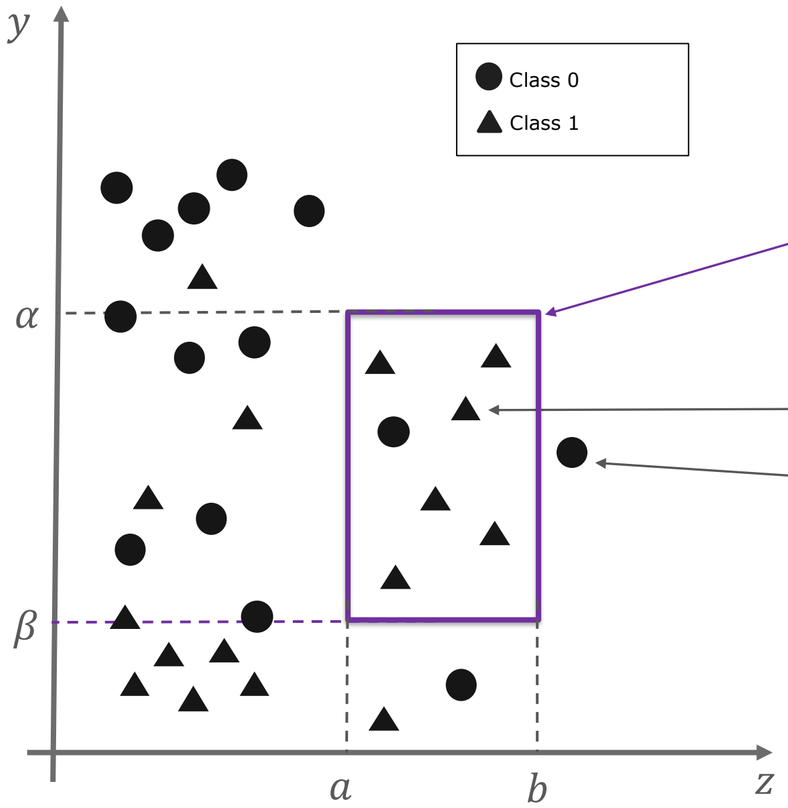


$\bar{\phi}_x(R)$	R	c
0.2174	prevalentHyp=1 AND sysBP $\geq$ 0.32	$\rightarrow$ 1
0.1473	sysBP $\geq$ 0.32	$\rightarrow$ 1
0.0828	male=1	$\rightarrow$ 1
0.0701	prevalentHyp=1	$\rightarrow$ 1

METHOD

GENERATION OF GLOBAL RULES | **EXTRACTION OF LOCAL RULES** | **EVALUATION OF GLOBAL RULES**





subgroup :
$z \in [a, b]\ and\ y \in [\alpha, \beta] \rightarrow 1$

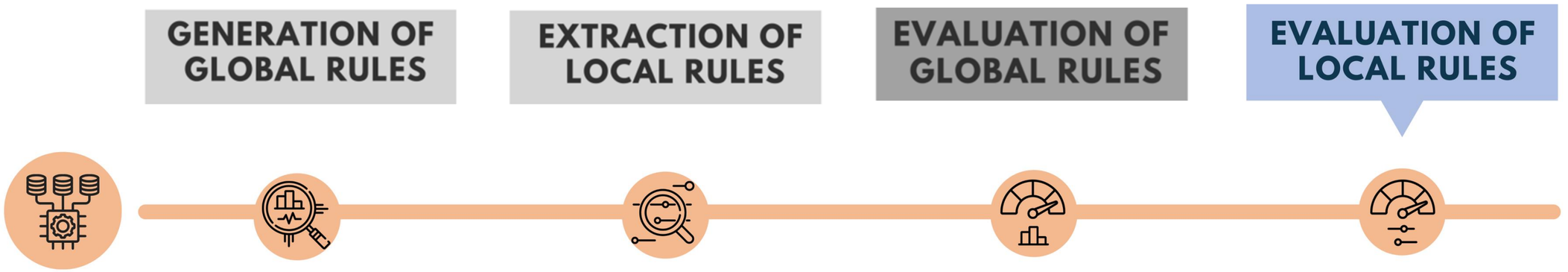
Target class

Negative (non-target) classes

$$lift\ (R) = \left(\frac{\#}{\#} \div \frac{\#}{\#}\right)$$

$$WRAcc\ (R) = \frac{\#}{\#}\left(\frac{\#}{\#} - \frac{\#}{\#}\right)$$

METHOD



Fidelity : Fidelity evaluates the consistency between the explanation-based prediction and the black-box model. It is defined as the proportion of instances for which the explanation recovers the model's original output :

$$Fidelity = \frac{1}{|D|} \sum_{x \in D} \mathbf{1}_{\{\hat{y}_{expl}(x) = f(x)\}}.$$

Accuracy : While fidelity measures alignment with the model, accuracy assesses the faithfulness of explanations with respect to the ground-truth labels :

$$Accuracy = \frac{1}{|D|} \sum_{x \in D} \mathbf{1}_{\{\hat{y}_{expl}(x) = y(x)\}}.$$

Completeness : Completeness quantifies the proportion of instances for which at least one explanatory rule is available, thereby measuring the coverage of the rule set over the dataset :

$$Completeness = \frac{|\{x \in D \mid \mathcal{R}_{pos}(x) \neq \emptyset\}|}{|D|}.$$

