Amélioration de l'interprétabilité des explications de *SHAP* grâce à la découverte de sous-groupes

Résumé. L'intégration de modèles prédictifs en médecine nécessite des explications compréhensibles pour soutenir la décision clinique. Cette étude préliminaire propose une approche post hoc, model-agnostic combinant *SHAP* et la *découverte de sous-groupes* afin de générer des règles explicites de type *SI–ALORS*. Cette combinaison permet de produire simultanément des explications locales et globales, tout en offrant des explications plus précise que les facteur d'importance de *SHAP* et une compréhension plus riche des interactions entre variables. Les expériences réalisées sur quatre jeux de données médicaux montrent une large couverture et une caractérisation précise des classes, avec des valeurs élevées de *WRAcc* et de *lift*. Les explications locales obtenues présentent une fidélité supérieure à 90% pour les modèles binaires. Bien que développée dans un contexte médical, l'approche peut être appliquée à tout domaine nécessitant intelligibilité et confiance dans les modèles prédictifs.

1 Introduction

L'utilisation croissante de l'intelligence artificielle en médecine s'accompagne d'un besoin essentiel : disposer d'outils capables de fournir des informations compréhensibles et exploitables par les cliniciens. L'intelligibilité constitue un enjeu central pour analyser les données, interpréter des relations complexes et soutenir la décision médicale. C'est dans cette perspective que différentes approches cherchent à rendre visibles les mécanismes sous-jacents aux données ou aux modèles prédictifs.

Les modèles interprétables, issus de l'*exploration de données*, produisent des modèles directement lisibles, capables de faire émerger des motifs locaux et des relations entre variables, avec pour objectif d'extraire de nouvelles connaissances non triviales (Aggarwal, 2015). Elles s'avèrent pertinentes pour prendre en compte la forte variabilité inter-individuelle des données médicales (Moranges et al., 2021).

Les méthodes d'*intelligence artificielle explicable (XAI)* visent à éclairer le fonctionnement de modèles prédictifs complexes, dits « boîte noire ». Parmi elles, SHAP (SHapley Additive exPlanations) (Lundberg et Lee, 2017) est devenue la référence en médecine (Caterson et al., 2024), grâce à sa capacité à fournir des explications locales et globales, et à son indépendance au modèle prédictif utilisé (*model-agnostic*). Néanmoins, ces explications globales reposent principalement sur des valeurs moyennes d'importance, qui ne capturent ni les interactions entre variables, ni la spécificité de certains profils de patients. Elles peuvent ainsi masquer des

relations cliniques multifactorielles et réduire la richesse des phénomènes biologiques sous-jacents.

Face à ces limites, notre objectif est de rendre explicables des modèles prédictifs performants, en développant une approche *post hoc*, *model-agnostic* et compatible avec le raisonnement clinique. Pour cela, nous proposons de combiner les explications *SHAP* avec la *découverte de sous-groupes (DS)* (Wrobel, 1997), une technique issue de l'*exploration de données* permettant de générer des règles explicites de type *SI–ALORS*, proches du raisonnement clinique habituel. D'un côté, ces règles permettent de relier plusieurs variables entre elles. D'un autre côté, elles facilitent la discussion, la validation et l'appropriation des modèles par les experts du domaine.

Nous entrevoyons trois points de distinction par rapport à l'existant. Premièrement, la DS permet d'identifier des sous-groupes cohérents de patients, révélant des interactions multidimensionnelles et offrant des explications plus spécifiques et fines que les approches globales fondées sur des moyennes. Deuxièmement, l'intégration directe des valeurs de SHAP pour produire des règles qui reflètent fidèlement le comportement interne du modèle. Enfin, contrairement aux méthodes existantes basées sur des règles (Ribeiro et al., 2018; Guidotti et al., 2019; Yuan et al., 2022), notre approche fournit simultanément des explications globales et locales, un point essentiel en médecine : les explications globales permettent aux médecins de comprendre et de valider le fonctionnement général du modèle, condition indispensable pour instaurer la confiance et favoriser son adoption clinique, tandis que les explications locales offrent une justification précise et individualisée de chaque prédiction, facilitant la prise de décision et la communication avec le patient.

Ainsi, la combinaison de SHAP et de la DS constituerait une couche d'explicabilité contextuelle, complémentaire aux explications classiques de SHAP, et proposerait une manière nouvelle et exploitable de rendre les modèles complexes plus transparents et plus exploitables pour la décision médicale.

2 Méthode d'extraction de règles à partir de *SHAP*

2.1 Cadre classique de *SHAP*

Pour extraire des explications sous forme de règles, nous nous appuyons sur le cadre *SHAP* introduit par Lundberg et Lee (2017), fondé sur les valeurs de Shapley issues de la théorie des jeux coopératifs. *SHAP* fournit une méthode additive d'attribution des contributions des variables, satisfaisant trois propriétés clés : l'exactitude locale, la nullité et la cohérence.

Étant donné un modèle de prédiction f et une instance x, *SHAP* cherche à expliquer la prédiction $f(x)$ à l'aide d'un modèle explicatif simplifié g, défini par :

$$g(z) = \phi_0 + \sum_{j=1}^{M} \phi_j z_j,$$

où chaque z_j indique si la variable j est prise en compte dans l'explication ($z_j = 1$ si la variable est présente, $z_j = 0$ sinon), et où ϕ_j représente la contribution associée à cette variable. Les valeurs ϕ_j quantifient ainsi l'effet individuel de chaque caractéristique sur la prédiction du modèle.

Soit $\phi_{x,j}$ la contribution de la variable j à la prédiction $f(x)$. Pour une instance x, la valeur $\phi_{x,j}$ correspond à la valeur de Shapley classique :

$$\phi_{x,j} = \sum_{S \subseteq \mathcal{F} \setminus \{j\}} \frac{|S|!(M - |S| - 1)!}{M!} \left[f_x(S \cup \{j\}) - f_x(S) \right],$$

où $\mathcal{F} = \{1, \ldots, M\}$ désigne l'ensemble des variables, et S l'ensemble des indices non nuls de z, c'est-à-dire les variables considérées comme présentes dans l'entrée simplifiée. La fonction $f_x(S)$ représente la prédiction moyenne du modèle, lorsque seules les variables du sous-ensemble S sont fixées.

