Explications Probabilistes Préférées Pareto-Optimales : Équilibrer Contraintes Cognitives et Préférences Utilisateur

Louenas Bounia[1]

[1] LIPN-UMR CNRS 7030 Université Sorbonne Paris Nord, Villetaneuse, France

bounia@lipn.univ-paris13.fr

Résumé

*Les modéles d'apprentissage automatique déployés dans des domaines critiques nécessitent des explications compréhensibles. Les explications abductives identifient des ensembles minimaux de caractéristiques garantissant une décision, mais souffrent de deux limitations : dépasser les limites cognitives humaines et ignorer les préférences utilisateur (actionnabilité, équité, coût). Les explications probabilistes réduisent la taille via une erreur contrôlée, tandis que les explications préférées intègrent les préférences ; cependant, ces approches restent disjointes. Nous introduisons le premier cadre unifiant ces paradigmes : des explications probabilistes préférées équilibrant contraintes cognitives, précision probabiliste et préférences utilisateur. Nous formulons le problème via une scalarisation pondérée (*WPPE*) et une optimisation de Pareto (*PPPE*), prouvons la* NP*-dureté pour les arbres de décision, et exploitons la supermodularité pour établir des garanties d'approximation. Nous proposons trois algorithmes complémentaires : la descente gloutonne pondérée (*WGD*) avec ratio $(e^{p_w} - 1)/p_w$, l'énumération de la frontière de Pareto (*PFE*) pour l'exploration interactive des compromis non-dominés, et l'algorithme lexicographique stratifié (*LSA*) pour les préférences ordinales strictes. Notre cadre permet une navigation rigoureuse dans l'espace des compromis Pareto-optimaux entre limites cognitives et préférences utilisateur.*

Mots-clés

Explications abductives, Explications probabilistes, Optimisation multi-objectifs, Pareto-optimalité, Préférences utilisateur, IA explicable.

Abstract

*Machine learning classifiers deployed in critical domains require comprehensible explanations. Abductive explanations identify minimal feature sets guaranteeing a decision but suffer from two limitations : exceeding human cognitive limits and ignoring user preferences (actionability, fairness, cost). Probabilistic explanations reduce size via controlled error, while preferred explanations integrate preferences ; however, these approaches remain disjoint. We introduce the first framework unifying these paradigms : preferred probabilistic explanations that balance cognitive constraints, probabilistic accuracy, and user preferences. We formulate the problem via weighted scalarization (*WPPE*) and Pareto optimization (*PPPE*), prove* NP*-hardness for decision trees, and exploit supermodularity to establish approximation guarantees. We propose three complementary algorithms : weighted greedy descent (*WGD*) with ratio $(e^{p_w} - 1)/p_w$, Pareto frontier enumeration (*PFE*) for interactive exploration of non-dominated trade-offs, and lexicographic stratified algorithm (*LSA*) for strict ordinal preferences. Our framework enables rigorous navigation of the Pareto-optimal trade-off space between cognitive limits and user preferences.*

Keywords

Abductive explanations, Probabilistic explanations, Multi-objective optimization, Pareto optimality, User preferences, Explainable AI.

1 Introduction

1.1 Motivation et Contexte

Les Classifieurs d'apprentissage automatique sont de plus en plus déployés dans des domaines à enjeux élevés tels que la santé [21], la justice pénale [19], l'octroi de prêts et les décisions d'emploi [47]. Ces modèles, incluant les arbres de décision, les forêts aléatoires et les réseaux de neurones, ont démontré des performances prédictives remarquables dans diverses applications.

Cependant, avec leur déploiement dans des secteurs critiques, la capacité d'expliquer les décisions des modèles est devenue essentielle pour la transparence, la confiance et la conformité réglementaire (droit à l'explication du RGPD, loi européenne sur l'IA). Ce besoin a conduit au développement de l'IA Explicable (XAI), avec l'adoption généralisée de méthodes agnostiques post-hoc populaires telles que LIME [49], SHAP [37] et les explications contrefactuelles [57]. Ces méthodes reposent sur des perturbations locales ou la théorie des jeux sans tenir compte de la structure du modèle.

Malgré leur popularité, les méthodes agnostiques souffrent de limitations fondamentales : elles peuvent générer des explications identiques pour des prédictions opposées [52, 23, 22], manquent de fondements théoriques rigoureux [35], et présentent une sensibilité aux perturbations des entrées [1].

Ces faiblesses compromettent la fiabilité dans les contextes critiques où les décisions ont des conséquences significatives.

Notre travail s'inscrit dans le cadre de l'XAI Formelle [38], fournissant des explications mathématiquement validées avec des garanties de fidélité et de robustesse. La forme la plus prominente est celle des *explications abductives*, qui identifient des ensembles minimaux de caractéristiques dont la fixation des valeurs garantit la même prédiction indépendamment des autres caractéristiques. Chaque caractéristique est nécessaire et collectivement suffisante pour justifier la décision.

Cependant, deux défis interdépendants émergent :

Défi 1 : Surcharge Cognitive. Même les explications minimales peuvent contenir trop de caractéristiques pour être traitées cognitivement. [41] ont établi que les humains ne peuvent traiter simultanément que 7 ± 2 éléments, une limite affinée à seulement 4 éléments pour les tâches de raisonnement complexe [14].

Défi 2 : Hétérogénéité des Préférences. Toutes les caractéristiques ne sont pas également pertinentes, et la même explication peut convenir à un utilisateur mais pas à un autre. L'utilité d'une explication dépend du profil utilisateur, de son expertise et du contexte de décision, notamment en ce qui concerne l'interprétabilité, l'actionnabilité et les contraintes éthiques [42]. Ignorer cette hétérogénéité peut conduire à des explications formellement correctes mais non informatives pour les utilisateurs finaux, comme l'illustrent les scénarios suivants :

- **Scénario A (Diagnostic médical) :** Un patient non expert préfère des explications avec des termes médicaux courants, tandis qu'un médecin est à l'aise avec cette terminologie.
- **Scénario B (Demande de prêt) :** Certaines caractéristiques, telles que *âge* > 50, ne sont pas actionnables, tandis que d'autres, comme le *ratio dette/revenu* $> 0{,}6$, peuvent être modifiées.
- **Scénario C (Applications sensibles au coût) :** Dans les contextes industriels et médicaux, certaines variables reposent sur des capteurs ou des tests coûteux, limitant leur utilisation dans les explications.

1.2 Le Manque dans la Recherche Actuelle

Les travaux existants traitent la concision et la compréhensibilité de manière disjointe. Les **explications probabilistes** [26, 58, 31, 34, 59, 12, 13, 2] relaxent l'exactitude via une erreur contrôlée ϵ, réduisant la taille pour respecter les limites cognitives, mais traitent toutes les caractéristiques uniformément sans intégrer les préférences utilisateur. À l'inverse, les **explications préférées** [4] intègrent les préférences (coût, actionnabilité, équité) [42, 53, 36] mais requièrent $\epsilon = 0$, produisant de longues explications mal adaptées aux besoins diversifiés des parties prenantes [36].

1.3 Contributions

Nous proposons le premier cadre raisonné unifiant les explications probabilistes et les préférences utilisateur. Notre approche calcule des explications qui sont simultanément **courtes** (via une erreur contrôlée ϵ), **précises** (avec des garanties probabilistes explicites) et **alignées sur les préférences** (optimisant l'utilité).

Contributions algorithmiques. Nous développons trois algorithmes complémentaires exploitant la supermodularité : la **Descente Gloutonne Pondérée (WGD)** optimise un objectif scalarisé avec un ratio d'approximation $(e^{p_w} - 1)/p_w$; l'**Énumération de la Frontière de Pareto (PFE)** explore efficacement les compromis non-dominés sans recherche exhaustive ; et l'**Algorithme Lexicographique Stratifié (LSA)** gère les préférences ordinales via des niveaux de priorité avec des garanties en temps polynomial.

Contributions théoriques. Nous établissons : (i) la NP-difficile de WPPE pour les arbres de décision, (ii) la préservation de la supermodularité sous pondération des préférences — cruciale pour nos garanties d'approximation, et (iii) la tractabilité en temps polynomial pour les arbres de décision malgré la dureté générale, permettant une implémentation pratique.

2 Préliminaires

Problèmes de classification. Nous considérons une classification binaire booléenne avec d caractéristiques. Soit $[d] = \{1, \ldots, d\}$ les indices des caractéristiques et $X_d = \{x_1, \ldots, x_d\}$ l'ensemble des caractéristiques. Chaque instance $\boldsymbol{x} = (x_1, \ldots, x_d) \in \{0,1\}^d$ est classifiée par $h : \{0,1\}^d \to \{0,1\}$. Nous supposons que le lecteur est familier avec l'apprentissage supervisé et les arbres de décision.

