RFIViz : Random Forest Interactive Visualisation, un outil simple pour comprendre les modèles

Zeyneb M'Hamedi, Christel Dartigues-Pallez, Timothy Bell

RFIViz : Random Forest Interactive Visualisation, un outil simple pour comprendre les modèles

Timothy H. Bell*, Zeyneb M'hamedi*
Christel Dartigues-Pallez*

*Université Côte d'Azur, I3S, SPARKS team, F-06560 Sophia, France
timothy.bell@i3s.unice.fr, Christel.Dartigues-Pallez@univ-cotedazur.fr

Résumé. Les Random Forests (RF) constituent un outil d'apprentissage automatique largement adopté pour les tâches de classification en raison de leur facilité d'interprétation par rapport à des modèles complexes tels que les réseaux neuronaux. Cependant, dans les scénarios où les RF sont construits dans des espaces de caractéristiques étendus et des échantillons abondants, leur interprétabilité diminue au fur et à mesure que les paramètres précédents augmentent. Cette perte d'interprétabilité empêche les utilisateurs de comprendre la logique qui sous-tend les classifications d'échantillons.

Notre recherche relève ce défi en proposant un outil de visualisation pour une meilleure compréhension des modèles RF, quelle que soit leur taille, particulièrement adapté aux utilisateurs non experts. Nous reconnaissons la nécessité d'outils interactifs qui facilitent une compréhension plus approfondie du raisonnement algorithmique employé par les modèles de RF et en particulier les systèmes de filtrage et de recherche pour surmonter l'espace de caractéristiques de haute dimension.

1 Introduction

Les RF sont utilisées dans divers domaines d'applications comme outil de classification ou de régression. Les arbres de décision et les RF sont généralement interprétables. Cependant, avec des données à haute dimension, il peut être difficile d'examiner chaque caractéristique et chaque arbre de décision individuellement. Comme pour la plupart des applications, l'utilisateur final n'est pas toujours un expert de l'algorithme en question, d'où la nécessité d'obtenir des résultats facilement interprétables. Pour les plus initiés aux RF, la visualisation permet d'identifier les arbres les plus faibles ou les plus forts et de les traiter en conséquence.

L'objectif de ce travail est de se concentrer sur les scores et les informations pertinentes d'un modèle RF entraîné afin de comprendre et in fine d'améliorer la classification de certains échantillons. Dans cet article, nous allons présenter : (2) les travaux existant pour la visualisation des RF, (3) les tâches de l'utilisateur que l'on cherche à incorporer dans l'outil, (4) les visualisations créées pour trois niveaux différents : les

arbres de décision, la forêt et les échantillons d'entrée, (5) et pour finir l'application avec un jeu de données portant sur la réussite des étudiants.

2 État de l'art

Quelques approches de visualisation des RF ont été réalisées. Nous distinguerons ici les approches statiques et interactives. Les méthodes statiques visent à représenter graphiquement les structures, les relations et les caractéristiques des RF, permettant aux utilisateurs de mieux comprendre leur fonctionnement et leurs performances. Elles fournissent des informations précieuses sur les modèles, sans nécessiter d'interfaces utilisateur complexes et interactives.

Certaines méthodes statiques utilisent des approches de visualisation communes utilisées en statistiques telles que t-SNE ou le tracé des valeurs propres. Des travaux plus récents font appel à des outils plus complexes tels que RF-PHATE Rhodes et al. (2020) et Explanable Matrix Neto et Paulovich (2020), mais ces deux méthodes rendent difficile une visualisation détaillée de chaque arbre de la forêt.

Forest Floor est un autre outil qui permet de représenter des RF Welling et al. (2016). Il représente la structure du modèle RF par des projections géométriques, tout en mettant l'accent sur les scores importants. L'inconvénient de cette visualisation est qu'elle présente une image beaucoup trop compliquée pour les utilisateurs novices. L'encombrement et l'approche géométrique rendent difficile l'interprétation des modèles RF à haute dimension et ne constituent donc pas une solution adéquate.