Les valeurs *SHAP* offrent une approche unifiée et théoriquement fondée pour attribuer une importance aux variables dans les prédictions individuelles, en accord les principes d'équité issus de la théorie des jeux. En pratique, le calcul de $f_x(S)$ doit être approximée et plusieurs stratégies sont posssibles. La bibliothèque *SHAP* en Python propose par exemple les approches « *interventional* » (Lundberg et Lee, 2017) et « *tree path dependent* » (Lundberg et al., 2018). Ces stratégies influencent la manière dont les variables n'appartenant pas à S sont marginalisées ou échantillonnées, selon que les dépendances entre variables sont préservées ou rompues.

SHAP a la particularité de fournir à la fois des explications locales et globales (Hermosilla et al., 2025; Vimbi et al., 2024), un atout essentiel dans les contextes cliniques. Cependant, les sorties de *SHAP* peuvent être difficiles à interpréter correctement pour les cliniciens ne disposant pas d'une formation en apprentissage automatique (Al-Absi et al., 2024; Alkhanbouli et al., 2025). Chaque variable se voit attribuer une valeur – positive, négative ou nulle – indiquant sa contribution par rapport à une prédiction de référence. Sans contexte supplémentaire, cette information peut sembler abstraite pour les non-spécialistes (Salih et al., 2025). Par exemple, une valeur *SHAP* de $+0.8$ pour « hypertension » suggère une augmentation du risque prédit, sans préciser l'ampleur de l'effet ni le mécanisme biologique sous-jacent. De plus, les moyennes globales de *SHAP* peuvent masquer des effets spécifiques à certains sous-groupes : des variables déterminantes pour une sous-population peuvent être moins importantes pour d'autres (Lundberg et al., 2020). Enfin, bien que de nombreux facteurs de risque cliniques interagissent, *SHAP* fournit peu d'informations sur leurs dépendances. Les valeurs d'interaction ne sont disponibles que pour les modèles arborescents et ne capturent que des effets entre 2 facteurs, négligeant les interactions d'ordre supérieur et risquant ainsi de simplifier des mécanismes complexes (Lundberg et al., 2020).

Pour pallier ces limites de *SHAP*, nous proposons de combiner *SHAP* avec la *DS*, afin d'extraire des règles interprétables. Ces règles visent à fournir des explications globales plus précises, contextualisées et capables de capturer des interactions complexes entre facteurs.

2.2 Découverte de sous-groupes

La *découverte de sous-groupes* est une tâche d'exploration de données visant à identifier des sous-ensembles d'une population dont la distribution, par rapport à une variable cible, diffère significativement de celle observée dans l'ensemble du jeu de données. L'objectif est de révéler des motifs locaux à la fois statistiquement pertinents et interprétables, permettant de mieux comprendre les conditions dans lesquelles la classe cible se manifeste avec une fréquence atypique.

Formellement, soit D l'ensemble de données analysé. Dans un contexte de classification avec un espace de sortie C, soit $c \in C$ la classe cible d'intérêt. On note $D^{(c)}$ l'ensemble des instances appartenant à la classe c.

Un sous-groupe est défini par une règle R correspondant à une conjonction de conditions sur les variables :

$$R = \{(j_k, \circ_k, v_k) \mid k \in \mathcal{F}\}$$

où chaque triplet spécifie une condition sur la variable j_k à l'aide d'un opérateur relationnel $\circ_k$ (tel que $<$, $>$ ou $=$) et d'une valeur seuil v_k. L'ensemble des instances du jeu de données satisfaisant la règle R est noté $\text{cov}(R)$. Parmi ces instances couvertes, on distingue $\text{cov}_\text{c}(R) = \text{cov}(R) \cap D^{(c)}$, le sous-ensemble des instances appartenant à la classe cible c.

Pour évaluer la qualité de tels sous-groupes, la *WRAcc* (*Weighted Relative Accuracy*) (Lavrač et al., 2004) est définie comme suit :

$$\text{WRAcc}(R) = \frac{|\text{cov}(R)|}{|D|} \left(\frac{|\text{cov}_\text{c}(R)|}{|\text{cov}(R)|} - \frac{|D^{(c)}|}{|D|} \right)$$

Cette mesure quantifie la différence entre la proportion d'instances observée de la classe cible au sein du sous-groupe et sa proportion attendue dans l'ensemble du jeu de données. Cette différence est ensuite pondérée par le support du sous-groupe, c'est-à-dire sa taille relative par rapport au jeu de données, favorisant ainsi l'identification de sous-groupes significatifs et représentatifs.

2.3 Une mesure d'intérêt à partir des valeurs *SHAP*

Dans ce travail, nous proposons d'étendre la mesure classique de *WRAcc* en la pondérant par les valeurs de Shapley au niveau des instances.

Tout d'abord, nous normalisons les valeurs de *SHAP*. Ces valeurs n'étant pas bornées dans un intervalle fixe, et leur amplitude peut varier selon l'échelle des sorties du modèle. Pour cette raison, nous les normalisons dans l'intervalle $[0, 1]$ ainsi :

$$\widetilde{\phi}_{x,j} = \frac{\phi_{x,j} - \phi_{\min}}{\phi_{\max} - \phi_{\min}},$$

où $\phi_{\min} = \min_{a \in D,\ b \in \mathcal{F}} \phi_{a,b}$ et $\phi_{\max} = \max_{a \in D,\ b \in \mathcal{F}} \phi_{a,b}$.

Pour chaque règle R, nous définissons l'importance d'une instance x comme suit :

$$w_x^R = \sum_{j \in \mathcal{F}_R} \widetilde{\phi}_{x,j}$$

où $\mathcal{F}_R$ l'ensemble des variables impliquées dans la règle R. w_x^R agrège les contributions de toutes les variables utilisées dans la règle, reflétant ainsi le degré selon lequel l'instance x participe au sous-groupe capturé par R.