Explications Abductives et Probabilistes. Une *explication abductive*[1] pour $\boldsymbol{x}$ est un sous-ensemble de caractéristiques $S \subseteq [d]$ tel que toute instance $\boldsymbol{y}$ coïncidant avec $\boldsymbol{x}$ sur S satisfait $h(\boldsymbol{y}) = h(\boldsymbol{x})$. Une *raison suffisante* est une explication abductive minimale. Ces explications pouvant être longues, les *explications probabilistes* [2, 59] relaxent la correction via une erreur contrôlée. La *fonction d'erreur* $\mu_{h,x}(S)$ compte les instances $\boldsymbol{y}$ où $y_S = x_S$ mais $h(\boldsymbol{y}) \neq h(\boldsymbol{x})$. L'*erreur normalisée*

$$\epsilon_{h,x}(S) = \frac{\mu_{h,x}(S)}{2^{d-|S|}}$$

mesure la proportion de complétions incompatibles. Un ensemble S est une *explication probabiliste de niveau* ϵ si $\epsilon_{h,x}(S) \leq \epsilon$.

Explications Préférées. Toutes les explications ne sont pas également utiles. Les *explications préférées* [4] tiennent compte des préférences utilisateur (actionnabilité, coût, contraintes) [42, 56] via des pondérations de caractéristiques ou des préférences ordinales. Nous utilisons des

1. Contrairement à [23], nous ne nous restreignons pas aux explications abductives minimales au sens de l'inclusion ensembliste.

coûts additifs : x_i reçoit un poids $w(x_i) \in R^+$, donnant $\text{cost}(S) = \sum_{i \in S} w(x_i)$.

Fonctions Sous-modulaires. Une fonction $f : 2^{[d]} \to R$ est *monotone* si $f(S) \leq f(T)$ (croissante) ou $f(S) \geq f(T)$ (décroissante) pour tout $S \subseteq T$. Elle est *supermodulaire* si la perte marginale $L_f(i \mid S) = f(S) - f(S \setminus \{i\})$ satisfait $L_f(i \mid S) \geq L_f(i \mid T)$ pour tout $S \subseteq T$ et $i \in S$, intuitivement, retirer un élément impacte davantage les petits ensembles. Une fonction est *sous-modulaire* si l'inverse est vérifié ; *modulaire* si les contributions sont additives. La *courbure* mesure l'écart à la modularité et est cruciale pour les garanties d'approximation. Pour f positive et $I \subseteq [d]$ non-vide : $c_f(I) = 1 - \min_{i \in I} \frac{f(I) - f(I \setminus \{i\})}{f(\{i\}) - f(\emptyset)}$. On a $c_f(I) \in [0, 1]$, où $c_f(I) = 0$ indique la modularité et $c_f(I) = 1$ la super/sous-modularité maximale. Pour f supermodulaire décroissante, $c_f(I) < 1$ garantit une approximation à facteur constant sous contraintes de cardinalité [24, 55].

2.1 Travaux Connexes

Notre travail s'inscrit dans un courant de recherche connu sous le nom d'*IA explicable formelle* (XAI formelle) [38, 17, 16], qui étudie les explications dotées de garanties prouvables [51, 5, 8, 11, 10].

Explications Abductives et Probabilistes. [16] ont caractérisé les raisons suffisantes pour diverses classes de modèles, avec des algorithmes efficaces pour les arbres de décision [27, 2, 13], les forêts aléatoires [25, 30], les modèles linéaires [39, 54], les ensembles d'arbres [7] et les réseaux de neurones [11, 60]. Sur le versant probabiliste, Waldchen [61] ont prouvé la $\mathsf{NP}^{\mathsf{PP}}$-complétude pour les circuits booléens, Arenas et al [2] ont montré la NP-complétude pour les arbres de décision, et [12] ont exploité la sous-modularité pour l'approximation avec des garanties formelles.

Explications Préférées et Multi-Objectifs. [42, 48] souligne l'importance de l'interprétabilité personnalisée. [4] ont intégré les contraintes de coût, d'actionnabilité et d'équité [56, 32], mais se restreignent aux explications exactes ($\epsilon = 0$). Notre travail généralise cette approche en autorisant $\epsilon > 0$ via un compromis paramétré par λ, géré par WGD pour les préférences quantitatives et LSA pour les préférences ordinales. [44] a considéré l'optimisation multi-objectifs pour les contrefactuels ; notre PFE étend cela aux explications probabilistes sur l'espace de Pareto (ϵ, cost), en s'appuyant sur l'optimisation sous-modulaire [45, 55] pour une approximation efficace des problèmes NP-difficiles.

3 Formulations du Problème

3.1 Problème Principal

Le nombre d'explications abductives peut croître exponentiellement avec les caractéristiques, même pour des modèles simples comme les arbres de décision et les ensembles d'arbres [43, 27, 3, 29], rendant l'analyse exhaustive impraticable. Certaines sont longues et difficiles à interpréter [28, 6], tandis que d'autres peuvent ne pas refléter les préférences utilisateur. Pour résoudre ce compromis, nous considérons les *explications probabilistes préférées*, qui autorisent une erreur probabiliste contrôlée ϵ pour réduire la taille tout en intégrant explicitement les préférences utilisateur sur les caractéristiques.

Définition 3.1 (Explications Probabilistes Préférées)**.** *Étant donné un Classifieur $h : \{0,1\}^d \to \{0,1\}$, une instance $x \in \{0,1\}^d$ et une fonction de poids $w : X_d \to R^+$ modélisant les préférences utilisateur, une explication probabiliste préférée est un sous-ensemble $S \subseteq [d]$ minimisant l'erreur probabiliste $\epsilon_{h,x}(S)$ tout en optimisant le coût $\text{cost}(S) = \sum_{i \in S} w(x_i)$ sous la contrainte $|S| \leq k$.*

Le problème central calcule des sous-ensembles de taille au plus k réalisant des compromis optimaux entre précision et satisfaction des préférences, conduisant aux formulations par scalarisation pondérée et optimisation de Pareto ci-dessous.

3.2 Problème de Scalarisation Pondérée (WPPE)

La première approche agrège les deux objectifs — erreur probabiliste et coût de préférence — en une seule fonction objectif via un paramètre de compromis λ.

Problème 3.2 (**WPPE** – Explications Probabilistes Préférées Pondérées)**.** ***Entrée :***

- *Un Classifieur $h : \{0,1\}^d \to \{0,1\}$*
- *Une instance $x \in \{0,1\}^d$ à expliquer*
- *Un ensemble candidat de caractéristiques $I \subseteq [d]$*
- *Une limite de taille $k \leq |I|$ reflétant les contraintes cognitives*
- *Une fonction de poids $w : X_d \to R^+$ encodant les préférences*
- *Un paramètre de compromis $\lambda \geq 0$ contrôlant l'importance relative du coût*

Sortie : *Un sous-ensemble de caractéristiques S^* résolvant*

$$S^* \in \underset{S \subseteq I, |S| \leq k}{\text{argmin}} \left[\epsilon_{h,x}(S) + \lambda \cdot \text{cost}(S)\right]$$

Interprétation : Le paramètre λ agit comme un taux de change entre les objectifs : $\lambda = 0$ privilégie exclusivement la minimisation de l'erreur (en ignorant les préférences), tandis qu'un grand λ favorise les explications à faible coût au détriment de la précision. Puisque nous minimisons le coût de préférence, les caractéristiques non préférées reçoivent des poids élevés tandis que les caractéristiques préférées reçoivent des poids faibles. Cette formulation est particulièrement adaptée lorsque les utilisateurs peuvent exprimer explicitement leurs préférences sous forme de compromis quantitatif, comme c'est souvent le cas dans les contextes médicaux.

Remarque 3.3 (Cas particuliers et complexité)**.** WPPE *généralise les explications probabilistes classiques : lorsque $\lambda = 0$, ou lorsque les préférences sont uniformes ($w(x_i) = c$ pour tout i) et que la taille $|S|$ est fixée exactement à*

k, le problème se réduit à $\min_{S\subseteq I,|S|\leq k} \epsilon_{h,x}(S)$. Puisque ce problème est $\mathsf{NP}^{\mathsf{PP}}$-complet pour les classifieurs CNF *[61]—strictement au-dessus de* NP*—*WPPE *hérite de cette difficulté.*

Proposition 3.4 (Complexité)**.** *Le problème* WPPE *est* NP*-difficile lorsque le Classifieur h est représenté par un arbre de décision et I est une raison suffisante.*

3.3 Problème d'Optimisation de Pareto (PPPE)

La seconde formulation évite de choisir λ a priori en considérant l'ensemble complet des compromis optimaux, appelé frontière de Pareto [20]. Contrairement à la scalarisation pondérée qui agrège les deux objectifs en une fonction scalaire, l'approche de Pareto minimise simultanément l'erreur probabiliste $\epsilon_{h,x}(\cdot)$ et le coût de préférence $\mathrm{cost}(\cdot)$ sans privilégier l'un ou l'autre.

Définition 3.5 (Dominance de Pareto)**.** *Soient deux explications $S, S' \subseteq I$ avec $|S|, |S'| \leq k$. On dit que S* domine au sens de Pareto *S', noté $S \prec_P S'$, si et seulement si :*

$$\epsilon_{h,x}(S) \leq \epsilon_{h,x}(S') \quad \textit{et} \quad \mathrm{cost}(S) \leq \mathrm{cost}(S'),$$

avec au moins une inégalité stricte.