Colourful Trees Nsch et al. (2019) adopte une approche plus évolutive en représentant la structure, les paramètres et les caractéristiques des arbres par des couleurs, des angles de branches et des tailles. Cela permet une vue d'ensemble facile d'un arbre donné, quelle que soit sa taille, mais ne permet pas une analyse plus approfondie des composants d'un arbre (scores, échantillons utilisés pour la construction, seuils).

Pour surmonter les inconvénients des méthodes précédentes, notamment due à la complexité et aux connaissances requises, certaines méthodes introduisent des outils dynamiques pour simplifier l'interprétation des RF. Toutefois, la visualisation RAFT Cutler et Breiman (Random Forest Tool) est obsolète et très complexe. Une autre visualisation très complexe est PaintingClass Teoh et Ma (2003). Bien que codée pour les arbres de décision, elle peut être étendue aux FR. Cet outil est adapté à la construction d'arbres et en donne un aperçu ciblé, mais le manque d'informations apparentes sur la visualisation rend difficile à la fois la compréhension d'un arbre et la prise en main de cet outil.

Enfin, iForest Zhao et al. (2019) est un outil très intuitif qui met l'accent sur les voies qui ont mené à une décision, mais qui reste compliqué pour les utilisateurs novices en raison de l'utilisation de méthodes de traçage obscures.

3 Tâches utilisateur

Après une étude exhaustive des outils actuels utilisés pour la visualisation de RF, ainsi que des outils populaires actuels utilisés pour d'autres algorithmes d'apprentissage