Pour tout sous-ensemble $S \subseteq D$, nous définissons son poids total de Shapley comme suit :

$$W(S) = \sum_{x \in S} w_x^R.$$

En nous appuyant sur cette notation, nous étendons la mesure classique *WRAcc* afin d'y intégrer les pondérations d'instances basées sur les valeurs de Shapley. La mesure obtenue, que nous appelons *WRAcc pondérée par les valeurs de Shapley*, est définie comme suit :

$$\text{WRAcc}_{\phi}(R) \;=\; \frac{W\left(\text{cov}(R)\right)}{W\left(D\right)} \cdot \left(\frac{W\left(\text{cov}_{\text{c}}(R)\right)}{W\left(\text{cov}(R)\right)} - \frac{W\left(D^{(c)}\right)}{W\left(D\right)} \right)$$

Cette version pondérée évalue si un sous-groupe concentre une part disproportionnée de l'importance explicative dérivée des valeurs de Shapley pour la classe cible, par rapport à l'ensemble du jeu de données. Ce faisant, elle adapte la mesure *WRAcc* afin de mettre en évidence les sous-groupes qui ne sont pas seulement statistiquement significatifs, mais également étroitement alignés sur les schémas d'attribution internes du modèle. Pour une illustration détaillée de cette mesure, le lecteur est invité à consulter la Figure A1 des Annexes.

On peut noter ici l'importance de la normalisation des valeurs de *SHAP*. Des pondérations négatives inverseraient l'effet attendu des contributions des variables et conduiraient à une évaluation erronée des sous-groupes. En ramenant les valeurs dans l'intervalle $[0, 1]$, on améliore ainsi l'interprétabilité de la mesure pondérée.

2.4 Algorithme d'extraction de règles *SHAP*

Nous appliquons la *DS* en prenant, pour chaque instance x, la classe prédite $f(x)$ comme variable cible c. Pour orienter la sélection de ces règles, nous exploitons les valeurs *SHAP* et les intégrons directement dans le calcul de la $\text{WRAcc}_{\phi}(R)$. Les sous-groupes maximisant cette mesure de qualité sont ensuite recherchés au moyen d'une *Beam Search*, un algorithme heuristique qui explore l'espace des règles de manière progressive en ne conservant, à chaque itération, que les candidats les plus prometteurs. Cette stratégie permet d'identifier efficacement des règles cohérentes avec les contributions locales du modèle tout en limitant le coût computationnel.

Le nombre de règles à générer constitue un paramètre d'entrée de l'algorithme (*result_set_size*), qui sélectionne les règles les mieux classées selon leurs scores de *WRAcc* pondérés par Shapley. Un autre paramètre fixe le nombre maximal de conditions par règle (*depth*). Ainsi l'algorithme repose sur un deux paramètres intuitifs, le rendant accessible aux non-spécialistes tout en leur permettant de contrôler le niveau de complexité des règles et d'obtenir des explications à la fois pertinentes et interprétables. Un exemple d'explications globales est donné Fig.A2.b des Annexes.

2.5 Génération de règles locales

Nous cherchons à fournir des explications au niveau individuel en identifiant, pour chaque prédiction, lesquelles des règles globalement extraites sont activées par une instance donnée. Étant donnée une instance $x \in D$ à expliquer, cette procédure s'effectue en trois étapes :

1. **Filtrage des règles basé sur la couverture.** Nous sélectionnons d'abord l'ensemble des règles qui couvrent x :

$$\mathcal{R}_{\text{cov}}(x) = \{R \mid x \in \text{cov}(R)\}\,.$$

2. **Filtrage des contributions *SHAP* positives.** Parmi cet ensemble, nous ne conservons que les règles pour lesquelles toutes les variables de $\mathcal{F}_R$ présentent des valeurs de Shapley positives pour l'instance x :

$$\mathcal{R}_{\text{pos}}(x) = \{R \in \mathcal{R}_{\text{cov}}(x) \mid \phi_{x,j} > 0 \ \ \forall j \in \mathcal{F}_R\}\,.$$

3. **Classement des règles selon la valeur moyenne de *SHAP*.** Enfin, les règles sont classées selon leur valeur moyenne de Shapley normalisée pour les variables impliquées dans la règle. Pour une règle $R \in \mathcal{R}_{\text{pos}}(x)$, nous définissons :

$$\overline{\phi}_x(R) = \frac{1}{|\mathcal{F}_R|} \sum_{j \in \mathcal{F}_R} \widetilde{\phi}_{x,j}.$$

Les règles sont ensuite triées par ordre décroissant de $\overline{\phi}_x(R)$ afin d'être présentées comme explications.

Un exemple d'explications locales est donné Fig.A2.d des Annexes. Le fait que les explications locales soient directement issues des règles globales garantit l'adéquation entre ces deux niveaux d'interprétation.

3 Expériences

3.1 Métriques d'évaluation

Pour évaluer le cadre proposé, nous nous appuyons sur un ensemble de métriques complémentaires permettant de mesurer à la fois la qualité des explications et les propriétés des règles extraites. Ces métriques sont définies à deux niveaux : au niveau de l'instance, pour l'évaluation des explications locales, et au niveau de la règle, pour évaluer l'ensemble des règles globales.

Pour chaque instance $x \in D$, l'explication locale est donnée par l'ensemble des règles activées $\mathcal{R}\text{pos}(x)$. Ces règles produisent collectivement une prédiction fondée sur les explications, notée $\hat{y}\text{expl}(x)$, obtenue par un vote majoritaire pondéré :

$$\hat{y}_{\text{expl}}(x) = \arg\max_{c \in C} \sum_{R \in \mathcal{R}_{\text{pos}}(x)} \mathbb{1}_{\{c(R)=c\}}\, \overline{\phi}_x(R),$$

où $c(R)$ désigne la classe prédite par la règle R.

En cas d'égalité, la règle présentant le plus faible classement selon $\overline{\phi}_x(R)$ est retirée itérativement jusqu'à l'obtention d'une classe unique. Cette procédure garantit que les prédictions fondées sur les explications demeurent à la fois déterministes et cohérentes avec l'ordre induit par les valeurs *SHAP*.