Définition 3.6 (Optimalité de Pareto)**.** *Une explication $S \subseteq I$ avec $|S| \leq k$ est* Pareto-optimale *si aucune autre explication $S' \subseteq I$ de taille au plus k ne la domine, c'est-à-dire qu'il n'existe pas de S' tel que $S' \prec_P S$. L'ensemble de toutes les explications Pareto-optimales constitue la* frontière de Pareto*, notée $\mathcal{P}_k$.*

Problème 3.7 (PPPE – Explications Probabilistes Préférées de Pareto)**.** ***Entrée :***

- *Un Classifieur $h : \{0,1\}^d \to \{0,1\}$*
- *Une instance $x \in \{0,1\}^d$ à expliquer*
- *Un ensemble candidat de caractéristiques $I \subseteq [d]$*
- *Une limite de taille $k \leq |I|$ reflétant les contraintes cognitives*
- *Une fonction de poids $w : X_d \to R^+$ encodant les préférences*

Sortie : *La frontière de Pareto $\mathcal{P}_k$, ou une approximation de celle-ci, i.e., l'ensemble (ou sur-ensemble) de toutes les explications Pareto-optimales de taille au plus k par rapport aux objectifs de minimisation $\epsilon_{h,x}(\cdot)$ et $\mathrm{cost}(\cdot)$.*

Avantages de l'approche de Pareto. Cette formulation offre plusieurs avantages. Premièrement, elle ne nécessite aucune spécification a priori du compromis, adaptée aux préférences mal définies ou évolutives [40]. Deuxièmement, elle permet une exploration interactive : les utilisateurs examinent la frontière $\mathcal{P}_k$ et sélectionnent leur compromis préféré. Troisièmement, elle révèle les *points de coude* [18] où les taux de substitution erreur-coût changent brusquement. Enfin, contrairement à la scalarisation pondérée qui capture un seul point par λ, elle fournit une caractérisation complète des solutions non-dominées.

4 Algorithmes d'Approximation

Nous présentons maintenant trois algorithmes complémentaires pour résoudre efficacement les problèmes WPPE et PPPE, en exploitant systématiquement la structure mathématique sous-jacente.

4.1 Propriétés Structurelles

Notre approche algorithmique repose fondamentalement sur une propriété de supermodularité relaxée de $\mu_{h,x}$ établie dans [12], en notant que $\epsilon_{h,x}$ n'est ni supermodulaire ni monotone en général.

Lemme 4.1 (Préservation de la Supermodularité)**.** *Soit $\mu_{h,x} : 2^{[d]} \to R^+$ une fonction supermodulaire et monotone décroissante. Soit $w : X_d \to R^+$ une fonction de poids. Définissons la fonction objectif pondérée :*

$$f_\lambda(S) = \mu_{h,x}(S) + \lambda \sum_{i\in S} w(x_i)$$

Alors $f_\lambda(\cdot)$ est supermodulaire, positive et monotone décroissante pour tout $\lambda \geq 0$ satisfaisant :

$$\lambda \leq \lambda^* = \min_{S\subseteq[d],\, i\notin S} \frac{\mu_{h,x}(S) - \mu_{h,x}(S\cup\{i\})}{w(x_i)}$$

4.2 Algorithme 1 (WGD)

Notre premier algorithme résout WPPE via une stratégie de descente gloutonne, en retirant itérativement les caractéristiques qui minimisent l'augmentation de la fonction objectif pondérée.

Algorithm 1 Descente Gloutonne Pondérée (WGD)

Require: Classifieur h, instance x, ensemble initial I, limite k, poids w, paramètre λ
Ensure: Explication $S_{\text{WGD}} \subseteq I$ avec $|S_{\text{WGD}}| \leq k$

```
n ← |I|, S_n ← I
for j = n jusqu'à 1 do
    for chaque caractéristique i ∈ S_j do
        Calculer la marge : Δ_i = f_λ(S_j \ {i}) − f_λ(S_j)
    end for
    i* ← argmin_{i∈S_j} Δ_i
    S_{j−1} ← S_j \ {i*}
    Stocker la paire (S_j, ε_{h,x}(S_j)) pour évaluation ultérieure
end for
retourner S_WGD ∈ argmin_{S_j : |S_j| ≤ k} [ε_{h,x}(S_j) + λ · cost(S_j)]
```

Intuition algorithmique. L'algorithme démarre avec l'ensemble complet I (qui peut être une raison suffisante) et retire itérativement la caractéristique dont la suppression cause la plus petite dégradation de la fonction objectif pondérée. Ce processus génère une séquence emboîtée $I = S_n \supset S_{n-1} \supset \cdots \supset S_1 \supset S_0 = \emptyset$, et l'algorithme retourne finalement la meilleure solution parmi celles-ci.

Proposition 4.2 (Garantie d'Approximation)**.** *Soit $\lambda \leq \lambda^*$ tel que défini dans le Lemme 4.1, de sorte que f_λ est supermodulaire, monotone décroissante et positive sur 2^I. Soit $S^* \in \operatorname{argmin}_{S\subseteq I, |S|\leq k} f_\lambda(S)$ une solution optimale de* WPPE*, et soit c_w la courbure de f_λ sur 2^I. Alors l'algorithme* WGD *retourne une solution S_{WGD} satisfaisant :*

$$f_\lambda(S_{WGD}) = \epsilon_{h,x}(S_{WGD}) + \lambda \cdot \text{cost}(S_{WGD}) \leq \alpha_w \cdot f_\lambda(S^*)$$

où $\alpha_w = \frac{e^{p_w}-1}{p_w}$ avec $p_w = \frac{c_w}{1-c_w}$.

Remarque 4.3. *En pratique, les ratios d'approximation observés sont significativement meilleurs que les bornes théoriques, comme le montrent nos expériences (Section 5). Cela s'explique par le fait que les bornes théoriques sont établies pour le pire cas, tandis que les instances réelles présentent généralement une structure plus favorable.*

4.3 Algorithme 2 : Énumération de la Frontière de Pareto (PFE)

Pour résoudre **PPPE** et construire une approximation de la frontière de Pareto, nous proposons une stratégie d'échantillonnage systématique de l'espace des paramètres λ.

Algorithm 2 Énumération de la Frontière de Pareto (PFE)

Require: Classifieur h, instance x, ensemble I, limite k, poids w, nombre d'échantillons L
Ensure: Frontière de Pareto approchée $\tilde{\mathcal{P}}$

Définir $\lambda_{\min} = 10^{-3}$
Calculer $\lambda^* = \min_{S\subseteq[d],\, i\notin S} \frac{\mu_{h,x}(S)-\mu_{h,x}(S\cup\{i\})}{w(x_i)}$
Calculer le facteur géométrique : $\gamma = \left(\frac{\lambda^*}{\lambda_{\min}}\right)^{1/(L-1)}$
Générer la séquence $\{\lambda_1, \ldots, \lambda_L\}$ avec $\lambda_1 = \lambda_{\min}$ et $\lambda_{\ell+1} = \gamma \cdot \lambda_\ell$
Solutions $\leftarrow \emptyset$
for chaque $\lambda \in \{\lambda_1, \ldots, \lambda_L\}$ **do**
    $S_\lambda \leftarrow \text{WGD}(h, x, I, k, w, \lambda)$
    Calculer $\epsilon_\lambda = \epsilon_{h,x}(S_\lambda)$ et $c_\lambda = \text{cost}(S_\lambda)$
    Solutions $\leftarrow$ Solutions $\cup \{(S_\lambda, \epsilon_\lambda, c_\lambda)\}$
end for
Supprimer les solutions dupliquées (paire (ϵ, c))
$\tilde{\mathcal{P}} \leftarrow$ FilterDominance(Solutions)
return $\tilde{\mathcal{P}}$

Procédure FilterDominance. Cette procédure itère sur l'ensemble des solutions et élimine celles dominées au sens de Pareto. Pour chaque paire de solutions (S, S'), elle vérifie si S domine S' (auquel cas S' est supprimée) ou vice versa. Elle peut être implémentée efficacement en temps $O(L^2)$ en triant préalablement les solutions selon une dimension.

Choix de la progression géométrique. L'utilisation d'une suite géométrique pour λ (plutôt qu'arithmétique) est motivée par deux considérations. Premièrement, elle assure une exploration uniforme de l'espace des compromis sur échelle logarithmique sur $[\lambda_{\min}, \lambda^*]$, ce qui est naturel car les ratios de coût sont souvent plus significatifs que leurs différences absolues. Deuxièmement, elle garantit une couverture complète de la frontière de Pareto dans le domaine de validité du Lemme 4.1, comme formalisé par le théorème suivant.