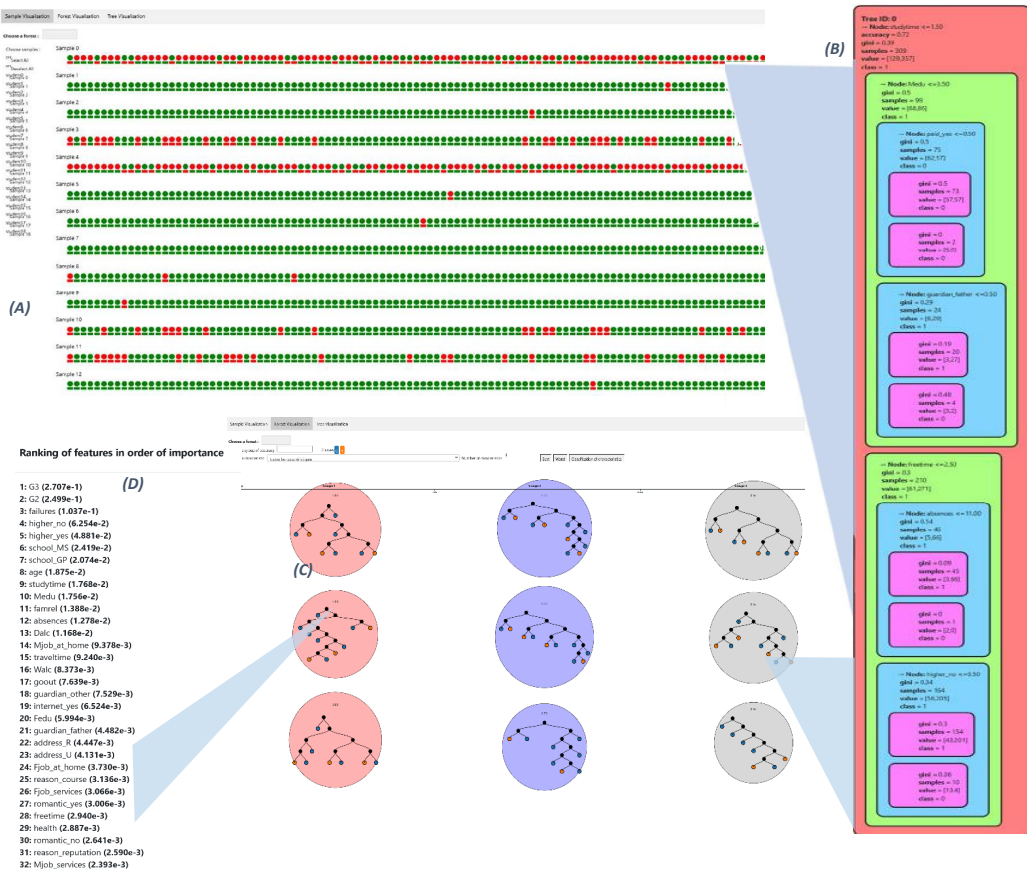

FIG. 1 – *Aperçu de l'outil de visualisation Random Forest. (A) Échantillons représentés par un cercle rempli et leurs classes prédites représentées par un rectangle rempli par rapport à leurs classes réelles. (B) Visualisation des structures arborescentes individuelles et de leurs informations, telles que la population divisée, le seuil de valeur et la classe prédite de chaque nœud. La classification de l'échantillon sélectionné peut également être affichée. (C) Une vue globale de tous les arbres construits dans le modèle forestier, leur précision correspondante et les caractéristiques utilisées. L'utilisateur peut également afficher tous les arbres utilisant les caractéristiques sélectionnées. (D) Une liste ordonnée de toutes les caractéristiques classées par ordre d'importance.*

automatique, nous avons retenu les tâches suivantes qui nous paraissent importantes. Nous supposons que le modèle a été construit et que cet outil de visualisation est utilisé pour les données de test/validation.

3.1 Tâche de l'utilisateur 1 : Comprendre le modèle en tant qu'utilisateur non averti

Comme nous l'avons vu dans les travaux connexes, la plupart des outils de visualisation supposent une bonne connaissance de l'algorithme (RF) ou de l'environnement nécessaire à la visualisation (codage, ingénierie). En outre, l'une des principales limites de ces visualisations existantes est la complexité visuelle, en particulier lorsqu'il s'agit de représenter un grand nombre d'arbres. Le besoin d'un outil simple mais explicable est évident. il nous parait essentiel de proposer à un utilisateur de notre outil, qui cherche à classifier des données, de pouvoir parcourrir tous les arbres du modèle pour pouvoir voir précisément comment chaque élément de son jeu de données a été classé, et notamment si la classification est compliquée (les arbres de la forêt ne sont pas tous d'accord entre eux) ou au contraire sans ambiguïté (les arbres de la forêt sont majoritairement unanimes).

3.2 Tâche de l'utilisateur 2 : Détecter les arbres les plus faibles

Certaines variantes des RF Zhang et Wang (2009)Yang et al. (2012)Kulkarni et Sinha (2012) nécessitent la suppression d'arbres afin d'améliorer les performances globales. Des approches utilisent des calculs complexes de seuils, mais il est essentiel de trouver un moyen rapide et visuellement simple d'identifier ces arbres. Nous souhaitons donc proposer un mode de visualisation qui permet de représenter tous les arbres de la forêt et d'immédiatement voir ceux qui sont le moins efficace. Il est également nécessaire, pour ces arbres aux performances mauvaises, de voir quelles caractéristiques y ont utilisées.

3.3 Tâche de l'utilisateur 3 : Comprendre la relation entre les caractéristiques, les prédictions et le modèle

Se débarrasser de la boîte noire des algorithmes d'apprentissage automatique est une tâche essentielle de la visualisation algorithmique moderne. L'affichage d'un flux d'information clair entre l'entrée (les caractéristiques) et la sortie (les prédictions) permet de justifier les classifications et d'ajuster éventuellement les valeurs des caractéristiques pour obtenir une classe préférée différente (cf section 5).

4 Approche proposée

Sur la base de la section précédente et des travaux antérieurs dans ce domaine, nous proposons une approche interactive principalement axée sur un utilisateur non-expert des RF. Notre outil -RFIViz- est conçu pour une compréhension rapide d'un modèle de FR et la classification de ses échantillons. Pour simplifier les RF, nous avons décidé

d'adopter une approche « Détails à la demande », en affichant les informations dans un flux en cascade pour permettre à tout utilisateur de comprendre une forêt donnée à différents niveaux. Pour surmonter la difficulté d'extraire des informations importantes (telles que l'importance des caractéristiques ou la précision de l'arbre), nous cherchons à utiliser des filtres dans les différentes scènes de visualisation de notre outil. Notre visualisation est disponible sur PyPi via « *pip install RFIViz* ».