Nous cherchons à évaluer la cohérence entre la prédiction fondée sur les explications et celle du modèle « boîte noire ». À cette fin, nous calculons la *fidélité*, définie comme la proportion d'instances pour lesquelles l'explication reproduit la sortie initiale du modèle :

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

Nous évaluons également la fidélité des explications par rapport aux étiquettes réelles au moyen de la métrique de *précision (accuracy)* :

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

Nous examinons également la généralisabilité des règles globales à l'ensemble du jeu de données en mesurant leur couverture. Celle-ci est évaluée par la *complétude*, qui quantifie la proportion d'instances pour lesquelles au moins une règle explicative est disponible.

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

Afin de garantir que les explications locales demeurent compréhensibles pour les utilisateurs, nous vérifions si l'ensemble des règles locales associées à chaque instance prédit une classe cohérente. À cette fin, nous utilisons la *cohérence*, qui mesure, parmi les instances couvertes par au moins une règle, la proportion de celles pour lesquelles toutes les règles explicatives s'accordent sur la même classe prédite :

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

Au niveau des règles, nous évaluons chaque sous-groupe individuellement. Nous rapportons d'abord sa valeur de *WRAcc*, déjà définie dans la Section 2.2. En complément de la *WRAcc*, nous utilisons également le *lift* afin de mieux caractériser la qualité des sous-groupes identifiés. Alors que la *WRAcc* évalue l'exceptionnalité d'un sous-groupe relativement à l'ensemble des données, le *lift* mesure l'augmentation relative de la fréquence de la classe cible au sein du sous-groupe par rapport à sa fréquence dans l'ensemble du jeu de données. Formellement, le *lift* d'une règle R est défini comme suit :

$$\text{lift}(R) = \frac{|\text{cov}_{\text{c}}(R)|}{|\text{cov}(R)|} \Big/ \frac{|D^{(c)}|}{|D|}.$$

Une valeur de *lift* supérieure à 1 indique que le sous-groupe est plus enrichi en instances de la classe cible que ce qui serait attendu par hasard.

3.2 Protocole expérimental

Notre évaluation est menée sur quatre jeux de données médicaux : *Framingham*, *Heart-attack*, *Covid19* et *Obesity*, résumés dans le Tableau 1. Chaque jeu de données est associé à un modèle prédictif distinct, entraîné pour atteindre une haute précision de prédiction. Le choix de jeux de données et de modèles hétérogènes vise à démontrer la généralité de l'approche proposée à travers différentes distributions de données et divers paradigmes de modélisation. Les valeurs *SHAP* ont été obtenues à l'aide de la bibliothèque `SHAP`, et nous avons réalisé une extension de la librairie `pysubgroup` en y intégrant la mesure WRAcc_{ϕ} pour réaliser la *DS*.

Jeux de données	|Classes|	|Caractéristiques|	|Instances|	Modèle	Précision du modèle
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

TAB. 1 – *Résumé des jeux de données utilisés pour l'évaluation, incluant le nombre de classes, de variables et d'instances, ainsi que les modèles prédictifs employés et leurs précisions respectives.*

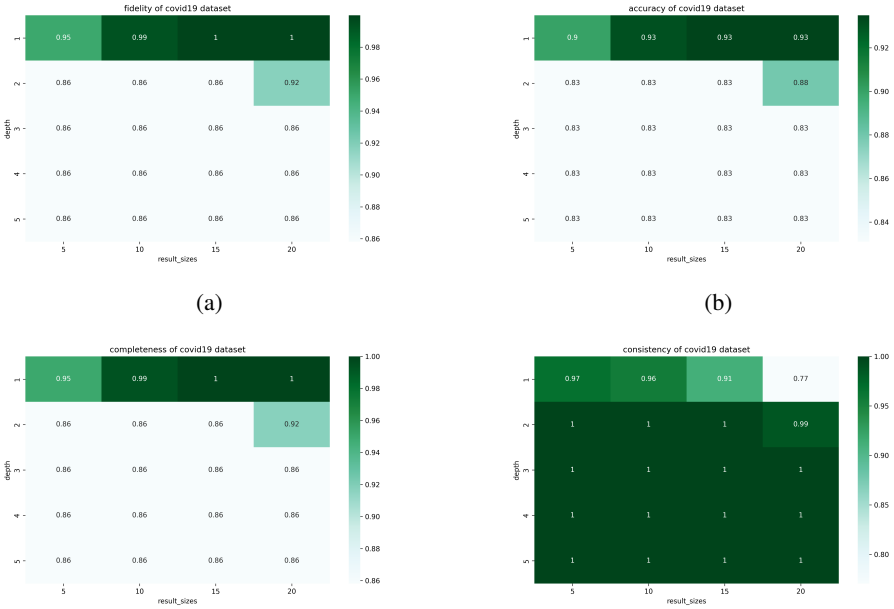
3.3 Résultats

Pour chaque jeu de données, nous avons calculé les métriques d'évaluation introduites dans la Section 3.1 selon différentes paramétrisations de l'algorithme de *DS*. Plus précisément, la profondeur maximale des règles (*depth*) a été ajustée de 1 à 5, et le nombre de règles générées par classe cible (*result_set_size*) variait de 5 à 20, par incréments de 5. Les résultats de ces expériences, illustrant l'évolution de la précision, de la fidélité, de la complétude et de la cohérence selon les différents paramètres, sont présentés pour le jeu de données *Covid19* dans la Fig. 1. Les résultats correspondants pour les jeux de données *Framingham*, *Heart-attack* et *Obesity* sont fournis dans le fichier Annexes (Fig. A3–A5).

Tendances générales. La précision et la fidélité demeurent constamment élevées, dépassant 0.8 pour l'ensemble des classifieurs binaires, quelle que soit la paramétrisation ou l'architecture du modèle. Cela confirme que les règles locales reproduisent fidèlement les prédictions du modèle (fidélité) tout en restant proches des étiquettes réelles (précision). La fidélité est généralement plus élevée et plus stable que la précision, indiquant que les explications capturent la logique interne du modèle avec une remarquable cohérence. La complétude est également élevée, toujours supérieure à 0.8, montrant que les sous-groupes découverts couvrent une large part du jeu de données et semblent bien se généraliser. En revanche, la cohérence varie davantage selon les jeux de données et les configurations de paramètres. Certaines instances restent couvertes par des règles contradictoires, ce qui suggère que la procédure de filtrage fondée sur *SHAP* atténue, sans toutefois éliminer totalement, les conflits entre règles.