Définition 4.4 (Approximation de Pareto (α, β))**.** *Une frontière de Pareto approchée $\tilde{\mathcal{P}}$ est une approximation (α, β) de la vraie frontière de Pareto $\mathcal{P}$ si, pour toute solution Pareto-optimale $S^* \in \mathcal{P}$, il existe une solution $\tilde{S} \in \tilde{\mathcal{P}}$ satisfaisant :*

$$\epsilon_{h,x}(\tilde{S}) \leq \alpha \cdot \epsilon_{h,x}(S^*) \quad \textit{et} \quad \text{cost}(\tilde{S}) \leq \beta \cdot \text{cost}(S^*)$$

Théorème 4.5 (Garantie pour PFE)**.** *Soit α_w le ratio d'approximation de* WGD *tel que défini dans la Proposition 4.2, sous la condition $\lambda \leq \lambda^*$ définie dans le Lemme 4.1. Alors l'algorithme* PFE *avec L échantillons tirés de $[\lambda_{\min}, \lambda^*]$ retourne une frontière $\tilde{\mathcal{P}}$ qui est une approximation de Pareto (α_w, α_w) de $\mathcal{P}$.*

4.4 Algorithme 3 : Approche Lexicographique Stratifiée (LSA)

Dans certains contextes applicatifs, les préférences utilisateur ne sont pas quantitatives mais ordinales : les utilisateurs peuvent classer les caractéristiques par niveaux de priorité, souhaitant utiliser en premier les caractéristiques de niveau 1, puis celles de niveau 2 uniquement si nécessaire, et ainsi de suite. Cette situation est fréquente en médecine (tests invasifs vs non-invasifs) ou en finance (données publiques vs privées).

Modèle de préférence lexicographique. Supposons que les caractéristiques soient partitionnées en m strates $I = I_1 \cup I_2 \cup \cdots \cup I_m$ avec $I_i \cap I_j = \emptyset$ pour $i \neq j$, où I_1 contient les caractéristiques les plus préférées et I_m les moins préférées. Une explication S est lexicographiquement préférée à S' si, en notant $n_i(S) = |S \cap I_i|$ le nombre de caractéristiques de S dans la strate i, il existe un niveau ℓ tel que :

$$n_i(S) = n_i(S') \quad \forall i < \ell \qquad \text{et} \qquad n_\ell(S) < n_\ell(S')$$

Principe de fonctionnement. L'algorithme LSA procède par niveaux de priorité, en commençant par les caractéristiques les moins préférées (niveau m) et en remontant progressivement vers les plus préférées (niveau 1). À chaque niveau ℓ, il retire itérativement des caractéristiques de I_ℓ tant que :

1. La limite de taille k n'est pas atteinte
2. L'erreur probabiliste reste inférieure à $\epsilon_{\max}$
3. Des caractéristiques restent à retirer dans le niveau courant

Ce processus garantit que les caractéristiques de faible priorité sont éliminées en premier, préservant autant que possible celles de haute priorité.

Théorème 4.6 (Garantie pour LSA)**.** *Pour chaque niveau de priorité ℓ, l'algorithme LSA produit une solution qui est α_w-approchée parmi toutes les solutions utilisant au plus le même nombre de caractéristiques des niveaux 1 à $\ell - 1$.*

Algorithm 3 Algorithme Lexicographique Stratifié (LSA)

Require: Classifieur h, instance x, strates $\{I_1, \ldots, I_m\}$, limite k, seuil d'erreur $\epsilon_{\max}$
Ensure: Explication lexicographiquement optimale S_{LSA}

$S \leftarrow I_1 \cup I_2 \cup \cdots \cup I_m$
if $\epsilon_{h,x}(S) > \epsilon_{\max}$ **then**
  retourner S {Infaisable}
end if
for niveau $\ell = m$ **jusqu'à** 1 **do**
  while $S \cap I_\ell \neq \emptyset$ **do**
    $i^* \leftarrow \text{argmin}_{i \in S \cap I_\ell} [\mu_{h,x}(S \setminus \{i\}) - \mu_{h,x}(S)]$
    $S_{\text{temp}} \leftarrow S \setminus \{i^*\}$
    if $\epsilon_{h,x}(S_{\text{temp}}) \leq \epsilon_{\max}$ **et** $|S_{\text{temp}}| \geq 1$ **then**
      $S \leftarrow S_{\text{temp}}$
    else
      break
    end if
  end while
end for
retourner S

4.5 Application aux Arbres de Décision

Les garanties d'approximation de la Proposition 4.2 et des Théorèmes 4.5 et 4.6 s'appliquent à toute classe d'hypothèses booléenne. Cependant, pour l'efficacité computationnelle, chaque appel à l'oracle $\mu_{h,x}(S)$ doit s'exécuter en temps polynomial. L'évaluation de $\mu_{h,x}$ dans le cas général est #P-difficile [15]. Pour les arbres de décision, [27, 33, 9] ont montré qu'elle peut être calculée en temps polynomial, tandis que [12] a démontré une évaluation en temps linéaire $O(|T| \cdot |S|)$, où $|T|$ est le nombre de nœuds de l'arbre.

Complexités résultantes. Nos algorithmes héritent des complexités suivantes dans le pire cas pour les arbres de décision. WGD effectue $O(n^2)$ appels à l'oracle (où $n = |I|$), chacun nécessitant $O(|T| \cdot d)$ opérations, donnant $O(n^2 \cdot |T| \cdot d)$. PFE effectue L appels à WGD sur $[\lambda_{\min}, \lambda^*]$, donnant $O(L \cdot n^2 \cdot |T| \cdot d)$ avec typiquement $L \in [20, 50]$. LSA effectue au plus n itérations nécessitant chacune $O(n)$ évaluations de $\mu_{h,x}(\cdot)$, donnant la même complexité $O(n^2 \cdot |T| \cdot d)$ que WGD. Cette tractabilité est particulièrement pertinente car les arbres de décision sont largement utilisés dans les domaines critiques (santé, finance, justice) en raison de leur interprétabilité [50], et nos algorithmes étendent cela en permettant des explications plus concises alignées sur les préférences tout en préservant l'efficacité computationnelle.

4.6 Analyse Comparative

Ayant établi la tractabilité en temps polynomial pour les arbres de décision, nous comparons maintenant les trois algorithmes :

Recommandations d'utilisation :

— **WGD** : Préférer lorsque le compromis souhaité est bien défini ($\lambda \leq \lambda^*$) et qu'une solution unique est recherchée. Idéal pour l'intégration dans des systèmes automatisés.

— **PFE** : Recommandé pour l'exploration interactive, lorsque les préférences utilisateur ne sont pas entièrement spécifiées a priori, ou pour présenter plusieurs options à l'utilisateur final.

— **LSA** : À utiliser dans les contextes où les préférences sont naturellement hiérarchiques (médecine, finance réglementée) et où les compromis quantitatifs sont difficiles à exprimer.

TABLE 1 – Comparaison des trois algorithmes proposés

Critère	WGD	PFE	LSA
Type de préférence	Quantitatif	Quantitatif	Ordinal
Sortie	Une solution	Frontière complète	Une solution
Choix de λ requis	Oui ($\leq \lambda^*$)	Non	Non
Complexité	$O(n^2\|T\|d)$	$O(Ln^2\|T\|d)$	$O(n^2\|T\|d)$
Ratio d'approximation	α_w	(α_w, α_w)	α_w par niveau
Exploration interactive	Non	Oui	Non

5 ÉVALUATION EXPÉRIMENTALE

Pour valider l'efficacité et la scalabilité de nos trois algorithmes proposés (WGD, PFE, LSA), nous avons mené des expériences extensives sur des Classifieurs à arbres de décision sur divers jeux de données de référence. Tous les algorithmes ont été implémentés en Python, en exploitant scikit-learn pour l'entraînement des modèles et NumPy pour les calculs numériques. Les expériences ont été réalisées sur une station de travail équipée d'un processeur Intel Core i9-9900 à 3,1 GHz et 64 Gio de RAM, sous Ubuntu 22,04 LTS.

Jeux de données. Nous avons considéré $B = 25$ jeux de données de classification binaire (10 à 10^5 caractéristiques) issus de Kaggle, OpenML et UCI, incluant MNIST38 et MNIST49, deux sous-ensembles binaires de MNIST opposant les chiffres 3 vs. 8 et 4 vs. 9 respectivement. Les tâches multi-classes ont été converties en binaire en considérant la classe dominante contre toutes les autres. Par manque d'espace, nous reportons les résultats sur $B = 10$.

Préférences Utilisateur. Nous avons considéré quatre modèles de préférences : (i) **préférences uniformes**, où toutes les caractéristiques reçoivent un poids égal, servant de référence ; (ii) **préférences basées sur SHAP**, utilisant les valeurs de Shapley comme poids d'importance locaux ; (iii) **préférences basées sur LIME**, dérivées des explications locales de LIME, et (iv) **importance des caractéristiques**, utilisant l'importance des caractéristiques de scikit-learn pour les arbres de décision. Par manque d'espace, nous ne reportons que les résultats pour les préférences basées sur SHAP dans l'article principal. Les résultats pour les autres scénarios sont fournis en matériel supplémentaire.

5.1 PROTOCOLE EXPÉRIMENTAL

Tâches d'Explication. Pour chaque benchmark b, une tâche d'explication est $(T, \boldsymbol{x}, I, k, w)$. Le Classifieur h est

représenté par un arbre de décision T entraîné via l'implémentation CART de scikit-learn [46]. L'instance $\boldsymbol{x}$ est sélectionnée aléatoirement dans l'ensemble de test ; l'ensemble candidat I est le chemin racine-feuille cohérent avec $\boldsymbol{x}$ (explication de chemin [27]). La fonction de poids w est dérivée des scénarios de préférences via une transformation monotone décroissante : $w(x_i) = \max\{s\} - s_i + \delta$ où s_i est le score d'importance brut et $\delta > 0$ assure la positivité. Cela mappe une importance plus élevée à des poids plus faibles, ce qui est cohérent avec notre objectif de minimisation du coût. Nous fixons la limite cognitive $k = 7 \pm 2$ [41] en général ; lorsque $|I|$ est petit, nous adaptons et fixons $k = 4 \pm 2$ pour éviter les cas triviaux où $k \geq |I|$. Nous imposons une limite de temps de 60 minutes par instance. Les performances sont évaluées en moyennant $\epsilon_{h,x}(S)$, $\mathrm{cost}(S)$ et $|S|$ sur $m = \min\{s, 150\}$ instances de test.