Pour faciliter l'utilisation, nous avons choisi de mettre en œuvre une visualisation à trois niveaux : une vue de l'échantillon, une vue de la forêt et une vue de l'arbre. Toutes ces vues permettent à l'utilisateur de disposer de flux d'information différents et adaptés à ses besoins.

4.1 Vue des échantillons

L'exemple de vue tel qu'il apparait dans la figure 1 (A) affiche tous les échantillons passés par le modèle et la classification de chaque échantillon par chaque arbre de la forêt.

Lorsque l'on traite des échantillons introduits dans un modèle, il est nécessaire de comprendre pourquoi certains échantillons peuvent être classés d'une manière ou d'une autre. Les utilisateurs plus avertis doivent généralement parcourir chaque arbre du modèle et afficher la classe choisie, ce qui est toujours une tâche fastidieuse. Ici, l'utilisateur peut utiliser cette vue pour des tâches multiples :

— En ce qui concerne la précision des arbres, les arbres qui prédisent rarement la bonne classe sont fréquents dans les RF, et certains travaux Yang et al. (2012) se débarrassent de ces arbres pour augmenter les performances de l'ensemble du RF ;

— Comprendre pourquoi l'échantillon peut être mal classé. Comme l'explique la figure 2, la classification incorrecte/correcte est immédiatement reconnue par le rectangle coloré, ce qui permet à l'utilisateur de passer rapidement en revue les échantillons qui sont très mal classés. L'utilisateur peut ensuite recueillir des informations supplémentaires en consultant l'arbre correspondant (voir la section 4.3). ;

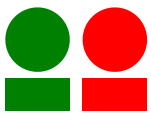

FIG. 2 – *Vue des échantillons utilisés dans le modèle, les cercles colorés représentent la classe prédite par l'arbre de la forêt (ici seulement deux couleurs dans cet exemple de classification binaire), les rectangles colorés sous les cercles indiquent si chaque arbre de la forêt a donné une bonne réponse (couleur verte) ou une mauvaise réponse (couleur rouge).*

4.2 Vue de la forêt

La vue suivante montre la forêt (Fig. 1 (C)) dans son ensemble et centre ses mécanismes sur le filtrage des caractéristiques. Chaque bulle représente un arbre de la forêt, trié horizontalement de gauche à droite en fonction de la précision de la prédiction. Une option de filtrage des caractéristiques est possible. L'utilisateur peut cliquer sur l'un des nœuds, ce qui met en évidence tous les autres nœuds de la forêt qui utilisent la même caractéristique pour diviser les échantillons, comme le montre la figure 4. Il peut également choisir d'afficher les arbres contenant certaines caractéristiques, choisies manuellement ou en fonction de leur importance. Cela permet de comprendre quelles sont les caractéristiques les plus importantes pour la prédiction. Un outil couramment utilisé dans divers progiciels RF consiste à afficher la liste des caractéristiques classées par ordre d'importance (Fig1 (D)), qui est également accessible dans cette vue. Enfin, l'utilisateur peut également sélectionner une option pour afficher les arbres en utilisant uniquement les N premières caractéristiques.

4.3 Vue de l'arbre

La dernière vue, l'arborescence (Fig. 1 (B), est une visualisation courante dans la plupart des implémentations RF. L'un des principaux avantages de notre interface est l'onglet consacré à la visualisation approfondie des informations d'un arbre. Cette fonctionnalité permet de comprendre en profondeur le fonctionnement d'un arbre spécifique, en offrant la possibilité de sonder chaque aspect de sa structure. Cette vue est construite de manière hiérarchique, chaque nœud est représenté par un rectangle, et les nœuds suivants sont construits à l'intérieur du rectangle précédent. Le nom du nœud (c'est-à-dire la caractéristique utilisée pour la division particulière) est placé en premier. Chaque nœud de l'arbre est détaillé, mettant en évidence des informations essentielles telles que :

- L'indice de Gini (score typique de notation des caractéristiques utilisé dans les RF), offrant un aperçu de l'impureté du nœud ;
- Le nombre total d'échantillons qui sont passés par ce nœud ;
- La valeur, qui indique le nombre d'échecs et de réussites à ce stade précis ;
- Le seuil calculé des valeurs des caractéristiques ;
- La classe prédite par le nœud, ce qui permet de comprendre la décision prise par le nœud sur la base des données qu'il a traitées.