Impact du nombre de règles générées. L'augmentation du nombre de règles par classe a un effet direct sur la complétude, qui croît logiquement à mesure qu'un plus grand nombre de règles couvre une part plus importante des données. La précision et la fidélité s'améliorent également avec le nombre de règles avant de se stabiliser, sans jamais diminuer. Cela indique que, même si des règles supplémentaires peuvent introduire une certaine redondance, elles ne dégradent pas la qualité explicative globale. Cependant, un ensemble de règles plus large tend à réduire la cohérence : à mesure que le nombre de règles augmente, la probabilité de contradictions entre elles s'accroît également. En résumé, la génération d'un plus grand nombre de règles renforce la couverture et la fidélité, mais au prix d'une cohérence interne réduite.

Impact de la profondeur des règles. La profondeur maximale autorisée des règles a des effets différenciés sur les métriques d'évaluation. Des règles plus profondes tendent à être plus spécifiques, ce qui réduit les contradictions et améliore ainsi la cohérence. Cependant, la



(a) (b)

(c) (d)

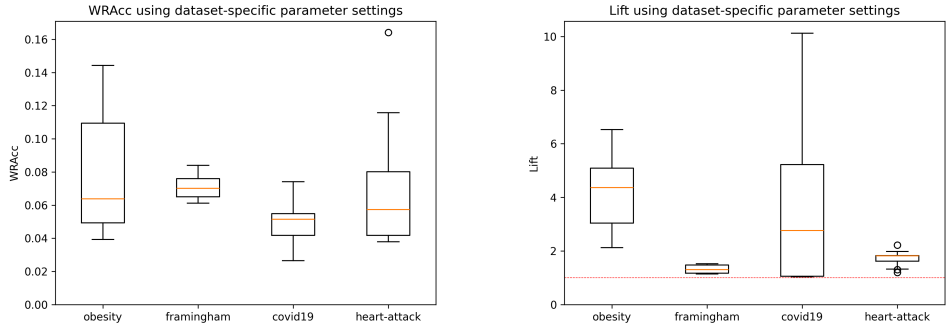
FIG. 1 – *Métriques d'évaluation pour le jeu de données Covid19 selon différents réglages de paramètres : (a) fidélité, (b) précision, (c) complétude et (d) cohérence.*

fidélité diminue généralement avec la profondeur, car des règles très spécifiques s'accordent moins bien avec les frontières de décision globales du modèle (à l'exception des classifieurs arborescents, pour lesquels une profondeur élevée reste cohérente avec la structure du modèle). L'exactitude suit une tendance similaire à celle de la fidélité, les règles trop spécialisées ayant une capacité de généralisation plus faible. Fait intéressant, les règles de profondeur égal à 1 offrent la meilleure précision et fidélité, ce qui s'explique par le fait que les valeurs *SHAP* sont calculées au niveau des variables individuelles et non des interactions entre variables. Dans l'ensemble, des règles excessivement profondes améliorent la cohérence interne mais réduisent la capacité des explications à reproduire fidèlement le comportement du modèle.

Stratégie de paramétrisation. Les résultats expérimentaux permettent de dégager plusieurs recommandations pour le choix des paramètres. Premièrement, le nombre de règles par classe doit être suffisamment élevé pour garantir une bonne couverture et une fidélité stable. Comme la précision et la fidélité atteignent un plateau sans diminuer, un ensemble de règles plus large peut être privilégié, à condition que la baisse de cohérence reste acceptable pour l'usage visé. Deuxièmement, la profondeur maximale des règles doit rester modérée. Des règles trop peu profondes peuvent engendrer des contradictions, tandis que des règles trop profondes dégradent la complétude, la fidélité et l'exactitude. Le compromis le plus efficace se situe à une profondeur intermédiaire, où la cohérence est déjà satisfaisante sans entraîner de perte substantielle de fidélité.

Dataset	depth	result_set_size	Fidelity	Accuracy	Completeness	Consistency
Framingham	2	10	0.9	0.88	0.97	0.8
Heart-attack	2	10	0.96	0.95	1	0.98
Covid19	2	20	0.92	0.88	0.92	0.99
Obesity	3	10	0.76	0.69	0.87	0.68

TAB. 2 – *Paramètres optimaux (profondeur maximale des règles et nombre de règles par classe) et métriques d'évaluation correspondantes (fidélité, exactitude, complétude et cohérence) pour chaque jeu de données.*



(a) (b)

FIG. 2 – *Boxplots of rule-level quality measures obtained with the selected parameters for each dataset : (a) Classic WRAcc and (b) Lift.*

Les choix de paramètres effectués pour chaque jeu de données, ainsi que les valeurs obtenues pour les différentes métriques, sont présentés dans le Tableau 2. Ces paramètres ont été sélectionnés afin de maximiser un compromis entre cohérence et fidélité, une forte fidélité sans cohérence conduisant à des explications peu interprétables. En complément des métriques d'évaluation globales, nous évaluons également la qualité propre des règles extraites. Deux mesures complémentaires sont mobilisées à cet effet : la *WRAcc* classique et le *Lift*. Les distributions de *WRAcc* et de *Lift* pour l'ensemble des règles, calculées selon les paramètres retenus pour chaque jeu de données, sont illustrées dans la Figure 2. Comme les valeurs de *WRAcc* sont systématiquement positives et que les valeurs de *Lift* demeurent strictement supérieures à 1, ces résultats démontrent que chaque règle extraite est bien spécifique à sa classe cible et qu'elle capture une déviation significative par rapport à la distribution globale.