Scalabilité et Qualité d'Approximation. En dehors du cas $\epsilon = 0$, où les explications abductives préférées exactes peuvent être calculées via MAXSAT [4], et du cas $\lambda = 0$ où les explications probabilistes exactes sont accessibles via un encodage SAT [2], aucune méthode exacte n'est connue pour le problème général des explications probabilistes préférées. Cela motive notre approche par approximation. Pour évaluer empiriquement sa qualité, nous comparons l'Algorithme 1 (WGD) contre deux références exactes. Premièrement, pour $\lambda = 0$, nous utilisons la méthode basée sur SAT de [2], étendue à la version décisionnelle du Problème 3.4 en ajoutant la clause $\bigvee\{x_i : (x_I)_i = 1\} \vee \bigvee\{\neg x_i : (x_I)_i = 0\}$, et résolue via une recherche binaire sur $(0, 1]$ avec une précision 10^{-3}, nécessitant au plus 10 appels à GLUCOSE 4 (PySAT, timeout 60 min). Deuxièmement, pour $\lambda > 0$, nous comparons contre une référence par énumération exhaustive, calculant $\mathrm{argmin}_{S \subseteq I, |S| \leq k} f_\lambda(S)$ sur tous les sous-ensembles possibles, sur des jeux de données de petite dimension où cela est tractable, comme *placement*, *tic-tac-toe* et d'autres.

Impact de λ sur WGD. Nous évaluons l'Algorithme 1 (WGD) sur $\lambda \in [10^{-3} \cdot 2^{d-k},\ 10^3 \cdot 2^{d-k}]$, où d est le nombre de caractéristiques et k la limite de taille, en utilisant des préférences normalisées basées sur SHAP avec $\sum_{i \in I} w(x_i) = 1$ pour assurer un taux d'échange significatif entre $\epsilon_{h,x}(\cdot)$ et $\mathrm{cost}(\cdot)$.

Exploration de la Frontière de Pareto via PFE. Une limitation clé de WGD est qu'il nécessite de spécifier λ a priori, ce qui peut être difficile en pratique lorsque les préférences utilisateur ne sont pas entièrement quantifiées. L'Algorithme 2 (PFE) répond à cela en calculant un ensemble de solutions Pareto-optimales couvrant l'espace complet des compromis entre $\epsilon_{h,x}(\cdot)$ et $\mathrm{cost}(\cdot)$, sans nécessiter de choix explicite de λ. Nous évaluons PFE avec $L = 30$ échantillons tirés de $\lambda \in [10^{-3}, 10^3]$ avec un pas géométrique, et reportons les frontières de Pareto approchées sous forme de graphiques 2D (ϵ vs. cost), mettant en évidence les points de coude où le taux de substitution marginal entre les deux objectifs change brusquement.

Préférences Ordinales et LSA. Tandis que WGD et PFE gèrent les préférences quantitatives via λ, certains domaines expriment naturellement les préférences en termes ordinaux. L'Algorithme 3 (LSA) répond à cela en partitionnant les caractéristiques en strates de priorité. Nous évaluons LSA sur COMPAS [56], et le comparons contre WGD en termes de $\epsilon_{h,x}(S)$, composition des strates $n_\ell(S)$ et $|S|$.

5.2 Résultats Expérimentaux

Le Tableau 2 reporte les résultats sur 10 des 25 benchmarks pour des Classifieurs à arbres de décision avec $k = 7$ et $\lambda = 0$. Les colonnes acc et d donnent la précision et le nombre de caractéristiques binaires. Les lignes sont triées par taille moyenne d'explication de chemin $|I|$ [27]. Nous reportons l'erreur moyenne $\epsilon_{h,\boldsymbol{x}}(S)$ et la taille d'explication $|S|$ pour WGD et la méthode exacte basée sur SAT, ainsi que le temps d'exécution de SAT. Les jeux de données en bleu indiquent un timeout partiel ; en magenta un timeout complet. WGD termine toujours en moins d'1 seconde. Pour le régime général $\lambda > 0$, le Tableau 3 traite le jeu de données *horse colic* sur 15 valeurs représentatives de $\lambda \in [10^{-3} \cdot 2^{d-k},\ 10^3 \cdot 2^{d-k}]$, comparant WGD contre l'énumération exhaustive ; les résultats confirment que WGD atteint un $f_\lambda(S)$ normalisé systématiquement inférieur à la méthode exacte pour λ petit et modéré, avec une borne d'approximation pratique bien en dessous de la garantie théorique de la Proposition 4.2. Cependant, pour de grandes valeurs de λ (au-delà de λ^*), les garanties théoriques ne tiennent plus et le ratio d'approximation peut se dégrader significativement. Cette dégradation est amplifiée lorsque I est une explication de chemin plutôt qu'une raison suffisante, car f_λ peut perdre sa structure de monotonicité dans ce régime, expliquant l'explosion occasionnelle de la borne observée dans le Tableau 3. Globalement, WGD démontre une scalabilité remarquable, restant stable en millisecondes sur de grands jeux de données comme *gisette* et *farm ads* où SAT dépasse le timeout — tandis que l'écart $\epsilon_{h,\boldsymbol{x}}(S_{\mathrm{WGD}}) - \epsilon_{h,\boldsymbol{x}}(S^*)$ reste négligeable pour $\lambda = 0$ et $|S_{\mathrm{WGD}}|$ est en moyenne inférieure à $|S^*|$.

Impact de λ sur WGD. La Figure 1 illustre l'effet de λ sur le jeu de données placement : à mesure que λ augmente, $\mathrm{cost}(S)$ et $|S|$ décroissent monotonicalement tandis que $\epsilon_{h,x}(S)$ augmente, confirmant que λ contrôle efficacement le compromis entre satisfaction des préférences et précision probabiliste. Pour $\lambda \leq \lambda^*$, le ratio d'approximation empirique reste bien en dessous de la borne théorique de la Proposition 4.2, confirmant la quasi-optimalité de WGD. Au-delà de λ^*, les garanties d'approximation ne tiennent plus et la qualité des solutions peut se dégrader, en particulier lorsque I n'est pas une raison suffisante. Cela souligne la nécessité d'explorer l'espace complet des compromis via PFE plutôt que de s'engager sur un λ unique a priori.

Exploration de la Frontière de Pareto via PFE. La Figure 2 reporte la frontière de Pareto approchée retournée par PFE sur COMPAS (préférences SHAP, $k = 7$, $|I| = 11$), comprenant 7 solutions non-dominées allant de haute fidélité ($\epsilon = 0{,}274$, $\mathrm{cost} = 0{,}998$, $|S| = 7$) à un ali-

Benchmark				$\epsilon_{h,\boldsymbol{x}}(S)$		$\lvert S\rvert$		Temps (s)
nom	acc	d	$\lvert I\rvert$	WGD	SAT	WGD	SAT	SAT
student perf.	92.04	30	5.37	0.27 ($\pm$0.10)	0.27 ($\pm$0.10)	2.03	2.03	2.11
hungarian	63.71	13	6.69	0.12 ($\pm$0.11)	0.12 ($\pm$0.09)	3.57	3.57	1.79
horse colic	75.94	40	6.74	0.14 ($\pm$0.06)	0.14 ($\pm$0.06)	4.09	4.09	11.71
loan eligibility	74.08	68	8.49	0.18 ($\pm$0.12)	0.22 ($\pm$0.13)	5.73	6.81	43.96
wine	70.11	11	9.05	0.09 ($\pm$0.09)	0.10 ($\pm$0.11)	5.66	5.64	35.48
employee attr.	83.37	63	10.58	0.07 ($\pm$0.08)	0.21 ($\pm$0.10)	6.44	7.04	1008.42
compas	68.02	40	10.97	0.05 ($\pm$0.08)	0.06 ($\pm$0.09)	5.86	6.82	1091.88
mnist49	96.28	784	15.60	0.38 ($\pm$0.13)	–	6.92	–	–
gisette	94.37	5000	21.39	0.33 ($\pm$0.10)	–	6.92	–	–
farm ads	81.12	54877	23.18	0.14 ($\pm$0.16)	–	6.34	–	–

TABLE 2 – Résultats expérimentaux pour $\lambda = 0$ (Problème 3.2) sur 10 benchmarks avec h arbre de décision, $k \in \{5, 6, 7\}$. En bleu : timeout partiel ; en magenta : timeout complet.