De même, lorsque l'on appelle cette vue à partir de la vue Échantillon, l'arbre met en évidence le chemin choisi par l'échantillon spécifique, comme le montre la figure 3. Cela permet aux utilisateurs d'étudier plus en détail le choix de la classification pour le modèle sélectionné.

5 Scénario d'utilisation

Pour ce scénario, nous utilisons l'ensemble de données publié dans un article sur la prédiction de la réussite des étudiants Cortez et Silva (2008). Il s'agit d'une compilation de données sur les étudiants dans le but de prédire la réussite de la note de fin d'année.

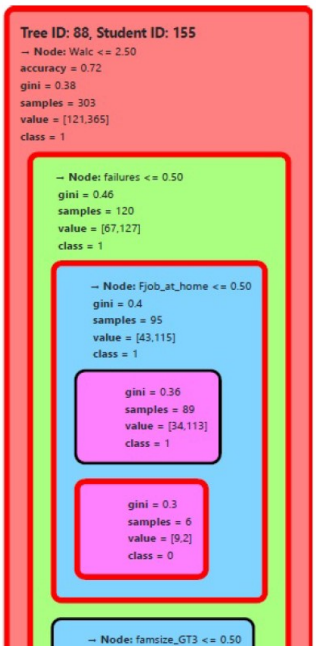

FIG. 3 – *Vue partielle d'un arbre singulier lorsqu'un échantillon a été préalablement sélectionné. Le chemin de l'échantillon classé est représenté par un rectangle rouge entourant les rectangles des nœuds.*

Lors du test du modèle pour cet ensemble de données particulier, nous avons obtenu un taux de précision de prédiction global de 70%. Nous examinons d'abord tous les échantillons dans la première scène de la visualisation (section 4.1). Nous remarquons ici que la majorité des échantillons présentent des cercles et des rectangles verts (représentant la classe de réussite correctement prédite), mais lorsqu'il y a un plus grand nombre de cercles rouges (classe d'échec), il y a également un plus grand nombre de rectangles rouges et verts (prédictions incorrectes et correctes). Cela indique que la forêt a tendance à classer facilement les élèves qui réussissent, mais qu'il y a plus de divergences pour les élèves qui échouent.

Nous avons choisi d'examiner un autre aspect, un échantillon 155 (voir Fig. 5, un échantillon est un étudiant) est classé avec précision pour la plupart des arbres, à l'exception d'un seul. D'un point de vue plus fonctionnel, étant donné le contexte de l'ensemble de données, cet élève peut être en danger d'échec en fonction de la raison pour laquelle l'arbre de décision a mal classé cet échantillon. Dans ce cas, et pour cet arbre, l'élève est considéré comme « en échec » en raison de la caractéristique « Fjob-at-home » (le parent n'a pas d'emploi). L'utilisateur peut alors décider si cette caractéristique favorise ou non la réussite de l'élève.

Une autre tâche consisterait à examiner les caractéristiques les plus importantes. En examinant la liste d'importance des caractéristiques (voir Fig.1 (D)), la caractéristique « G2 » apparaît comme la plus importante pour les divisions. Après un examen plus approfondi, en sélectionnant la caractéristique et en mettant en évidence sa présence dans tous les autres arbres, nous remarquons que la plupart des arbres dont la précision est supérieure à 80% utilisent cette caractéristique pour les divisions (voir la Fig.4).