Consistency : Consistency measures, among instances covered by at least one rule, the proportion for which all explanatory rules agree on the same predicted class :

$$Consistency = \frac{|\{x \in D \mid \mathcal{R}_{pos}(x) \neq \emptyset \ \wedge \ \exists c \in C \ \ \forall R \in \mathcal{R}_{pos}(x), \ c(R) = c\}|}{|\{x \in D \mid \mathcal{R}_{pos}(x) \neq \emptyset\}|}.$$

EXPERIMENTS

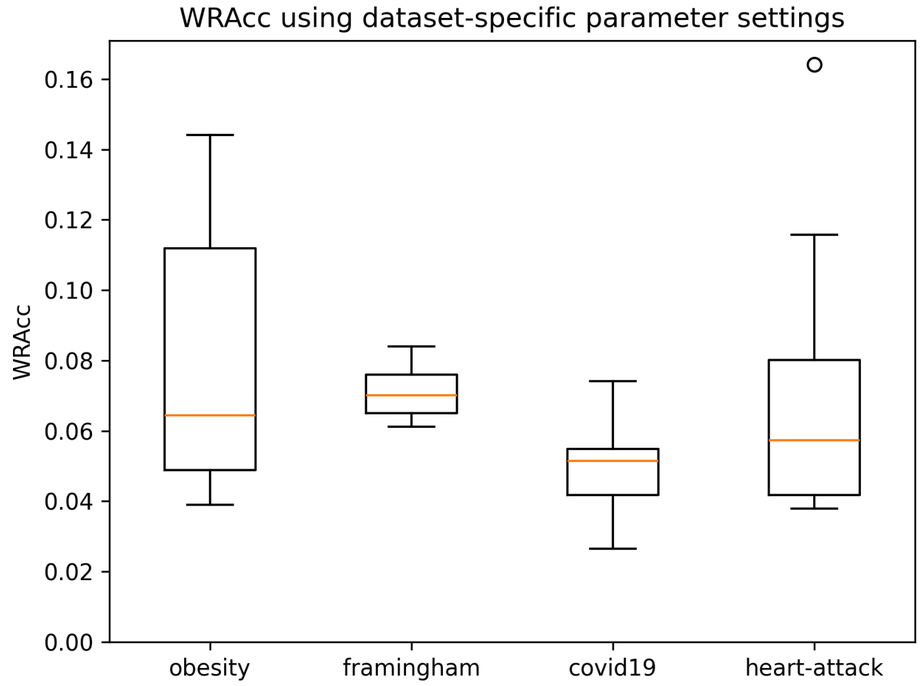
Dataset

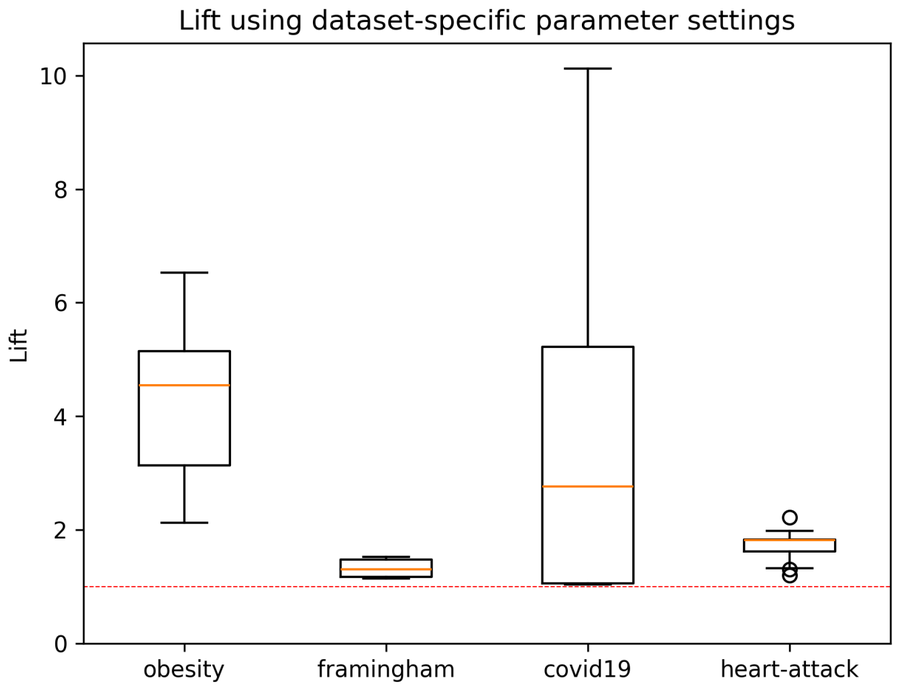
Dataset	|Classes|	|Features|	|Instances|	Model	Model Accuracy
Framingham	2	15	3,658	Random Forest	0.9758
Heart-attack	2	8	2,111	Decision Tree	0.9924
Covid19	2	19	1,048,575	Logic Regression	0.9384
Obesity	7	15	2,111	MultiLayer Perceptron	0.8511

Results

after selecting the setting

Dataset	depth	result_set_size	Fidelity	Accuracy	Completeness	Consistency
Framingham	2	10	0.9	0.88	1	0.39
Heart-attack	2	10	0.96	0.95	1	0.65
Covid19	2	20	0.92	0.88	1	0.79
Obesity	3	10	0.76	0.69	0.92	0.42





Discussion

Limitations

- Local explanations may contain conflicting rules
- Incomplete coverage
- Redundancy from overlapping rules
- Restricted to classification on tabular data

Perspective

- Develop conflict-aware rule extraction mechanisms
- Diversity-based pruning for compact explanations
- Extension to regression tasks
- Generalization to non-tabular modalities

Merci de votre attention