λ	$\epsilon_{h,\boldsymbol{x}}(S)$		$\lvert S\rvert$		$f_\lambda(S)$ (norm.)	
	WGD	Exact	WGD	Exact	WGD	Exact
0.0010	0.1105 ($\pm$0.0689)	0.1105 ($\pm$0.0689)	4.7387 ($\pm$0.6536)	4.7387 ($\pm$0.6536)	0.0587	0.0587
0.2507	0.1105 ($\pm$0.0689)	0.2268 ($\pm$0.0789)	4.7387 ($\pm$0.6536)	3.0270 ($\pm$1.3046)	0.0587	0.1205
0.5005	0.1122 ($\pm$0.0684)	0.2401 ($\pm$0.0719)	4.7297 ($\pm$0.6837)	2.8919 ($\pm$1.3244)	0.0596	0.1275
0.7502	0.1122 ($\pm$0.0684)	0.2626 ($\pm$0.0904)	4.7297 ($\pm$0.6837)	2.7658 ($\pm$1.2445)	0.0596	0.1395
10^0	0.1291 ($\pm$0.0614)	0.2672 ($\pm$0.0882)	4.6396 ($\pm$0.9280)	2.7387 ($\pm$1.2499)	0.0686	0.1419
2.15×10^6	0.2409 ($\pm$0.1115)	0.3825 ($\pm$0.1694)	3.1171 ($\pm$1.4316)	1.7748 ($\pm$0.9743)	0.1280	0.2032
6.98×10^6	0.2490 ($\pm$0.1174)	0.3905 ($\pm$0.1730)	3.0180 ($\pm$1.4205)	1.7027 ($\pm$0.9260)	0.1323	0.2074
1.18×10^7	0.2566 ($\pm$0.1157)	0.3944 ($\pm$0.1743)	2.9009 ($\pm$1.4203)	1.6847 ($\pm$0.9201)	0.1363	0.2095
1.66×10^7	0.2626 ($\pm$0.1214)	0.3949 ($\pm$0.1740)	2.8198 ($\pm$1.4155)	1.6847 ($\pm$0.9201)	0.1395	0.2098
2.15×10^7	0.2632 ($\pm$0.1230)	0.3949 ($\pm$0.1740)	2.8108 ($\pm$1.4111)	1.6847 ($\pm$0.9201)	0.1398	0.2098
2.15×10^{11}	0.3388 ($\pm$0.1526)	0.4342 ($\pm$0.1916)	2.0991 ($\pm$1.2003)	1.2432 ($\pm$0.5061)	0.2766	0.2306
3.82×10^{11}	0.3549 ($\pm$0.1562)	0.4351 ($\pm$0.1911)	2.0270 ($\pm$1.1505)	1.2252 ($\pm$0.4966)	0.3320	0.2311
6.79×10^{11}	0.3549 ($\pm$0.1562)	0.4400 ($\pm$0.1934)	2.0270 ($\pm$1.1505)	1.1982 ($\pm$0.4615)	0.4436	0.2337
1.21×10^{12}	0.3555 ($\pm$0.1561)	0.4402 ($\pm$0.1933)	2.0180 ($\pm$1.1546)	1.1802 ($\pm$0.4492)	0.6425	0.2338
2.15×10^{12}	0.3639 ($\pm$0.1581)	0.4428 ($\pm$0.1927)	1.9730 ($\pm$1.1347)	1.1622 ($\pm$0.3923)	1.0000	0.2352

TABLE 3 – WGD vs. méthode exacte sur Horse Colic ($d = 36$, $k = 5$, acc. $= 77{,}48\%$, préférences SHAP, $|R_d| \approx 6$, $\lambda \in [10^{-3} \cdot 2^{d-k},\ 10^{3} \cdot 2^{d-k}]$). $f_{\tilde{\lambda}}(S)$ (avec $\tilde{\lambda} = \lambda / 2^{d-k}$).

gnement maximal des préférences ($\epsilon = 0{,}517$, cost ≈ 0, $|S| = 1$). Un point de coude particulièrement attractif est la deuxième solution de Pareto ($\epsilon = 0{,}285$, cost $= 0{,}119$, $|S| = 7$), qui réalise une réduction de 88% du coût de préférence au prix d'une augmentation marginale de l'erreur de $0{,}011$, illustrant comment PFE permet aux utilisateurs d'identifier des solutions fortement alignées sur leurs préférences sans connaissance explicite de λ.

Préférences Ordinales et LSA. L'Algorithme 3 (LSA) est évalué sur COMPAS [56] avec $\epsilon_{\max} = 0{,}1$, $k = 7$ et trois strates : I_1 (actionnables : *Number_of_Priors*, *score_factor*, *Age_**), I_2 (partiels : *Misdemeanor*), et I_3 (attributs démographiques non-actionnables). LSA élimine lexicographiquement trois caractéristiques non-actionnables de I_3 tout en préservant $\epsilon_{h,x}(S) = 0$, donnant $n_1 = 4$, $n_2 = 1$, $n_3 = 2$ avec $|S| = 7$. Cela est significatif dans le contexte COMPAS : l'explication repose principalement sur des caractéristiques actionnables (*Number_of_Priors*, *score_factor*) tout en minimisant les attributs démographiques sensibles. En contraste, WGD ($\lambda = 0$) retourne $|S| = 5$ avec seulement $n_1 = 3$ et $n_2 = 0$, ignorant complètement la structure ordinale. LSA atteint un ratio non-actionnable strictement inférieur à erreur égale, confirmant qu'il produit des explications à la fois probabilistiquement valides et maximalement alignées sur les priorités utilisateur.

6 Conclusion et Perspectives

Nous avons introduit le premier cadre unifié pour les *explications probabilistes préférées*, reliant précision probabiliste, contraintes cognitives et préférences utilisateur via trois algorithmes complémentaires : WGD pour les préférences quantitatives avec ratio d'approximation formel $(e^{p_w} - 1)/p_w$, PFE pour l'exploration interactive de la frontière de Pareto sans engagement a priori sur λ, et LSA pour les préférences ordinales lexicographiques. Les trois exploitent la structure supermodulaire de $\mu_{h,x}$ pour assurer une tractabilité en temps polynomial et des certificats de qualité formels pour les arbres de décision. Les expériences sur 25 benchmarks confirment que WGD monte en charge sur des jeux de données allant jusqu'à $54{,}877$ caractéristiques en millisecondes là où les méthodes exactes basées sur SAT dépassent le timeout, que PFE révèle des points de coude significatifs permettant aux utilisateurs de sélectionner interactivement leur compromis précision-coût préféré, et que LSA produit systématiquement des explications avec moins de caractéristiques non-actionnables que WGD à erreur probabiliste égale, comme démontré sur COMPAS.

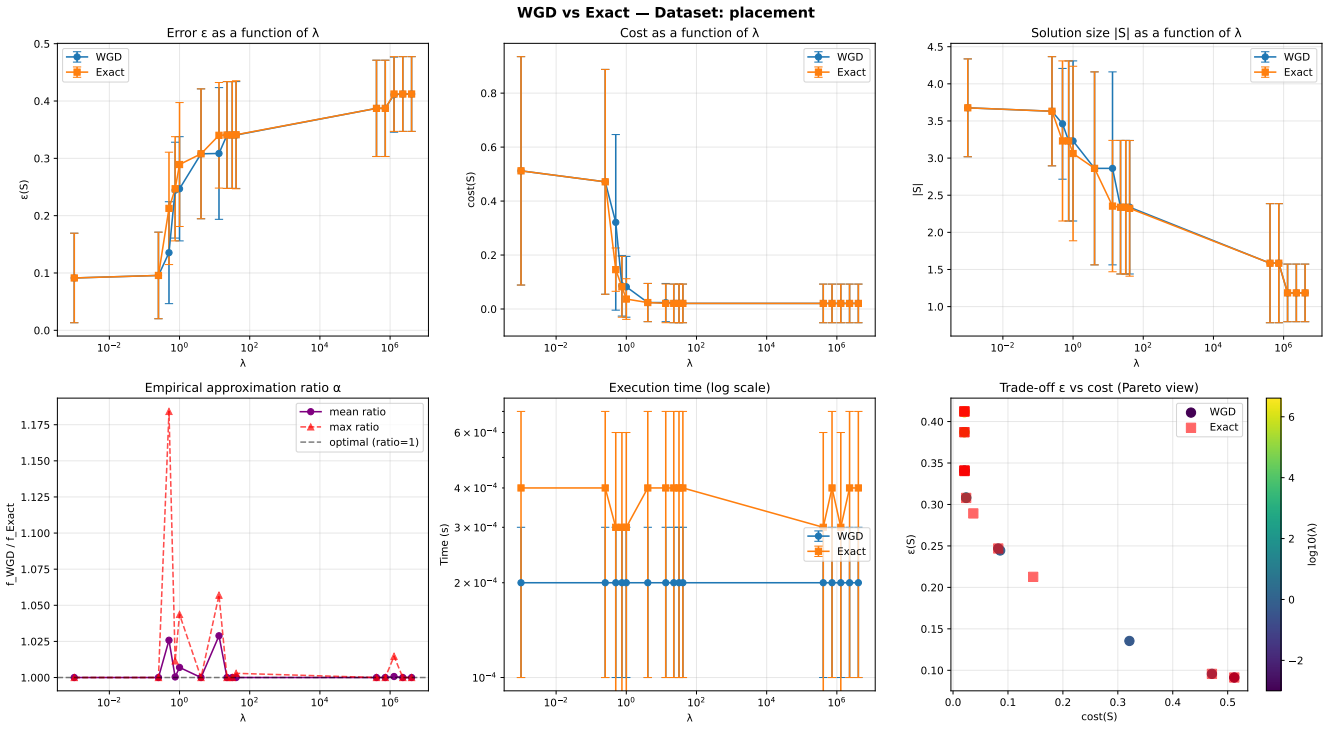

FIGURE 1 – WGD vs. méthode exacte sur le jeu de données placement (préférences SHAP) et $k = 4$.