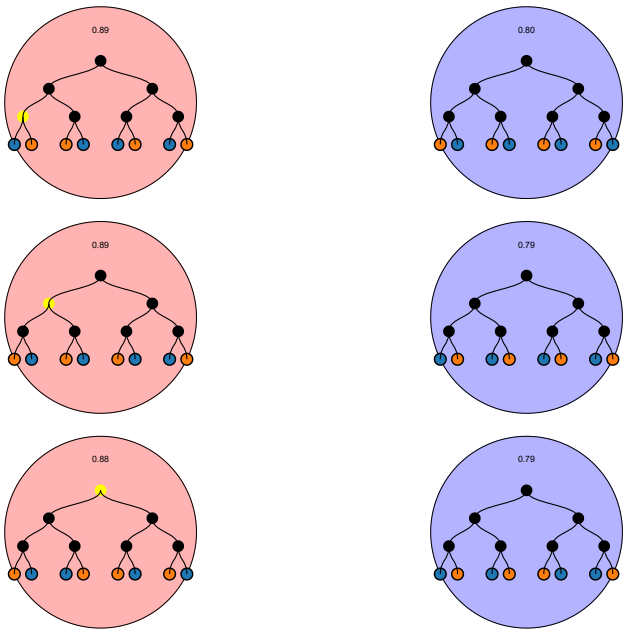

FIG. 4 – *Vue globale de la forêt (seuls quelques arbres sont représentés ici). À gauche, les arbres dont la précision est supérieure à 80% et à droite, les arbres dont la précision est comprise entre 60% et 80%. L'utilisateur a sélectionné une caractéristique, qui est surlignée en jaune lorsqu'elle est présente.*

6 Discussion et travaux futurs

Notre approche répond aux limites des approches déjà existantes que nous avons évoquées au début de cet article. en premier lieu, nous offrons une visualisation simple et intuitive : les arbres de la forêt, les caractétistiques les plus importantes ainsi que les échantillons sont rapidement visualisables. L'autre point fort auquel nous avons particulièrement fait attention est le fait que cette visualisation ne doit pas seulement être faite pour des experts mais également pour les utilisateurs finaux, en lleur permettant de voir, pour chaque échantillon, les réponses de tous les arbres de la forêt. D'un simple coup d'oeil un utilisateur non expert peut ainsi voir si, pour un échantillon donné, une

grande majorité des arbres sont d'accord entre eux ou si au contraire la décision est plus partagée.

Les travaux futurs permettront d'afficher davantage de valeurs telles que les scores f, l'entropie, afin que l'utilisateur puisse avoir une meilleure idée de la performance du modèle sélectionné. La visualisation sera également étendue à d'autres types de méthodes d'ensemble d'arbres de décision : Extremely Randomized Trees Geurts et al. (2006), XGBoost Chen et Guestrin (2016).

Nous prévoyons également d'ajouter une connexion entre la visualisation et l'instanciation RF, permettant à l'utilisateur d'interagir directement avec le modèle construit. Cela permettrait de mettre en œuvre Yang et al. (2012) et d'étendre la tâche de l'utilisateur en interagissant avec l'entrée et le modèle sur la base d'un résultat requis. Un autre problème souvent rencontré avec les approches interactives est le temps de calcul, en particulier pour la construction de la vue de la forêt (en raison de la création des nœuds et des branches). Nous cherchons à améliorer ce point en réduisant le temps de latence du client lors de la génération des éléments web.

Enfin, la visualisation résout les tâches de l'utilisateur énumérées dans la section 3, d'autres tâches de l'utilisateur ont été traitées dans d'autres outils tels que Zhao et al. (2019). Une compilation des différentes tâches des RF existantes permettra d'obtenir un outil plus complet. Cependant, notre objectif est toujours de faire en sorte que RFIViz soit destiné à des utilisateurs non experts, c'est pourquoi il faut encore travailler à la simplification des flux d'information.