4 Limites et travaux futurs

Nous identifions plusieurs limites à l'approche proposée. Premièrement, la méthode actuelle est restreinte aux tâches de classification, car elle repose sur la distribution de classes cibles discrètes de la *DS*. Deuxièmement, les explications locales peuvent inclure des règles prédisant des classes différentes, notamment lorsque la profondeur maximale des règles est

fixée à 1. De telles règles conflictuelles peuvent nuire à l'interprétabilité et compliquer la compréhension du raisonnement décisionnel du modèle. Troisièmement, les explications locales ne sont disponibles que pour les instances satisfaisant au moins une règle extraite et présentant des valeurs *SHAP* positives. Les instances non couvertes demeurent donc inexpliquées, ce qui peut limiter l'utilité de la méthode dans des contextes à forte dimensionnalité où la couverture des sous-groupes est faible. L'augmentation du nombre de règles générées peut atténuer ce problème, mais au prix d'une redondance accrue. La redondance constitue elle-même une autre limite : la *DS* peut produire des règles chevauchantes ou imbriquées – par exemple « $X > 15$ » et « $X > 15 \wedge Y = \text{True}$ » – qui réduisent la clarté sans apporter d'information supplémentaire. Des travaux futurs devraient ainsi explorer des stratégies d'élagage et de sélection basées sur la diversité afin de maintenir un ensemble de règles compact et non redondant. Enfin, le cadre est applicable seulement sur des données tabulaires. L'étendre à d'autres modalités de données, telles que les séries temporelles, les images ou les graphes, élargirait considérablement son champ d'application.

5 Conclusion

Ce travail propose une approche préliminaire combinant *DS* et *SHAP* de produire des règles explicatives cohérentes avec le comportement interne d'un modèle prédictif. La méthode fournit une intelligibilité plus contextualisée que les approches fondées sur les facteurs d'importances, et se rapproche du raisonnement clinique grâce à des règles explicites et discutables. Cette contribution ouvre des perspectives pour une explicabilité mieux alignée avec les besoins des praticiens et offre un cadre flexible, applicable à tout modèle prédictif et à tout domaine où transparence et compréhension sont essentielles.

Références

Aggarwal, C. C. (2015). An introduction to data mining. In *Data mining : The textbook*, pp. 1–26. Springer.

Al-Absi, D. T., M. C. E. Simsekler, M. A. Omar, et S. Anwar (2024). Exploring the role of artificial intelligence in acute kidney injury management : a comprehensive review and future research agenda. *BMC Medical Informatics and Decision Making 24*(1), 337.

Alkhanbouli, R., H. Matar Abdulla Almadhaani, F. Alhosani, et M. C. E. Simsekler (2025). The role of explainable artificial intelligence in disease prediction : a systematic literature review and future research directions. *BMC medical informatics and decision making 25*(1), 110.

Caterson, J., A. Lewin, et E. Williamson (2024). The application of explainable artificial intelligence (xai) in electronic health record research : A scoping review. *Digital health 10*, 20552076241272657.

Guidotti, R., A. Monreale, F. Giannotti, D. Pedreschi, S. Ruggieri, et F. Turini (2019). Factual and counterfactual explanations for black box decision making. *IEEE Intelligent Systems 34*(6), 14–23.

Hermosilla, P., S. Berríos, et H. Allende-Cid (2025). Explainable ai for forensic analysis : A comparative study of shap and lime in intrusion detection models. *Applied Sciences 15*(13), 7329.

Lavrač, N., B. Kavšek, P. Flach, et L. Todorovski (2004). Subgroup discovery with cn2-sd. *Journal of Machine Learning Research 5*(Feb), 153–188.

Lundberg, S. M., G. Erion, H. Chen, A. DeGrave, J. M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal, et S.-I. Lee (2020). From local explanations to global understanding with explainable ai for trees. *Nature machine intelligence 2*(1), 56–67.

Lundberg, S. M., G. G. Erion, et S.-I. Lee (2018). Consistent individualized feature attribution for tree ensembles. *arXiv preprint arXiv :1802.03888*.

Lundberg, S. M. et S.-I. Lee (2017). A unified approach to interpreting model predictions. *Advances in neural information processing systems 30*.

Moranges, M., C. Rouby, M. Plantevit, et M. Bensafi (2021). Explicit and implicit measures of emotions : Data-science might help to account for data complexity and heterogeneity. *Food Quality and Preference 92*, 104181.

Ribeiro, M. T., S. Singh, et C. Guestrin (2018). Anchors : High-precision model-agnostic explanations. In *Proceedings of the AAAI conference on artificial intelligence*, Volume 32.

Salih, A. M., Z. Raisi-Estabragh, I. B. Galazzo, P. Radeva, S. E. Petersen, K. Lekadir, et G. Menegaz (2025). A perspective on explainable artificial intelligence methods : Shap and lime. *Advanced Intelligent Systems 7*(1), 2400304.

Vimbi, V., N. Shaffi, et M. Mahmud (2024). Interpreting artificial intelligence models : a systematic review on the application of lime and shap in alzheimer's disease detection. *Brain Informatics 11*(1), 10.

Wrobel, S. (1997). An algorithm for multi-relational discovery of subgroups. In *European symposium on principles of data mining and knowledge discovery*, pp. 78–87. Springer.

Yuan, J., B. Barr, K. Overton, et E. Bertini (2022). Visual exploration of machine learning model behavior with hierarchical surrogate rule sets. *IEEE Transactions on Visualization and Computer Graphics 30*(2), 1470–1488.

Summary

The integration of predictive models in medicine requires understandable explanations to support clinical decision-making. This preliminary study introduces a post-hoc, model-agnostic approach that combines *SHAP* with *DS* to generate explicit IF–THEN rules. This combination provides both local and global explanations, while offering more precise insights than *SHAP* feature importance and a richer understanding of interactions between variables. Experiments conducted on four medical datasets demonstrate broad coverage and accurate class characterization, with high *WRAcc* and *lift* values. The associated local explanations achieve over 90% fidelity for binary models. Although developed in a medical context, the approach can be applied to any domain requiring intelligibility and trust in predictive models.