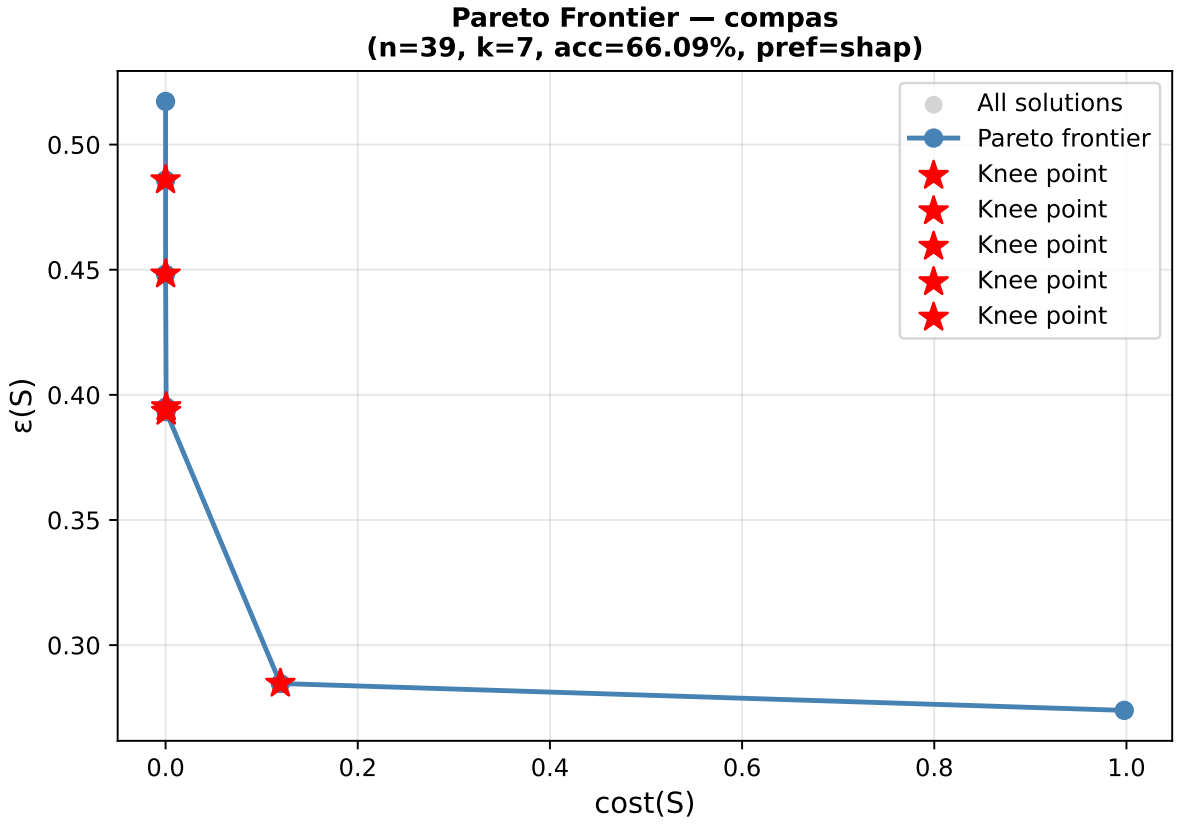

FIGURE 2 – Frontière de Pareto obtenue par PFE sur COMPAS.

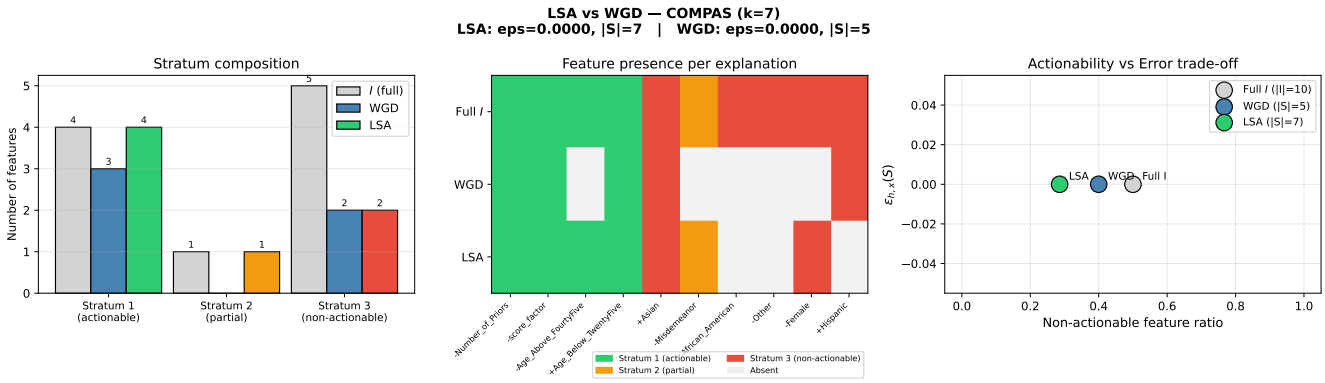

FIGURE 3 – LSA vs. WGD sur compas ($\epsilon_{max} = 0{,}1$, $k = 7$).

Plusieurs directions étendent naturellement ce travail. **(i)** L'extension des oracles efficaces pour $\mu_{h,x}$ aux forêts aléatoires, ensembles d'arbres et réseaux de neurones élargirait significativement l'applicabilité. **(ii)** Des bornes adaptatives à l'instance exploitant la structure du jeu de données ou de l'arbre pourraient donner des garanties plus serrées que le ratio de courbure dans le pire cas. **(iii)** Remplacer l'échantillonnage exhaustif de λ dans PFE par des stratégies bayésiennes permettrait d'identifier les points de coude plus efficacement. **(iv)** L'intégration de métriques d'équité de groupe comme contraintes dures dans WPPE renforcerait l'applicabilité dans les domaines réglementés tels que la justice pénale et le crédit. **(v)** Enfin, des études utilisateur réelles validant l'alignement des préférences au-delà des proxies synthétiques (SHAP, LIME, importance des caractéristiques) demeurent une étape essentielle vers le déploiement dans des systèmes de décision orientés vers l'humain.

Références

[1] David Alvarez-Melis and Tommi S. Jaakkola. On the robustness of interpretability methods. In *Proceedings of the 2018 ICML Workshop on Human Interpretability in Machine Learning (WHI 2018)*, Stockholm, Sweden, 2018. arXiv :1806.08049.

[2] Marcelo Arenas, Pablo Barceló, Miguel Romero Orth, and Bernardo Subercaseaux. On computing probabilistic explanations for decision trees. *Advances in Neural Information Processing Systems*, 35 :28695–28707, 2022.

[3] G. Audemard, S. Bellart, L. Bounia, F. Koriche, J-M. Lagniez, and P. Marquis. Trading complexity for sparsity in random forest explanations. In *Proc. of AAAI'22*, 2022.

[4] G. Audemard, S. Bellart, L. Bounia, F. Koriche, J.M. Lagniez, and P. Marquis. On preferred abductive explanations for decision trees and random forests. In *Proc. of IJCAI'22*, 2022.

[5] Gilles Audemard, Steve Bellart, Louenas Bounia, Frédéric Koriche, Jean-Marie Lagniez, and Pierre Marquis. On the computational intelligibility of boolean classifiers. In *Proc. of KR'21*, 2021.

[6] Gilles Audemard, Steve Bellart, Louenas Bounia, Frédéric Koriche, Jean-Marie Lagniez, and Pierre Marquis. On the explanatory power of boolean decision trees. *Data Knowledge Engineering*, 142, 2022.

[7] Gilles Audemard, Jean-Marie Lagniez, Pierre Marquis, and Nicolas Szczepanski. On the computation of example-based abductive explanations for random forests. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence (IJCAI)*, pages 3679–3687, 2024.

[8] Steve Azzolin, SAGAR MALHOTRA, Andrea Passerini, and Stefano Teso. Beyond topological self-explainable GNNs : A formal explainability perspective. In *Forty-second International Conference on Machine Learning*, 2025.

[9] Pablo Barceló, Mikaël Monet, Jorge Pérez, and Bernardo Subercaseaux. Model interpretability through the lens of computational complexity. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 15487–15498, 2020.

[10] Shahaf Bassan, Xuanxiang Huang, and Guy Katz. Unifying formal explanations : A complexity-theoretic perspective. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2026. Poster.

[11] Shahaf Bassan and Guy Katz. Towards formal xai : Formally approximate minimal explanations of neural networks. In *Proceedings of the 29th International Conference on Tools and Algorithms for the Construction and Analysis of Systems*, 2022.

[12] Louenas Bounia and Frederic Koriche. Approximating probabilistic explanations via supermodular mi-

nimization. In *Uncertainty in Artificial Intelligence (UAI 2023)*., 2023.