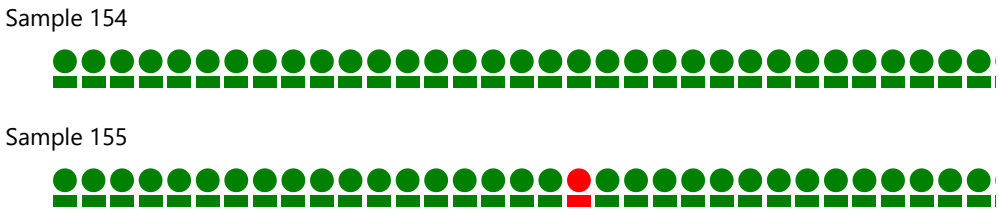

FIG. 5 – *Prédiction sur 2 échantillons de l'ensemble de données Cortez et Silva (2008). Un échantillon est correctement prédit par tous les arbres, l'autre est correctement prédit par tous les arbres sauf un.*

7 Conclusion

Dans cette étude, nous avons présenté RFIViz, un outil de visualisation interactif et convivial conçu pour combler le fossé entre la complexité des modèles de forêt aléatoire et la capacité d'interprétation de l'utilisateur. En combinant les connaissances des méthodes existantes et les besoins des utilisateurs, RFIViz offre une approche "Détails à la demande" à travers trois vues clés - échantillon, forêt et arbre - permettant une compréhension complète à des niveaux de profondeur variables.

RFIViz simplifie les subtilités des modèles complexes, tant pour les experts que pour les novices, en aidant à identifier les arbres faibles ayant un impact sur les performances

globales et en démêlant la nature de la boîte noire des algorithmes d'apprentissage automatique.

L'application à un ensemble de données sur la réussite des étudiants a mis en évidence les prouesses de RFIViz dans l'identification des modèles de classification, l'offre d'informations exploitables et l'aide à la prise de décision.

Les améliorations futures incluront de nouvelles mesures, l'extension à d'autres méthodes d'ensemble, et une intégration plus profonde avec l'instanciation du modèle, réaffirmant notre engagement à affiner RFIViz en tant qu'outil accessible et complet pour l'interprétation des modèles.

Références

Chen, T. et C. Guestrin (2016). Xgboost : A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16. ACM.

Cortez, P. et A. Silva (2008). Using data mining to predict secondary school student performance. *EUROSIS*.

Cutler, A. et L. Breiman. Raft (random forest tool).

Geurts, P., D. Ernst, et L. Wehenkel (2006). Extremely randomized trees. *Mach. Learn. 63*(1), 3–42.

Kulkarni, V. Y. et P. K. Sinha (2012). Pruning of random forest classifiers : A survey and future directions. In *2012 International Conference on Data Science Engineering (ICDSE)*, pp. 64–68.

Neto, M. et F. Paulovich (2020). Explainable matrix—visualization for global and local interpretability of random forest classification ensembles. *27*(2), 1427–1437.

Nsch, R. H., P. Wiesner, S. Wendler, et O. Hellwich (2019). Colorful trees : Visualizing random forests for analysis and interpretation. pp. 294–302.

Rhodes, J. S., A. Cutler, G. Wolf, et K. R. Moon (2020). Supervised visualization for data exploration. *ArXiv abs/2006.08701*.

Teoh, S. et K. Ma (2003). PaintingClass : Interactive construction, visualization and exploration of decision trees. pp. 667–672.

Welling, S., H. Refsgaard, P. Brockhoff, et L. Clemmensen (2016). Forest floor visualizations of random forests.

Yang, F., W. hang Lu, L. kai Luo, et T. Li (2012). Margin optimization based pruning for random forest. *Neurocomputing 94*, 54–63.

Zhang, H. et M. Wang (2009). Search for the smallest random forest. *Statistics and its interface 2*, 381.

Zhao, X., Y. Wu, et D. Lee (2019). iForest : Interpreting random forests via visual analytics. *25*(1), 407–416.

Summary

Random Forests (RFs) stand as a widely adopted machine learning tool for classification tasks due to their interpretability relative to complex models like neural networks. However, in scenarios where RFs are constructed within expansive feature spaces and abundant samples, their interpretability diminishes as the previous parameters grow. This loss of interpretability impedes users from comprehending the rationale behind sample classifications. Our research addresses this challenge by proposing a visualisation tool for better understanding of RF models whatever their size, particularly tailored for non-expert users. We recognize the necessity for interactive tools that facilitate a deeper grasp of the algorithmic reasoning employed by RF models and especially filter and find systems to overcome the high dimensional feature space.