Annexes

A Proposition

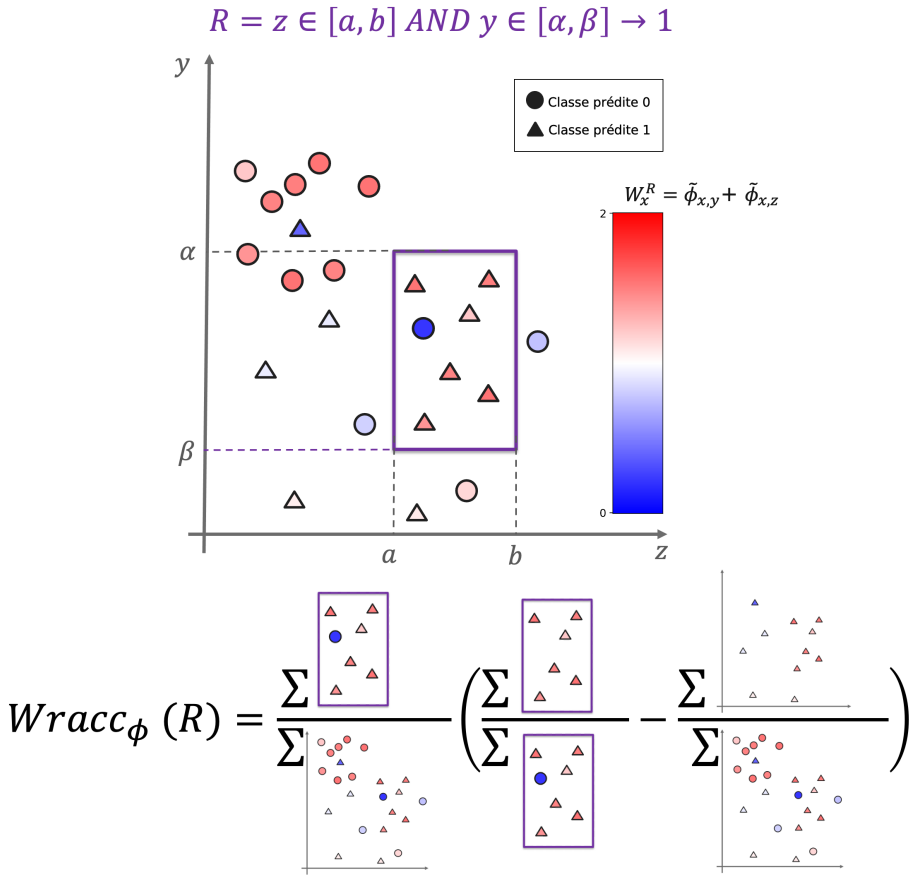
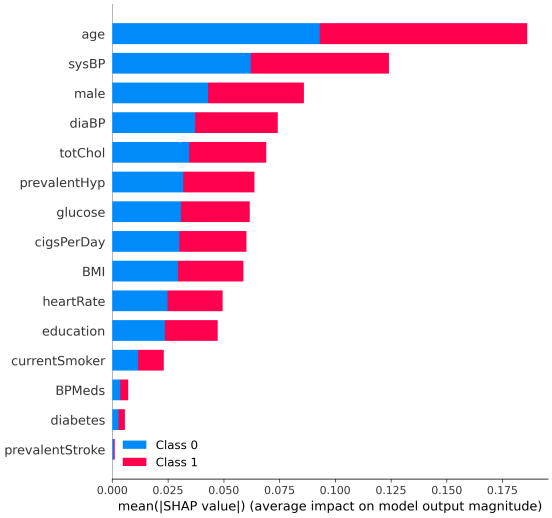


Figure A1: Illustration de la mesure de qualité proposée nommée $\text{WRAcc}_{\phi}(R)$

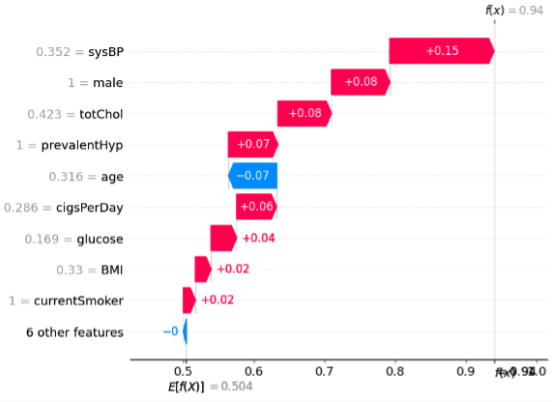
B Exemples d'explications



(a) Visualisation globale classique des valeurs SHAP.

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

(b) Règles globales extraites par notre méthode (10 par classe).



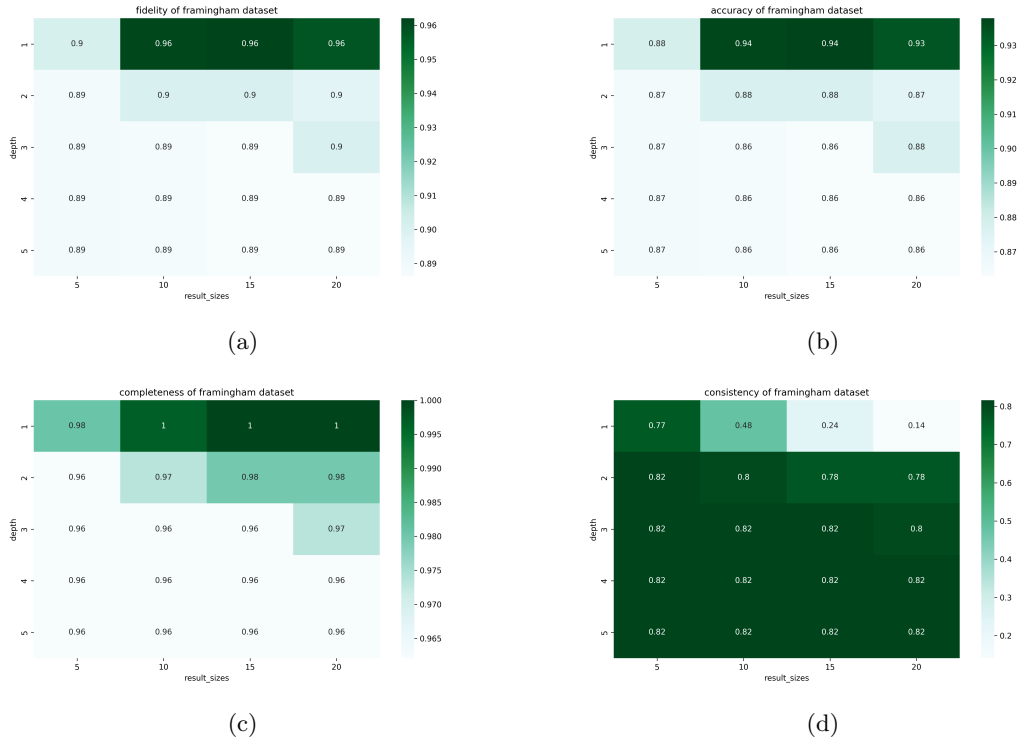
(c) Explication locale SHAP pour la première instance du jeu de test.