[13] Louenas Bounia and Insaf Setitra. Enhancing the intelligibility of decision trees with concise and reliable probabilistic explanations. *Data & Knowledge Engineering*, 2025.

[14] Nelson Cowan. The magical number 4 in short-term memory : A reconsideration of mental storage capacity. *Behavioral and Brain Sciences*, 24(1) :87–114, 2001.

[15] Yves Crama and Peter L Hammer. *Boolean functions : Theory, algorithms, and applications*. Cambridge University Press, 2011.

[16] A. Darwiche and A. Hirth. On the reasons behind decisions. In *Proc. of ECAI'20*, 2020.

[17] Adnan Darwiche. Logic for explainable AI. In *Proceedings of the 38th Annual ACM/IEEE Symposium on Logic in Computer Science (LICS)*, pages 1–11. IEEE, 2023.

[18] Indraneel Das and J. E. Dennis. A closer look at drawbacks of minimizing weighted sums of objectives for pareto set generation in multicriteria optimization problems. *Structural Optimization*, 14(1) :63–69, 1997.

[19] Julia Dressel and Hany Farid. The accuracy, fairness, and limits of predicting recidivism. In *Science Advances*, volume 4, page eaao5580, 2018.

[20] Matthias Ehrgott. *Multicriteria Optimization*. Springer-Verlag Berlin Heidelberg, 2nd edition, 2005.

[21] Andre Esteva, Brett Kuprel, Roberto A Novoa, Justin Ko, Susan M Swetter, Helen M Blau, and Sebastian Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639) :115–118, 2017.

[22] Xuanxiang Huang and João Marques-Silva. On the failings of shapley values for explainability. *Int. J. Approx. Reason.*, 2024.

[23] A. Ignatiev, N. Narodytska, and J. Marques-Silva. Abduction-based explanations for machine learning models. In *Proc. of AAAI'19*, pages 1511–1519, 2019.

[24] Rishabh Iyer and Jeff Bilmes. Submodular optimization with submodular cover and submodular knapsack constraints. *NeurIPS*, 2013.

[25] Y. Izza and J. Marques-Silva. On explaining random forests with SAT. In *Proc. of IJCAI'21*, pages 2584–2591, 2021.

[26] Yacine Izza, Xuanxiang Huang, Alexey Ignatiev, Nina Narodytska, Martin C. Cooper, and Joao Marques-Silva. On computing probabilistic abductive explanations. *Int. J. Approx. Reason.*, 2022.

[27] Yacine Izza, Alexey Ignatiev, and Joao Marques-Silva. On explaining decision trees. *ArXiv*, abs/2010.11034, 2020.

[28] Yacine Izza, Alexey Ignatiev, and Joao Marques-Silva. On tackling explanation redundancy in decision trees. *J. Artif. Intell. Res.*, 75 :261–321, 2022.

[29] Yacine Izza, Alexey Ignatiev, Sasha Rubin, Joao Marques-Silva, and Peter J. Stuckey. Most general explanations of tree ensembles (extended version). *IJCAI 2025*, 2025.

[30] Yacine Izza, Alexey Ignatiev, Peter J. Stuckey, and Joao Marques-Silva. Delivering inflated explanations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024.

[31] Yacine Izza, Kuldeep S. Meel, and João Marques-Silva. Locally-minimal probabilistic explanations. *ECAI*, 2024.

[32] Amir-Hossein Karimi, Gilles Barthe, Bernhard Schölkopf, and Isabel Valera. A survey of algorithmic recourse : Contrastive explanations and consequential recommendations. *ACM Computing Surveys*, 55(5) :1–29, 2023.

[33] F. Koriche, J.-M. Lagniez, P. Marquis, and S. Thomas. Knowledge compilation for model counting : Affine decision trees. In *Proc. of IJCAI'13*, pages 947–953, 2013.

[34] Frédéric Koriche, Jean-Marie Lagniez, and Chi Tran. Probabilistic explanations for regression models. In *Proceedings on Uncertainty in Artificial Intelligence*, 2025.

[35] M. Suresh Kumar, V. Soundarya, S. Kavitha, E.S. Keerthika, and E. Aswini. Credit card fraud detection using random forest algorithm. In *2019 3rd International Conference on Computing and Communications Technologies (ICCCT)*, 2019.

[36] Markus Langer, Daniel Oster, Timo Speith, Holger Hermanns, Lena Kästner, Eva Schmidt, Andreas Sesing, and Kevin Baum. What do we want from explainable artificial intelligence (xai) ? – a stakeholder perspective on xai and a conceptual model guiding interdisciplinary xai research. *Artificial Intelligence*, 296 :103473, 2021.

[37] S. Lundberg and S-I. Lee. A unified approach to interpreting model predictions. In *Proc. of NIPS'17*, 2017.

[38] Joao Marques-Silva. Logic-based explainability in machine learning. *ArXiv*, abs/2211.00541, 2023.

[39] Joao Marques-Silva, Thomas Gerspacher, Martin C. Cooper, Alexey Ignatiev, and Nina Narodytska. Explaining naive bayes and other linear classifiers with polynomial time and delay. *Neural Information Processing Systems*, 2020.

[40] Kaisa Miettinen. *Nonlinear Multiobjective Optimization*, volume 12 of *International Series in Operations Research & Management Science*. Kluwer Academic Publishers, Boston, 1999.

[41] G. A. Miller. The magical number seven, plus or minus two : Some limits on our capacity for processing information. *The Psychological Review*, 63(2) :81–97, 1956.

[42] T. Miller. Explanation in artificial intelligence : Insights from the social sciences. *Artificial Intelligence*, 2019.

[43] C. Molnar. *Interpretable Machine Learning*. Leanpub, 2020.

[44] Ramaravind K. Mothilal, Amit Sharma, and Chenhao Tan. Explaining machine learning classifiers through diverse counterfactual explanations. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 607–617. ACM, 2020.

[45] George Nemhauser and Laurence Wolsey. *Integer and Combinatorial Optimization*. Wiley-Interscience Series in Discrete Mathematics and Optimization. John Wiley & Sons, Inc., New York, 1988. Advisory Editors : Ronald L. Graham, Jan Karel Lenstra, Robert E. Tarjan.

[46] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn : Machine learning in Python. *Journal of Machine Learning Research*, 12 :2825–2830, 2011.

[47] Manish Raghavan, Solon Barocas, Jon Kleinberg, and Karen Levy. Mitigating bias in algorithmic hiring : Evaluating claims and practices. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 469–481, 2020.

[48] Yanou Ramon, Tom Vermeire, Olivier Toubia, David Martens, and Theodoros Evgeniou. Understanding consumer preferences for explanations generated by xai algorithms. *arXiv*, 2021.

[49] M. T. Ribeiro, S. Singh, and C. Guestrin. "why should I trust you ?" : Explaining the predictions of any classifier. In *Proc. of SIGKDD'16*, pages 1135–1144, 2016.

[50] Cynthia Rudin. Please stop explaining black box models for high stakes decisions. *ArXiv*, abs/1811.10154, 2018.

[51] A. Shih, A. Choi, and A. Darwiche. A symbolic approach to explaining bayesian network classifiers. In *Proc. of IJCAI'18*, 2018.

[52] Dylan Slack, Sophie Hilgard, Emily Jia, Sameer Singh, and Himabindu Lakkaraju. Fooling lime and shap : Adversarial attacks on post hoc explanation methods. In *Proceedings of the 2020 AAAI/ACM Conference on AI, Ethics, and Society (AIES)*, pages 180–186. ACM, 2020.

[53] Kacper Sokol and Peter A. Flach. One explanation does not fit all : The promise of interactive explanations for machine learning transparency. *KI-Künstliche Intelligenz*, 34(2) :235–250, 2020.

[54] Bernardo Subercaseaux, Marcelo Arenas, and Kuldeep S. Meel. Probabilistic explanations for linear models. In *Proceedings of the Thirty-Ninth AAAI Conference*, 2025.

[55] Maxim Sviridenko, Jan Vondrák, and Justin Ward. Optimal approximation for submodular and supermodular optimization with bounded curvature. *Mathematics of Operations Research*, 42(4) :1197–1218, 2017.

[56] Berk Ustun, Alexander Spangher, and Yang Liu. Actionable recourse in linear classification. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 10–19, 2019.

[57] Sandra Wachter, Brent Mittelstadt, and Chris Russell. Counterfactual explanations without opening the black box : Automated decisions and the GDPR. *Harvard Journal of Law & Technology*, 31(2) :841–887, 2018.

[58] Stephan Wäldchen. *Towards explainable artificial intelligence : Interpreting neural network classifiers with probabilistic prime implicants*. PhD thesis, Technische Universität Berlin, 2022. Ph.D. thesis.

[59] Eric Wang, Pasha Khosravi, and Guy Van den Broeck. Probabilistic sufficient explanations. In *Proceedings of the 30th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 3082–3088, 2021.

[60] Min Wu, Haoze Wu, and Clark Barrett. Verix : Towards verified explainability of deep neural networks. In *NeurIPS*, 2023.

[61] Stephan Wäldchen, Jan MacDonald, Sascha Hauch, and Gitta Kutyniok. The computational complexity of understanding binary classifier decisions. *Journal of Artificial Intelligence Research*, 70 :351–387, 2021.