$\phi_x(R)$	R	c
0.2174	prevalentHyp=1 AND sysBP ≥ 0.32	$\to 1$
0.1473	sysBP ≥ 0.32	$\to 1$
0.0828	male=1	$\to 1$
0.0701	prevalentHyp=1	$\to 1$

(d) Explication locale générée par notre méthode pour cette même instance.

Figure A2: Comparaison des explications globales (haut) et locales (bas) pour le jeu de données Framingham, en opposant les explications fournies par SHAP classique (gauche) à celles générées par notre méthode (droite).

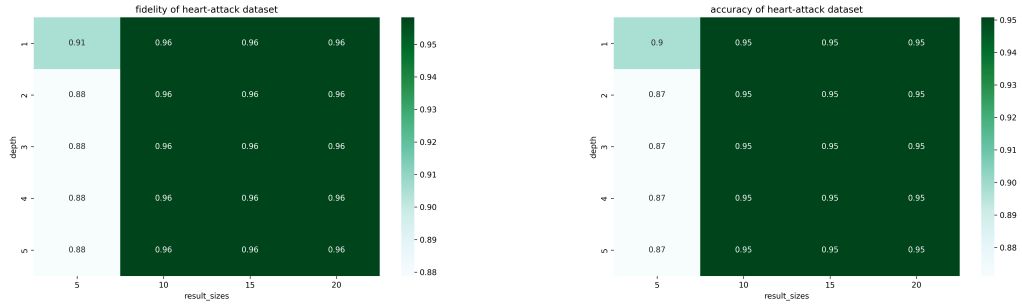
C Résultats du paramètrage



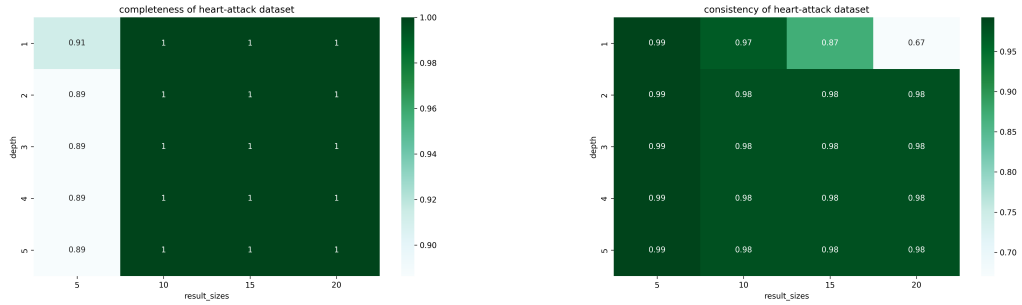
(a) (b)

(c) (d)

Figure A3: Métriques d'évaluation pour le jeu de données *Framingham* selon différents réglages de paramètres : (a) fidélité, (b) précision, (c) complétude et (d) cohérence.

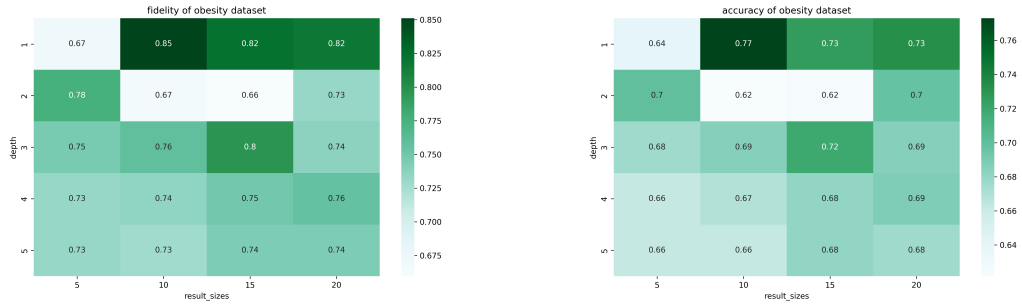


(a) (b)

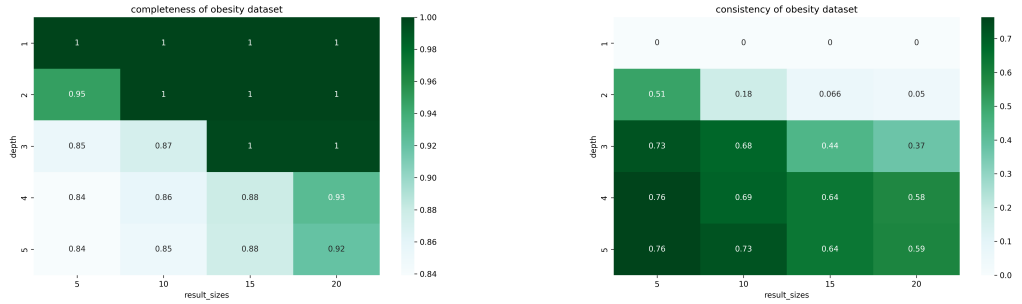


(c) (d)

Figure A4: Métriques d'évaluation pour le jeu de données *Heart-attack* selon différents réglages de paramètres : (a) fidélité, (b) précision, (c) complétude et (d) cohérence.



(a) (b)



(c) (d)

Figure A5: Métriques d'évaluation pour le jeu de données *Obesity* selon différents réglages de paramètres : (a) fidélité, (b) précision, (c) complétude et (d) cohérence.